universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Algorithmic Venture Insights
Exploration of Machine Learning in Predicting Startup Success and Enhancing Investment Decision-Making

verfasst von | submitted by

Tomáš Gallo

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 974
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Banking and Finance
Betreut von | Supervisor:	Dipl.-Inform. Dipl.-Volksw. Dr. Christian Westheide

Declaration

The work contained in this thesis is my own work unless otherwise stated.

ACKNOWLEDGMENTS

This work would have not been originated without consultations with my supervisor, Dipl.-Inform. Dipl.-Volksw. Dr. Christian Westheide. He challenged my thinking on the strategy, methodology, and concept of the work, as well as helped me to achieve exactness and clarity in the text.

*I would also like to express my gratitude to **all** my teachers who had shown great performance when pushing me to my current position.*

Additionally, I would like to thank to Berthold Baurek-Karlic from Venionaire Capital AG for giving me the opportunity to gain practical experience in the world of Venture Capital.

Finally, I would like to express my gratitude to my parents for their unwavering support and understanding. It is important to acknowledge that the world we live in today is vastly different from the world they experienced in their youth.

ZUSAMMENFASSUNG

Die vorliegende Studie hat zum Ziel, die Einsatzmöglichkeiten moderner Methoden des maschinellen Lernens (ML) für Venture-Capital-Geber (VCs) im Rahmen ihrer Entscheidungsprozesse zu demonstrieren. Die finanzielle Unterstützung potenzieller Start-ups birgt das Risiko, dass nur ein geringer Anteil der Gründungen erfolgreich sein wird und Gewinne erzielen wird. In der Realität scheitern die meisten Start-ups, was zu einem Verlust des investierten Kapitals führt. Um die Wahrscheinlichkeit von Fehlentscheidungen zu reduzieren, wurden verschiedene ML-Methoden eingesetzt, um praktische Szenarien zu simulieren. Die hier präsentierten Ergebnisse sind vielversprechend. Die empirischen Ergebnisse deuten darauf hin, dass Modelle wie XGBoost effektiv eingesetzt werden können, um datengestützte Entscheidungen im VC-Bereich zu treffen. Es sei jedoch darauf hingewiesen, dass keine Methode die Möglichkeit von Täuschungen durch Antragsteller sowie die Rolle des Zufalls vollständig ausschließen kann. Vor diesem Hintergrund werden trotz der Integration von Spitzencomputingmodellen in den Bewertungsprozess der Kandidaten auch in Zukunft Schätzungen bleiben, die auf Basis von Wahrscheinlichkeitsaussagen getroffen werden.

ABSTRACT

The objective of this study is to demonstrate how modern machine learning methods can assist VCs in their decision-making process. When considering potential financial support, it is difficult to predict which start-up will be successful and generate a profit. In reality, the majority of start-ups are unsuccessful, resulting in a loss of invested capital. To reduce the likelihood of incorrect decisions, we utilized various ML methods to simulate practical scenarios. The results presented here are promising. The empirical results indicate that models such as XGBoost can be effectively used to make data-driven decisions in VC. However, it is important to acknowledge that no method can completely eliminate the possibility of deception by applicants as well as the role of chance. Therefore, predictions will remain estimations for the foreseeable future, despite the integration of top computational models in the candidate evaluation process.

LIST OF THE MOST FREQUENTLY USED ABBREVIATIONS

AI=artificial intelligence

AUC=Area Under Curve

DL=deep learning

ML=machine learning

ROC=receiver operating characteristic curve

XGBoost= eXtreme Gradient boosting

VC=venture capital

VCst(s)=venture capitalist(s)

Note: References were formatted according to the requirements of the Journal of Finance.

Content

1 INTRODUCTION 9

1.1. Background 9

1.2. Problem statement 10

1.3. Research questions 10

1.4. Thesis outline 11

2 LITERATURE REVIEW 12

2.1 Overview of venture capital: macro level look at venture capital 12

2.2 Venture-capitalist's decision-making: micro level look at venture capital 13

2.3 Unveiling machine learning 14

2.3.1 Overview and definition of machine learning 14

2.3.2 Brief history of machine learning 15

2.3.3 Fundamentals of machine learning 16

2.3.4 Types of machine learning 16

2.3.5 Hyperparameter tuning 17

2.3.6 Importance of data quality 18

2.4 Augmenting venture-capitalist's decision-making: the algorithmic revolution 19

2.4.1 The human-machine synergy: combining human judgment with algorithmic insights 19

2.4.2 Reducing biases and broadening horizons: counteracting human biases in venture capital 19

2.4.3 Existing forays: machine learning in venture capital 20

3 METHODOLOGY 23

3.1 Data Collection 24

3.2 Exploratory data analysis 24

3.3 Data cleaning and feature engineering 26

3.4 Defining target variable 27

3.5 Overview of selected machine learning models 27

3.5.1 Introduction to model selection and training 27
3.5.2 Dataset split 27
3.5.3 Hyperparameter tuning and model fitting overview 28
3.5.4 Random Forest Classifier 28
3.5.5 XGBoost 28
3.5.6 Logistic Regression 29

4 EMPIRICAL RESULTS 30

4.1 Model comparison metrics 30

4.2 Visual analysis of further parameters 32
4.2.1 Confusion matrix 32
4.2.2 Receiver operating characteristic curve and area under curve 33
4.2.3 Precision-Recall curves 34

4.3 Feature importance analysis 35

4.4 Interpretation of results 38
4.4.1 Effectiveness in predicting startup outcomes 38
4.4.2 Practical application in venture capital 39

4.5 Limitations of the study 39

4.6 Summary of the outcomes 41

5 DISCUSSION 42

6 CONCLUSION 45

7 FUTURE DEVELOPMENTS 46

8 REFERENCES 48

1 INTRODUCTION

1.1. Background

In finance, a shift in paradigm towards data-driven decision-making has transformed investment strategies. This evolution started in the hedge fund sector in the 1980s, signifying a transformative shift from conventional, human judgment-based methods to those prioritizing quantitative and algorithmic strategies. Pioneering hedge funds Renaissance Technologies and D. E. Shaw & Co. employed computational analysis, utilizing machine learning and complex algorithms to navigate financial markets (Ionescu and Diaconita (2023)). The success of these quantitative hedge funds demonstrated the potential of machine learning and data-driven strategies in finance, paving the way for the widespread adoption of these technologies in the industry (Emerson (2019)).

This seismic shift in the public markets, marked by the successful integration of data analytics and algorithmic methods, established a precedent that is now gaining traction in the private equity realm, specifically in venture capital (VC), (Ferrati and Muffatto (2021)). VC as an alternative investment asset class, contributes significantly to economic growth by supporting innovative start-ups, yet unlike the technological transformation of the public markets, VC remains a largely people-driven industry. The decision-making process in VC involves evaluating both quantifiable elements such as financial projections, market trends, and the competitive landscape, as well as less tangible elements such as the quality of the founding team's "jockey" and the viability of the business model's “horse”(Desarbo, Macmillan and Day (1987); Gompers, Gornall, Kaplan and Strebulaev (2020); Gompers, Gornall, Kaplan and Strebulaev (2021); Kaplan and Lerner (2010)).

As the financial industry continues to develop, venture capitalists (VCst) are considering utilizing the techniques and concepts that transformed hedge funds to improve their own investment strategies and decision-making processes. This consideration is motivated by the inherent complexities of assessing startups, with high stakes and limited information (Sommer, Loch and Dong (2009)). A standard VCst assesses roughly two hundred startups each year but invests in only about four, demonstrating the rigorous and selective character of this method (Gompers, Gornall, Kaplan and Strebulaev (2021)). The challenge stems from the dearth of historical data on emerging businesses and the intricate interplay of traits unique to entrepreneurs and startups. Identifying the most promising ventures from a pool of numerous contenders is an immensely arduous task that is often susceptible to errors, given the subjective nature of human judgment and the time constraints associated with each potential deal. In spite of its obvious practical importance, the issue lacks adequate exposure in both the research community and the financial industry (Davidsson (2008)).

From an industrial standpoint, numerous venture funds such as Google Ventures, Correlation Ventures, EQT, and Earlybird have achieved substantial success in implementing smart algorithms for assistance in their decision-making processes. However, the operations of their algorithms are confidential and diligently guarded, and publicly accessible research endeavors to replicate and enhance their advancement would provide significant worth to the research community (Cao (2023); Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov and de Rijke (2018)).

1.2. Problem statement

The VC industry has historically relied on human intuition and networks, which has proven beneficial. However, this dependence poses significant challenges to scalability, limits the diversity of opportunity sourcing, and introduces potential biases in decision-making processes. Integrating machine learning (ML) offers an opportunity to address these issues by enhancing predictive precision and revealing latent patterns in startup success. VC firms can use ML to uncover valuable insights that may be difficult for humans to analyze. This can lead to more informed and equitable investment decisions.

1.3. Research questions

The purpose of this thesis is to investigate the role of ML algorithms in enhancing decision-making in the VC sector. The research goal is to narrow the gap between traditional VC methods that rely on human judgment and intuition, and the developing field of ML-based data-driven decision-making.

The study aims to examine the efficacy of ML algorithms in predicting outcomes for startups. Specifically, it explores how these algorithms can potentially outperform or supplement conventional methods employed by human investors.

Through this research, the thesis aims to contribute to the existing body of knowledge by:

1. Demonstrating the potential of ML in enhancing the predictive accuracy in venture investment decisions.

2. Identifying specific areas within the VC process where ML algorithms can provide significant value-addition.

3. Highlighting the unique insights and patterns that ML can unveil, which could revolutionize the approach to startup evaluation and investment.

1.4. Thesis outline

- Chapter 2: Literature Review - Examines the historical context of VC, introduces ML, and explores its applications in VC.
- Chapter 3: Methodology - Describes the research methodology, including data collection, machine learning models used, and analytical techniques.
- Chapter 4: Empirical Analysis - Presents the results of the machine learning models applied to VC data.
- Chapter 5: Discussion - Discusses the implications of the findings, comparing ML-driven approaches with traditional VC decision-making.
- Chapter 6: Conclusion – summarizes the outcomes of the study.
- Chapter 7: Future developments – outlines potential areas for future study.

2 LITERATURE REVIEW

2.1 Overview of venture capital: macro level look at venture capital

VC is a distinct form of private equity that concentrates on investing in young firms exhibiting high-growth potential, thus fostering innovation in the economy (Greenwood, Han and Sánchez (2022)). Such investments in young innovative companies bring a significant amount of risk into the VC 's portfolio. Despite significant advancements in technology and business practices, the failure rate of startups has remained relatively constant over the past few decades. It is estimated that around 90% of startups fail within their first five years, indicating a high rate of failure in the startup industry (Potanin (2023)).

VC returns follow a power law distribution, where a small number of investments yield returns far greater than the rest, often exceeding them combined (Crawford, Aguinis, Lichtenstein, Davidsson and McKelvey (2015); Retterath (2020)). Research has shown that VC-backed ventures have a higher survival rate compared to non-VC-backed ones (Sandberg and Hofer (1987); Zacharakis and Meyer (1998)). VC investors play a crucial role in the growth and success of startups, providing not only funding but also expertise and strategic networks (Davila, Foster and Gupta (2003)). Kaplan and Strömberg assert that VCsts are particularly adept at addressing a crucial principal-agent problem in market economies, which involves connecting entrepreneurs with innovative ideas (but no financial resources) with investors who possess capital (but lack ideas), (Kaplan and Strömberg (2001)).

The allocation of resources from the hands of VCs to ambitious startup founders is bound by incentive contracts designed to ensure that entrepreneurs are rewarded for strong performance while also allowing investors to assume control if the entrepreneur underperforms (Kaplan and Strömberg (2003)). These incentive contracts, referred to as term sheets, are formally documented agreements that outline the terms of funding and investment in a company. Gompers conducted interviews with 885 industry professionals and discovered that VCs exhibit less flexibility in negotiating certain contractual terms compared to others during the funding deal (Gompers, Gornall, Kaplan and Strebulaev (2020)). His findings reveal that VCs prioritize pro-rata investment rights (rights allowing investors to maintain their ownership percentage in subsequent financing rounds), liquidation preferences (the right to receive a specified amount before other shareholders in the event of a sale or dissolution of the company), anti-dilution protection (provisions that protect investors from a decrease in ownership percentage in future financing rounds), vesting (a schedule determining when founders or employees earn their stock or options, encouraging long-term

commitment), valuation (the agreement on the monetary value of the company, impacting the share of ownership exchanged for investment), and board control (the ability to influence or make decisions by holding board seats, impacting the company's strategic direction) in their investment agreements.

2.2 Venture-capitalist's decision-making: micro level look at venture capital

During a deal flow, particularly in the screening and evaluation phases, equity investors encounter two types of risks: adverse selection and moral hazard (Fried and Hisrich (1994)). Adverse selection arises when an investor chooses to invest in a company that ultimately fails (type 1 error) or fails to invest in a company that would have been successful. Moral hazard is a concern as entrepreneurs may decrease their commitment or make the project riskier over time. To reduce the risk of both type 1 and type 2 errors, investors aim to minimize information asymmetry between themselves and the entrepreneurial team (Jaimovich (2010)).

When evaluating a funding application, an investor must accurately estimate the success or failure of the company to make an informed decision on whether to invest or not, minimizing type 1 and type 2 errors (Cao (2023)). The investor can only rely on directly observable factors such as the characteristics of the entrepreneurs and management team, the product or service, the market, and financials to form an estimate of the venture's outcome (Ferrati and Muffatto (2021)).

This chapter explores the basic operations of VCsts, following the framework of sourcing, screening, selection, and post-investment activities (Tyebjee and Bruno (1984)). Sourcing potential investment opportunities represents a critical initial step in the VC process. Gompers and colleagues emphasize in their research that VCs heavily rely on their professional networks, highlighting the significance of established relationships and reputation in the industry. This reliance on networks suggests that deal sourcing involves more than merely securing any opportunity; it requires sourcing deals aligned with the VC's expertise, investment thesis, and sector preferences. Moreover, this heavy reliance on personal networks, while leveraging established relationships and reputation, may inadvertently narrow the scope of opportunities, potentially overlooking innovative ventures outside these networks (Gompers, Gornall, Kaplan and Strebulaev (2020)).

The diligent screening process that follows sourcing is essential in pinpointing feasible investment prospects. Hsu et al. emphasize that angel investors and VCs assess various factors, with VCs placing special significance on ventures with robust economic prospects and competent

management teams (Hsu, Haynie, Simmons and McKelvie (2014)). This stage indicates a concentration on both the business's viability and the team's ability to execute.

During the selection phase, VCs prioritize an in-depth evaluation of the management or founding team's skills and passion over other factors, as suggested by Gompers et al. (Gompers, Gornall, Kaplan and Strebulaev (2020); Gompers, Gornall, Kaplan and Strebulaev (2021)). While business model, market size, and industry dynamics also hold significance, they are subordinate considerations in comparison to the quality of the management team. While this focus aligns with the high-risk nature of venture investing, it may also result in overlooking quantifiable aspects of a startup's potential, such as market size or technological innovation, due to human preference for familiar or easily understandable traits. Furthermore, the evaluations of VCsts can be skewed by cognitive limitations such as local bias, overconfidence, and loss aversion, which are inherent human biases (Blohm, Antretter, Sirén, Grichnik and Wincent (2022)). In addition, it is important to evaluate the compatibility between the personalities of the entrepreneurs and the VCst, as they will be working closely together in the post-investment phase (Ferrati and Muffatto (2021)).

In post-investment activities, VCs' roles extend beyond capital provision to strategic guidance and network building, as described by Gompers et al. After investing in a startup, VCsts typically assume roles as board members and advisors with the goal of accelerating the company's growth and overall performance (Gompers, Gornall, Kaplan and Strebulaev (2020)). Furthermore, VC investors who have previously invested in a startup assist with fundraising by connecting the startup to potential investors for follow-on funding rounds (Retterath (2020)). However, this hands-on approach, while beneficial, can be limited by the VCs' personal capacity, expertise, and biases, potentially impacting the breadth and depth of the support provided to the portfolio companies.

2.3 Unveiling machine learning

2.3.1 Overview and definition of machine learning

ML, a branch of artificial intelligence (AI), has revolutionized data comprehension and utilization. It's now a critical tool in our technological era, allowing systems to learn, evolve, and (fully or semi-) autonomously make decisions or predictions. This shift from traditional programming – which relies on explicit instructions – to self-learning approaches marks a fundamental change in how we interact with data and technology (Raschka (2019)).

Today, we exist in an era abundant with both structured and unstructured data presenting an exceptional chance to turn raw data into useful knowledge - "Giving machines the ability to learn from data" (Raschka (2019)). This contrasts with the traditional approach of manually extracting rules and building models from large datasets and offers a more efficient alternative. ML allows the acquisition of knowledge embedded in data, facilitating gradual improvement of predictive models, and enhancing data-driven decision-making processes. The importance of ML exceeds academic research in computer science; it has become a crucial aspect of our daily lives and influences how we interact with technology powered by data (Israel (2020)).

ML algorithms are fashioned to learn from and conclude decisions based on data, with no manual derived rules by humans. This innovative approach has opened new possibilities in various domains, enabling effective handling and analysis of large quantities of data. The development of ML subfield within AI signifies a critical juncture in technological advancement (Israel (2020); Mitchell (1997)).

2.3.2 Brief history of machine learning

The development of ML started in the 1950s, aiming to design machines that could learn and imitate human intelligence. The early stages of this journey involved groundbreaking inventions like the perceptron, created by Frank Rosenblatt in 1957, which paved the way for neural networks. Unfortunately, the field faced a period of limited development during the "AI winter," where a lack of computational power and access to data resources restrained progress (Raschka (2019)).

The development of ML began in the 1950s with a goal to create machines capable of learning and emulating human intelligence. The early stages of this effort included innovative inventions such as the perceptron, invented by Frank Rosenblatt in 1957, which set the foundation for neural networks. Regrettably, the field encountered a period of restricted development known as the "AI winter," during which limited access to data resources and computational power hindered progress (Raschka (2019)).

The continuous improvement of algorithms and exponential surge in data availability has transformed ML into a dynamic and rapidly advancing field. As ML progresses, it holds the promise of unlocking even more transformative applications that could reshape industries and influence future technological advancements (Jordan and Mitchell (2015)).

2.3.3 Fundamentals of machine learning

Statistical learning theory is the basis of ML, providing a framework that allows systems to analyze data, detect patterns, and generate predictions. The ability to accurately forecast future outcomes makes this theory crucial for enhancing ML algorithms. It enables systems to understand the relationships within vast amounts of data and predict outcomes with precision. Statistical learning theory essentially furnishes ML algorithms with the ability to learn from data, a capability that has become indispensable in numerous industries, such as finance (Israel (2020); Jordan and Mitchell (2015)).

2.3.4 Types of machine learning

I. Supervised learning: Supervised learning is a critical part of ML that involves training models on labeled data to predict outcomes for new data. There are two primary types of supervised learning: classification and regression. Classification tasks involve categorizing data into distinct classes, such as distinguishing between spam and non-spam emails with email filters. On the other hand, regression focuses on predicting continuous outcomes, such as forecasting stock prices based on economic indicators. This methodology enables accurate forecasts in diverse applications, spanning from digital communication to financial analysis. Thus, indicating its extensive relevance and significance in data-driven decision-making (Raschka (2019)).

   Deep learning (DL; a subset of supervised learning): in DL employ artificial neural networks (ANNs) to extract progressively more abstract features from raw input. For instance, when detecting financial fraud, the initial layers may handle fundamental transaction details, while deeper layers can detect intricate patterns that indicate fraudulent operations. ANNs typically have thousands (often millions) of parameters, leading to extremely complicated nonlinear relationship between the input features/factors and output predictions. As a result, DL models suffer from the criticism that they are mysterious black-boxes, as opposed to some white-box ML models like logistic regression. However, explaining why a DL model comes up with certain predictions for startups (especially diverging predictions for comparable startups) is crucial for investment professionals (Cao (2023)).

II. Unsupervised Learning: Unsupervised learning analyzes unlabeled or structurally unknown data to uncover significant insights without predetermined outcomes or rewards. Its methodology encompasses clustering and dimensionality reduction techniques. Clustering groups data according to similarities and creates meaningful subgroups or clusters; it is commonly

used in fields such as marketing to identify customer groups with shared interests. Dimensionality reduction, another facet of unsupervised learning, compresses high-dimensional data to address storage and computational efficiency challenges. It filters out noise and preserves relevant information in a condensed form, thereby enhancing the predictive performance of ML algorithms by focusing on the most significant data aspects (Raschka (2019)).

III. Reinforcement learning: Reinforcement learning is a distinct type of ML where a system, termed as an 'agent', learns to make decisions by interacting with its environment. Unlike supervised learning, reinforcement learning doesn't rely on correct ground truth labels; instead, it uses a reward signal as feedback, indicating the effectiveness of the agent's actions. The agent aims to maximize a reward by implementing trial-and-error or strategic planning, continuously refining strategies based on environmental feedback. A typical case of reinforcement learning is a chess engine, where the agent determines a sequence of moves based on the chessboard's current state. The reward is determined by the game's outcome, whether it be a victory or a loss. This approach enables the agent to acquire optimal strategies and decision-making abilities in dynamic and intricate settings (Raschka (2019)).

2.3.5 Hyperparameter tuning

To improve ML models, hyperparameter tuning is utilized. Hyperparameter tuning in ML involves adjusting the settings of a model to optimize its performance. It is comparable to fine-tuning the controls of a complex gadget to achieve optimal results. The process of improving a ML model's performance by adjusting its non-data-learned parameters is known as fine-tuning. This involves setting the controls of the model, which it cannot learn on its own and must be configured by the user. Techniques such as GridSearchCV systematically explore various parameter combinations to identify the most effective settings for a given dataset. Optimal hyperparameter settings are crucial for improving a model's ability to detect subtle patterns and make accurate predictions, thereby enhancing its overall effectiveness. It is important to fine-tune these settings to achieve the best results *(Fig. 1)*.

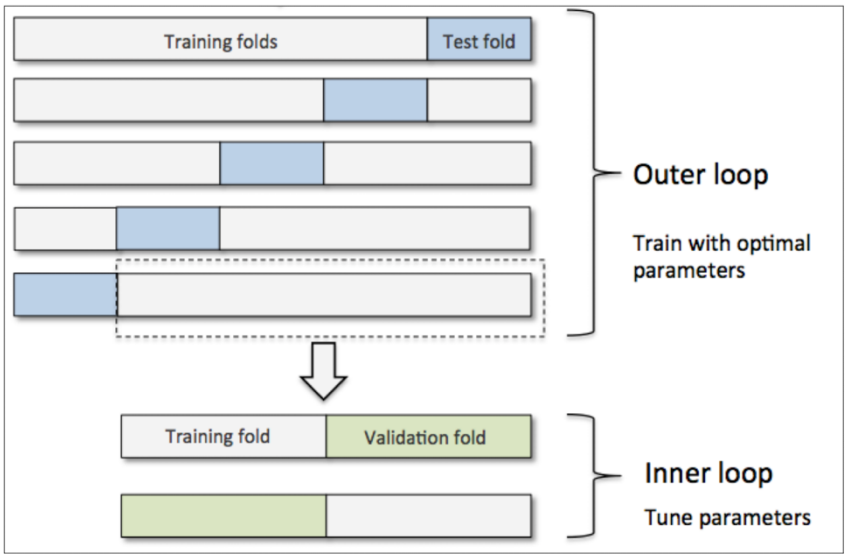

Figure 1 Cross validation (sample of a 5x2 nested cross validation)

2.3.6 Importance of data quality

The quality of data is crucial to the performance of ML models, as it significantly affects their efficiency, accuracy, and ability to handle complex tasks. To ensure high data quality, it is necessary to rigorously assess its accuracy, completeness, consistency, and relevance. Inadequate data quality can result in misleading analytics and unreliable decisions, which is especially problematic in fields such as VC where data-driven decisions have significant financial implications (Gupta (2023); Gupta, Mujumdar, Patel, Masuda, Panwar, Bandyopadhyay, Mehta, Guttula, Afzal, Mittal and Munigala (2021); Jain, Patel, Nagalapatti, Gupta, Mehta, Guttula, Mujumdar, Afzal, Mittal and Munigala (2020)). The importance of data quality in ML cannot be overstated. It is fundamental in algorithm development and deployment, especially as ML is increasingly used in critical decision-making processes. Robust models are needed to effectively handle data quality challenges and ensure reliable and unbiased outcomes. Integrating comprehensive data quality assessments into the ML workflow is not only an enhancement but a necessity to maintain the integrity and reliability of the resulting analyses and predictions (Budach (2022); Sessions (2006)).

2.4 Augmenting venture-capitalist's decision-making: the algorithmic revolution

"The VC industry is on the brink of a transformative era, where the combination of human insight and algorithmic analysis revolutionizes traditional investment decision-making models" (Retterath (2022)). This section explores the significant implications of this symbiotic relationship.

ML provides a computational advantage in the inherently subjective nature of VC decision-making. This synergy is not meant to replace human judgment but to supplement it with data-driven insights. This is supported by Nagar, who demonstrated that combining human and machine predictions can lead to more accurate and robust outcomes (Nagar (2011)).

2.4.1 The human-machine synergy: combining human judgment with algorithmic insights

Human comprehension of market intricacies, entrepreneurial mindset, and strategic compatibility is now augmented by the ability of ML to analyze massive datasets, identify patterns, and predict trends with an unrivaled accuracy that surpasses human capabilities. For example, while VCsts are dependent on their experience to assess the potential of a startup's managerial team, ML models can evaluate past data from comparable startups to anticipate the probability of success. This combination allows for a more extensive evaluation, utilizing the strengths of both human insight and machine accuracy (Ferrati and Muffatto (2021)).

2.4.2 Reducing biases and broadening horizons: counteracting human biases in venture capital

The incorporation of ML in VC extends beyond the primary startup screening and selection phases; it greatly influences the complete investment process, comprising the crucial post-investment phase. With the integration of ML, VCs can address human biases in the initial stages while improving their post-investment value addition significantly.

Retterath (2020) observes that geography and traditional network-based deal sourcing are becoming less relevant. The competition for high-potential deals is driving a shift towards active sourcing models, indicating an evolving landscape where ML can provide a competitive edge (Retterath (2020)).

Recent studies reveal the potential of ML in improving the alignment of VC investment theses. For example, a study used six ML models to measure the fit of developed companies to a VC

firm's investment thesis, with planning strategy and team management being the most determinant factors in investment decisions (Bai and Zhao (2021)).

Arroyo et. al conclude that ML can support venture investors in decision-making by predicting various outcomes, including subsequent funding rounds and company closure (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019)). Another study compared the performance of ML algorithms with that of VC investment professionals in screening early-stage companies, finding that the XGBoost algorithm outperformed the median VC by 25% and the average VC by 29% (Retterath (2020)).

Arroyo and Tuli both highlight the potential of ML in providing real-time insights and predictive analytics, enabling VCs to identify potential issues and growth opportunities (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019); Tuli (2023)). Recent studies have demonstrated the potential of ML algorithms in refining market positioning and product offerings for startups. Chen found that ML can accurately predict sustainable patterns in consumer product preferences, outperforming traditional forecasting tools (Chen (2023)). Paruchuri further emphasized the role of ML in market segmentation, targeting, and positioning, particularly through the use of k-means clustering algorithm (Paruchuri (2019)).

2.4.3 Existing forays: machine learning in venture capital

The integration of ML into VC is a rapidly evolving area of research with transformative potential for the industry. Over the past two decades, data-driven approaches have dominated research on predicting startup success (i.e. identifying startups that eventually become unicorns). However, most of the research has been analytical and statistical, rather than utilizing ML approaches. The literature review below critically examines key studies that have explored the application of ML algorithms in predicting startup success and enhancing investment strategies. These studies collectively underscore the shift towards data-driven decision-making in VC and highlight the nuanced approaches and methodologies adopted in recent research.

Dellermann et al explores the integration of machine and collective intelligence to predict early-stage startup success, especially in high-risk environments. It employs a design science research approach that combines ML with human judgment. The implications of this study for VC include improving decision-making accuracy in high-risk scenarios by presenting a new method of combining objective data analysis and subjective human judgment (Dellermann (2017); Dellermann, Lipusch, Ebel and Leimeister (2019)).

Sharchilev et al explores the prediction of startups' ability to attract further investment after initial funding using web-based open sources. The methodology focuses on predicting future funding events with an investor-centric approach and data augmentation. This approach significantly enhances predictive accuracy and provides deeper market insights, which are beneficial for investors and entrepreneurs (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov and de Rijke (2018)). Blohm et al. compares Business Angels and ML algorithms in early-stage investing. The study focused on overcoming cognitive biases affecting human decision-making. The study concluded that while experienced BAs can outperform ML algorithms, integrating human and machine decision-making could lead to optimal investment performance (Blohm, Antretter, Sirén, Grichnik and Wincent (2022)). Thirupathi et al (2021) aims to identify factors linked to the future success of diverse startups using an XGBoost model. The study demonstrates the effectiveness of ML models in assessing small venture viability. Further research is suggested into legal founding documents and valuation predictions (Thirupathi, Alhanai and Ghassemi (2021)).

Ross creates models to predict the success of startups, with a focus on exit scenarios and follow-on funding. The study employs ML ensembles and demonstrates significant implications for VC by predicting startup exits with higher accuracy than traditional methods (Ross (2021)).

Retterath compares the performance of ML algorithms with that of VC professionals in screening startups. The study emphasizes the efficiency and accuracy of ML models over traditional human-based methods in the VC sector (Retterath (2020)). Arroyo et al evaluates the potential of ML algorithms to predict success levels for early-stage companies. The study's findings are important for VCsts, demonstrating that ML can improve investment success rates and provide a more nuanced investment strategy (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019)). Corea's ML-based framework enhances VC decision-making by identifying key features that influence investment decisions. This suggests a shift towards a more data-driven model for VC investments (Corea (2021)). Ang et al predicts the post-money valuation and success probability of startups. The study demonstrates the predictive power of ML in evaluating startup success. It also suggests further research into alternative performance metrics and combining predictive analytics with prescriptive methodologies (Ang (2022)).

Bai et Zhao uses ML methods to determine valuable metrics for VCs. The study highlights the importance of qualitative factors in investment decisions, which contrasts with traditional methods that heavily rely on quantitative data (Bai and Zhao (2021)).

Źbikowski focuses on creating a predictive model to forecast business success without bias. The study takes a realistic approach to predicting business success, emphasizing the significance of geographical and industrial factors (Zbikowski and Antosiuk (2021)).

Stahl developed a predictive model for multi-stage startup success using time-series data. The research offers VC investors a tool for better investment decision-making, showcasing the potential of ML in enhancing the VC investment process (Stahl (2021)).

Caruso et al examined the advantages of utilizing ML, specifically DL, in VC. The research offers valuable insights into the use of ML in VC, indicating a shift towards more data-driven and automated screening processes (Caruso (2015)).

This literature review highlights the versatile and evolving role of ML in VC. The studies discussed demonstrate the increasing importance of data-driven approaches in enhancing predictive accuracy for startup success and refining investment strategies. The presented methodologies, datasets, and perspectives contribute to a deeper understanding of the role of ML in VC and lay the groundwork for future research. As the VC ecosystem evolves, the integration of ML promises to profoundly redefine its landscape. In agreement with Bill Maris, former managing partner at Google Ventures, the availability of extensive data sets offers VCsts an exceptional opportunity to move beyond intuition-based investments towards more informed, algorithmic decision-making. As technology advances, the feasibility of utilizing sophisticated algorithms becomes more accessible, prompting investors to adopt AI for more strategic investment choices (Antretter (2020); Antretter, Siren, Grichnik and Wincent (2020)).

3 METHODOLOGY

After presenting the theoretical framework and potential of ML as a transformer of VC decision-making, this chapter now focuses on its practical application. We explain the methodology used in this research to explore the application of ML in VC, detailing the systematic and empirical approach taken.

This study goes beyond the conceptual understanding of ML's potential in VC. It seeks to empirically test, analyze, and validate an algorithm 's capability to be trained on data about startup companies.

In this chapter we will define the methodological framework used in this research, describing each step from data collection to model development and validation. My approached is inspired by a framework for ML process introduced by Raschka (Raschka (2019)), *Fig. 2*. The framework consists of the following steps:

- (3.2) Data Collection: This section introduces the primary datasets used in the study, including their sources and relevance to the research objectives. It details the origin, scope, and limitations of the data.

- (3.3) Exploratory Data Analysis: (EDA) is an essential step in understanding the characteristics of the dataset, identifying patterns and anomalies, and gaining insights that inform subsequent data processing and model fitting.

- (3.4) Data Cleaning and Feature Engineering: This section describes the procedures for data preprocessing, including handling missing values. Additionally, it covers the creation of new features that are expected to enhance model understanding and performance.

- (3.5) Defining the Target Variable: This section elaborates on the process of defining the target variable for the study, defining what our ML models will predict. It explains the rationale behind the chosen definition, its alignment with research objectives, and potential implications.

- (3.6) Overview of Selected ML Models: This section provides a discussion of the ML models chosen for this study, including their theoretical foundations, strengths, limitations, and applicability to the research problem.

- (3.7) Model Development: Insights into the model development process, focusing on the steps of hyperparameter tuning, cross-validation strategies, and the rationale behind the chosen parameters.

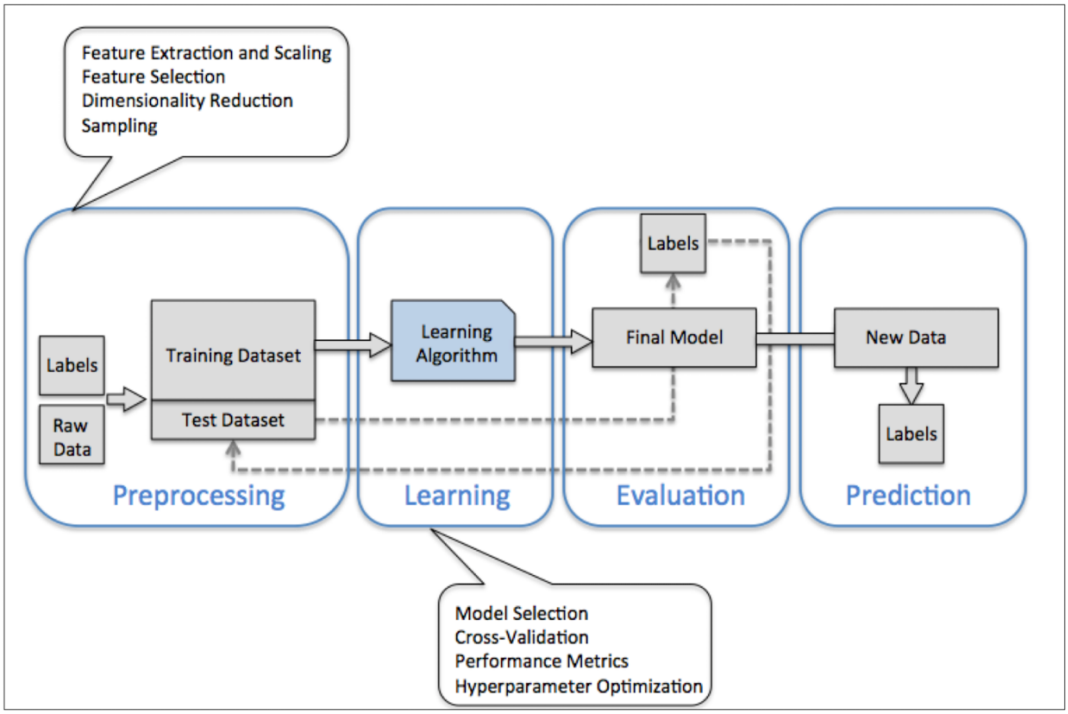

Figure 2 A workflow diagram for building a ML model (Raschka, 2015)

3.1 Data Collection

The primary dataset for this research was obtained from Crunchbase, accessed through Kaggle.com which is known for its extensive repository of datasets. This dataset contains information on companies and their funding rounds, with data dating back to December 2015. While the dataset's cutoff date limits the analysis to historical data, it provides a substantial basis for understanding startup dynamics and VC decisions during the specified time period.

3.2 Exploratory data analysis

The "Companies.csv" is our main dataset *(Fig. 3)*, which comprises 66,368 entries and provides a comprehensive overview of various companies. It includes 14 columns of data on company details such as name, URL, category, total funding raised, status, location, and funding dates.

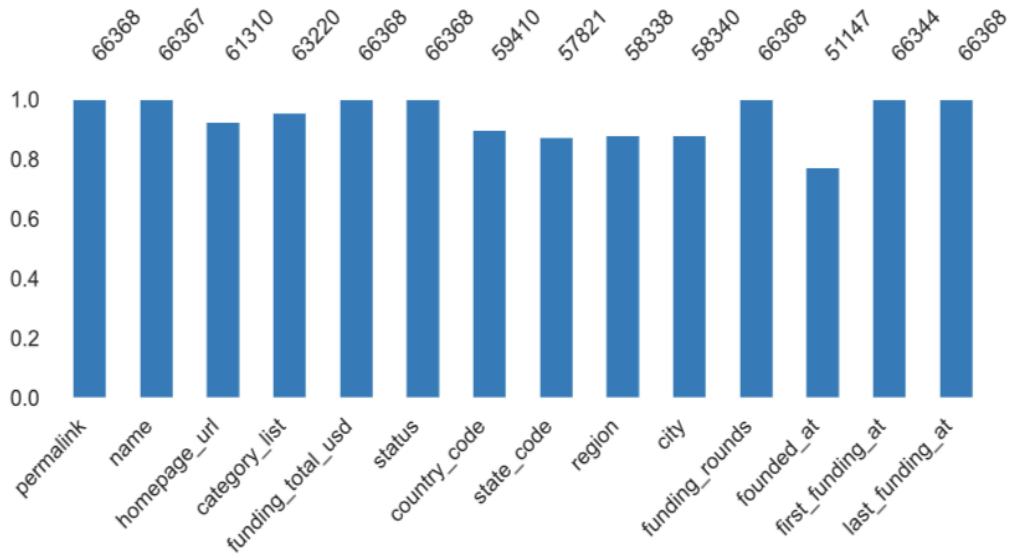

Figure 3 Companies.csv raw dataset

The "Rounds.csv" dataset comprises 114,949 entries and offers comprehensive insights into various aspects of funding rounds *(Fig. 4)*. It complements our main dataset about companies, as it provides additional information about the companies' funding and allows for engineering of useful features.

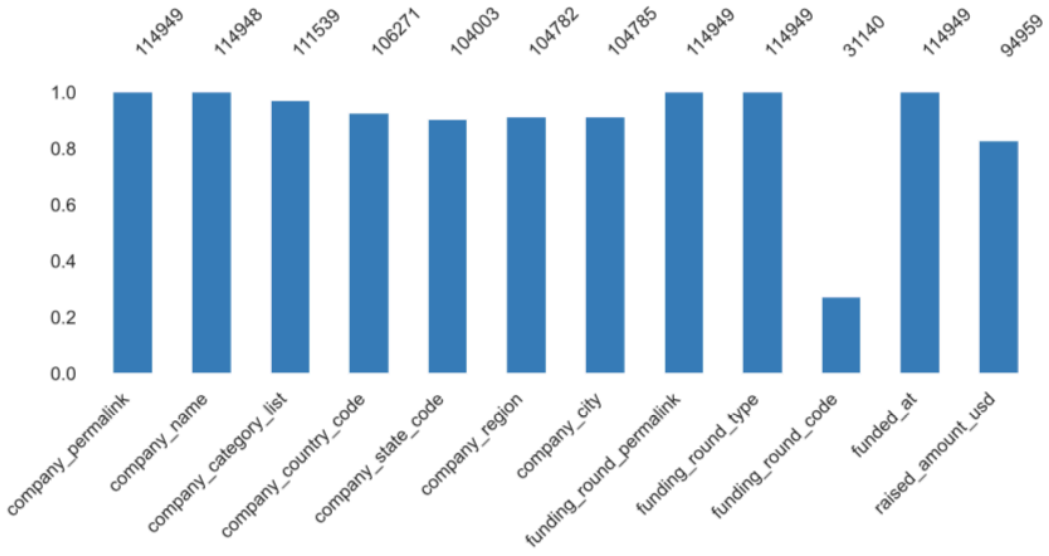

Figure 4 Rounds.csv raw dataset

3.3 Data cleaning and feature engineering

During the initial phase of data preparation, addressing missing values was crucial. To avoid losing valuable information, missing values were replaced with zero value placeholders. Additionally, supplementary columns were included to explicitly denote the presence of such placeholder values. This approach ensured the integrity of the dataset. Moreover, we converted date columns to datetime objects to enable precise temporal analysis. Additionally, we refined our dataset to include only companies founded between 1998 and 2015, ensuring consistency and relevance. Special attention was given to companies that were still operating as of the dataset cutoff. To ensure a more accurate assessment of success indicators, only companies founded before 2013 were considered, providing a minimum two-year operational window.

During the subsequent phase of feature engineering, four features were developed: ‘avg_time_funding’ (the average time between funding rounds), ‘operating_time’ (operating time of companies), ‘num_of_rounds’ (number of funding rounds), and ‘single_category’ (unique industry category). The calculation of the average time between funding rounds is based on the funding dates from the Rounds.csv dataset. This metric indicates a company's ability to attract frequent investments and reflects robust market confidence. The operating time feature quantifies the duration a company has been operational, providing insight into its maturity and potentially its risk level. The 'number of rounds' feature tallies the instances a company received funding, serving as another indicator of market trust. To streamline analysis, the category list was managed by tokenizing category tags and identifying the top 25 most common industries. Companies that did not fit into these categories were classified as 'other,' simplifying the dataset while maintaining its richness for predictive modeling. This simplification was necessary to make the data manageable for analysis, although it introduced a potential bias toward over-represented categories.

The data was prepared for machine learning analysis through meticulous feature engineering and significant data transformations and encoding. All non-numeric data was systematically converted to either integer or float data types for computational compatibility. Furthermore, we transformed key categorical variables, such as single category, country code, city, and founded at, using one-hot encoding. This process converts categorical entries into a binary matrix, which is essential for incorporating qualitative data into our quantitative analysis framework. It allows for effective integration and interpretation by the machine learning models.

3.4 Defining target variable

The target variable was defined based on the status of the companies in the dataset, labelled as "successful" (1) and "unsuccessful" (0). The successful category included companies that were acquired, went public (IPO), or companies that are marked as “operating” in our dataset, are at least two years since founding date and had three or more rounds of financing. Conversely, companies classified as unsuccessful were either closed or operating with two or fewer rounds of financing. This classification involved multiple assumptions, particularly for "operating" companies, that balanced the need for a clear indicator of success with the limitations of the dataset. To achieve a clear binary classification of the company's status, we must make certain assumptions, even if they cannot be made in real-world applications.

A major challenge in defining the outcome variable was dealing with data imbalance. The final dataset included 42,848 companies scored as '0' (unsuccessful) and 14,354 scored as '1' (successful). This imbalance posed a potential risk to the model's performance, particularly in terms of its ability to accurately predict the minority class (successful firms). Addressing this issue was critical to ensuring the validity and reliability of the ML models' results.

3.5 Overview of selected machine learning models

3.5.1 Introduction to model selection and training

In the field of machine learning, choosing the appropriate models for a given task is crucial for the success of any predictive analysis. Each model has its own strengths and weaknesses, making them suitable for different types of data and prediction tasks. To adhere to the No-Free-Lunch (NFL) theorem, which posits that no single model can universally perform best across all scenarios (Köppen, Wolpert and Macready (2001); Wolpert (2002); Wolpert (2023)), we utilized three ML models to conduct a comprehensive analysis of the dataset. This approach recognizes that although a model may perform well in certain conditions, its assumptions may not hold true in every context.

3.5.2 Dataset split

A consistent approach was used to divide the dataset into training, validation, and test sets, with a split of 60-20-20, respectively. This method guarantees that each model is trained, validated, and tested on distinct portions of the dataset, minimizing the risk of overfitting and enhancing the model's ability to predict unseen data reliably (Liu and Cocea (2017)).

3.5.3 Hyperparameter tuning and model fitting overview

The hyperparameters for each model were tuned uniformly to ensure a fair comparison between models. The process involved adjusting parameters, observing their effects on model performance, and selecting the optimal combination (Raschka (2019)). In the next section, we will introduce the models that were trained during this empirical study, and we will closely look at their training parameters.

3.5.4 Random Forest Classifier

The Random Forest Classifier (RFC) is an advanced ensemble learning technique that constructs multiple decision trees during training. It predicts outcomes by averaging or taking the majority vote of these trees, which significantly reduces overfitting and increases the model's accuracy on unseen data. RFC's ability to perform both classification and regression makes it adaptable to diverse datasets. Its built-in feature for assessing variable importance is particularly useful in identifying significant predictors within complex datasets (Amornsiripanitch, Gompers and Xuan (2019); Raschka (2019)).

The study conducted a grid search using a parameter grid that focused on the number of trees in the forest (n_estimators), the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf). The parameters were chosen to balance model complexity and computational efficiency, ensuring a thorough yet feasible hyperparameter tuning process. The GridSearchCV was configured with a cross-validation fold of 3 to optimize for accuracy while considering computational constraints (Raschka (2019)).

3.5.5 XGBoost

XGBoost (eXtreme Gradient boosting), short for Extreme Gradient Boosting, is a cutting-edge refinement of gradient boosting techniques that prioritizes speed and performance. It optimizes both the loss function and regularization component to efficiently address overfitting. The model enhances prediction accuracy by sequentially correcting errors made by preceding trees, with a strong emphasis on computational efficiency and scalability. XGBoost is a versatile option for various predictive modeling challenges, ranging from traditional classification to complex regression tasks (Dorogush (2018); Chen and Guestrin (2016)).

To optimize the XGBoost model, hyperparameters such as the number of gradient boosted trees (n_estimators), the maximum depth of a tree (max_depth), and the learning rate (learning_rate) needed to be adjusted. To balance computational efficiency with the robustness of the tuning process, we used a smaller number of 2 cross-validation folds. The aim of choosing these parameters was to explore a range of model complexities while managing computational load on my computer.

3.5.6 Logistic Regression

Logistic Regression is a fundamental model in the field of statistical ML, particularly suited for binary classification tasks. It models the probability of a binary outcome via the logistic function, providing a direct relationship between predictor variables and the probability of the target event. Logistic regression is often the preferred choice for making probabilistic predictions when understanding the influence of predictors is crucial. This is due to its simplicity, interpretability, and ease of implementation (Raschka (2019)).

The tuning of the Logistic Regression model focused on the regularization strength (C) and the choice of the solver algorithm (solver). The parameter grid was designed with a limited number of options for the regularization parameter to reduce computational burden. The 'liblinear' solver is an efficient choice for smaller datasets and binary classification problems. To balance the need for a robust model assessment with computational efficiency, we used a 3-fold cross-validation (Raschka (2019)).

4 EMPIRICAL RESULTS

This chapter presents an empirical analysis of the ML models developed, building upon the methodologies outlined in Chapter 3. The analysis aims to bridge the gap between theoretical model training and practical application, revealing insights into the predictive capabilities of each model in the context of VC. In this analysis, we evaluate the performance of the Random Forest, XGBoost, and Logistic Regression models in predicting the success and failure (Class 1 and Class 0) of startups. The purpose of this evaluation is to determine the practical utility of these models in enhancing decision-making processes within the VC industry.

4.1 Model comparison metrics

The performance of the model is evaluated using several key metrics: accuracy, precision, recall, and the F1 score *(Fig. 5, 6)*. These metrics provide a comprehensive view of each model's predictive ability, enabling us to comprehend their strengths and weaknesses in classifying startup outcomes (Yacouby (2020)).

Before discussing the empirical results, it is important to understand the theoretical background of the key performance metrics.

- Accuracy: Accuracy is a metric that reflects the overall correctness of the model and is calculated as the ratio of correctly predicted observations to the total observations.
- Precision: Precision indicates the proportion of positive identifications that were truly correctly predicted. In contexts where the cost of false positives is high, it is especially crucial to consider this metric.
- Recall (sensitivity): Recall measures the proportion of actual positives that were correctly identified. It is critical in contexts, where missing a positive (falsely misclassifying as negative) is detrimental.
- F1 score: The F1 Score is the harmonic mean of precision and recall. When dealing with uneven class distribution, F1 Score is used to balance precision and recall.

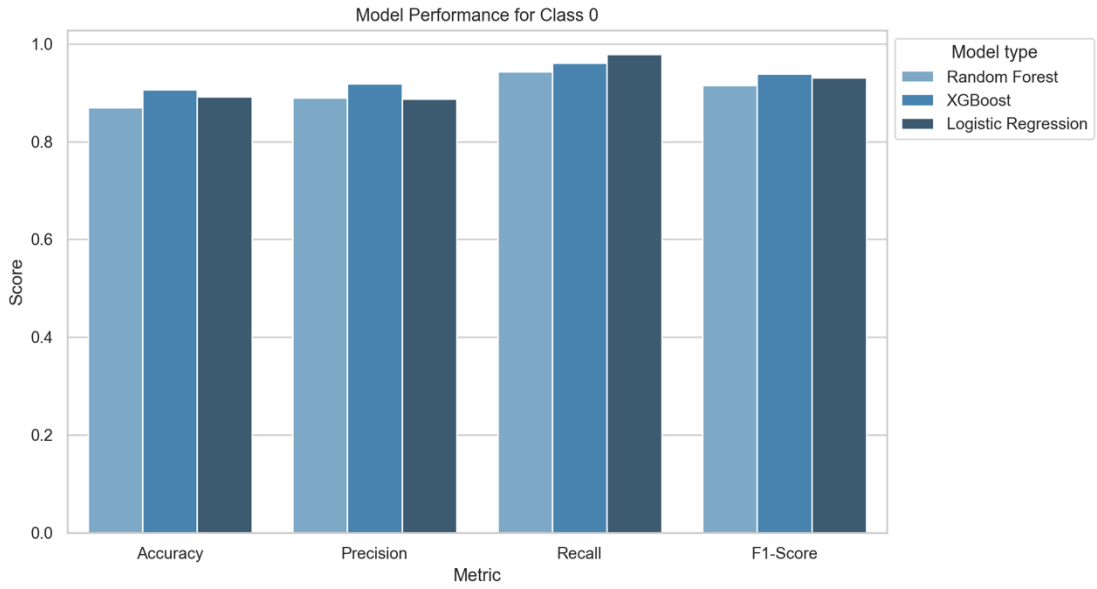

Figure 5 Performance metric for predicting startup failure (Class 0)

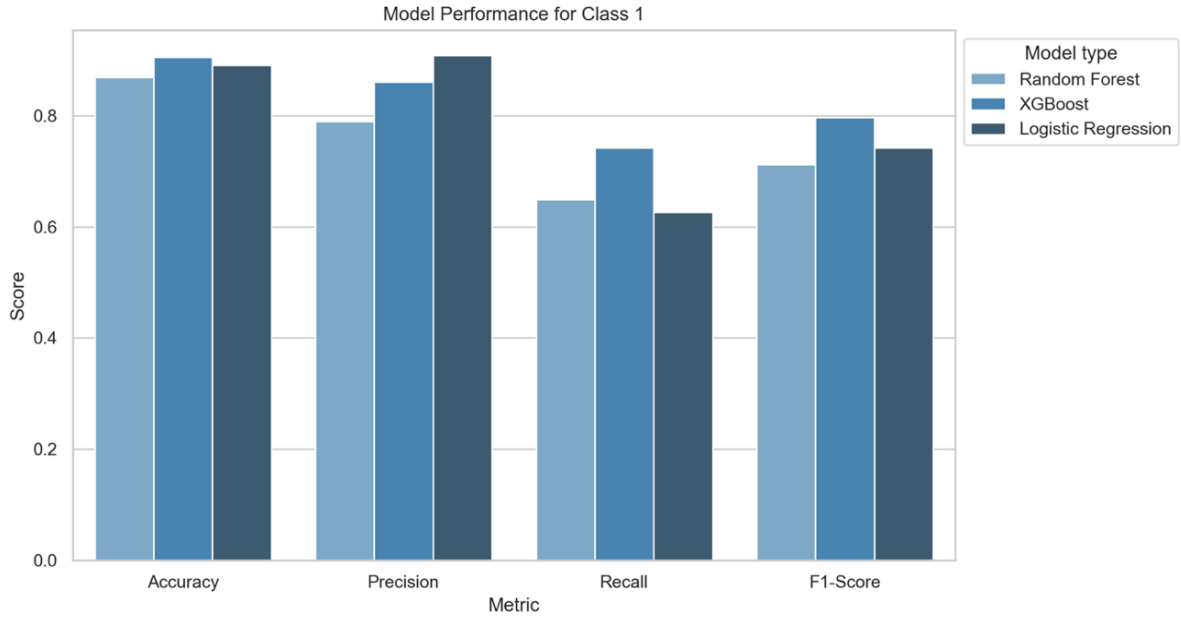

Figure 6 Performance metric for predicting startup success (Class 1)

The XGBoost model shows superior performance, with the highest accuracy and recall rates for successful startups. Notably, its recall rate of 0.74 for Class 1 predictions is higher than the rates of 0.65 and 0.62 for the Random Forest Classifier and Logistic Regression, respectively. This suggests that the XGBoost algorithm is less likely to miss out on a promising startup and less likely to falsely classify a venture as a negative prospect. This text demonstrates the strong capability of

correctly identifying potential successful ventures, which is arguably the most critical aspect for VC's decision-making.

The Logistic Regression model, known for its high precision in predicting successful startups, suggests its effectiveness in minimizing false positives, which is essential for investment strategies focused on 'picking winners' and increasing the probability of returning capital on startup investment.

The Random Forest Classifier, although displaying the lowest overall accuracy among our models, still exhibits significant precision and recall for unsuccessful startups.

4.2 Visual analysis of further parameters

Visual representations of model performance provide an intuitive understanding of their predictive capabilities. In this part, we will visualize each model 's:

- Confusion matrix
- ROC-AUC curves
- Precision-recall curves

4.2.1 Confusion matrix

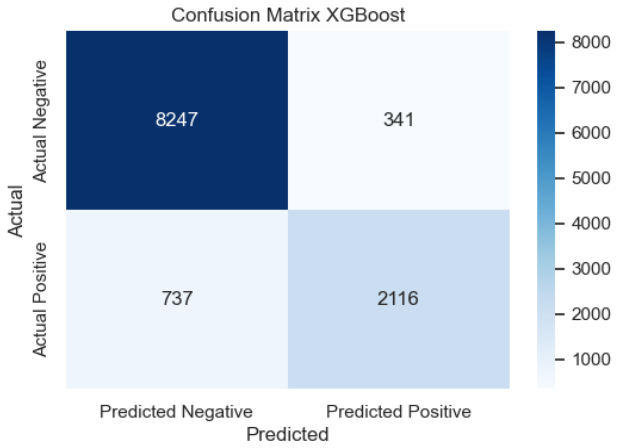

Figure 7 CM XGBoost

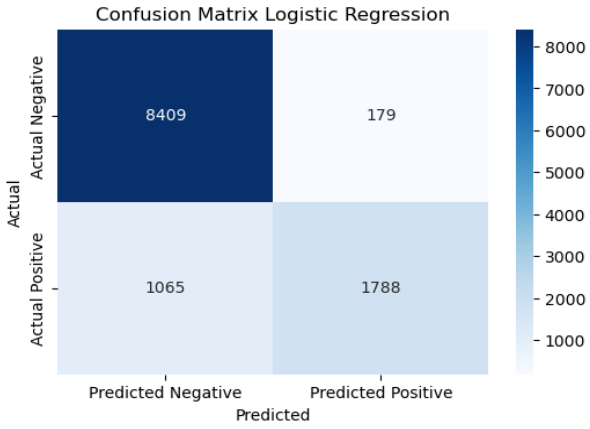

Figure 8 CM Logistic regression

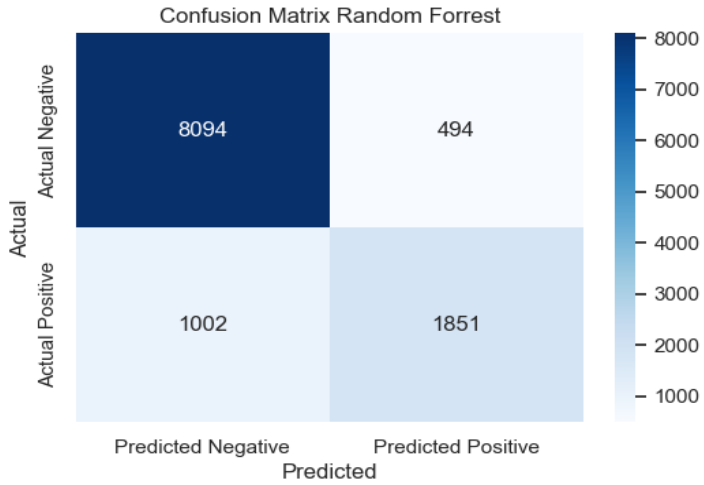

Figure 9 CM RFC

This matrix provides a snapshot of the true positive, false positive, true negative, and false negative predictions of each model *(Fig. 7, 8, 9)*. It is important to interpret these results to determine the models' ability to accurately classify startup outcomes.

4.2.2 Receiver operating characteristic curve and area under curve

The Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) are used to evaluate the true positive rate against the false positive rate (Raschka (2019)), *Fig. 10*. A higher AUC indicates superior model performance.

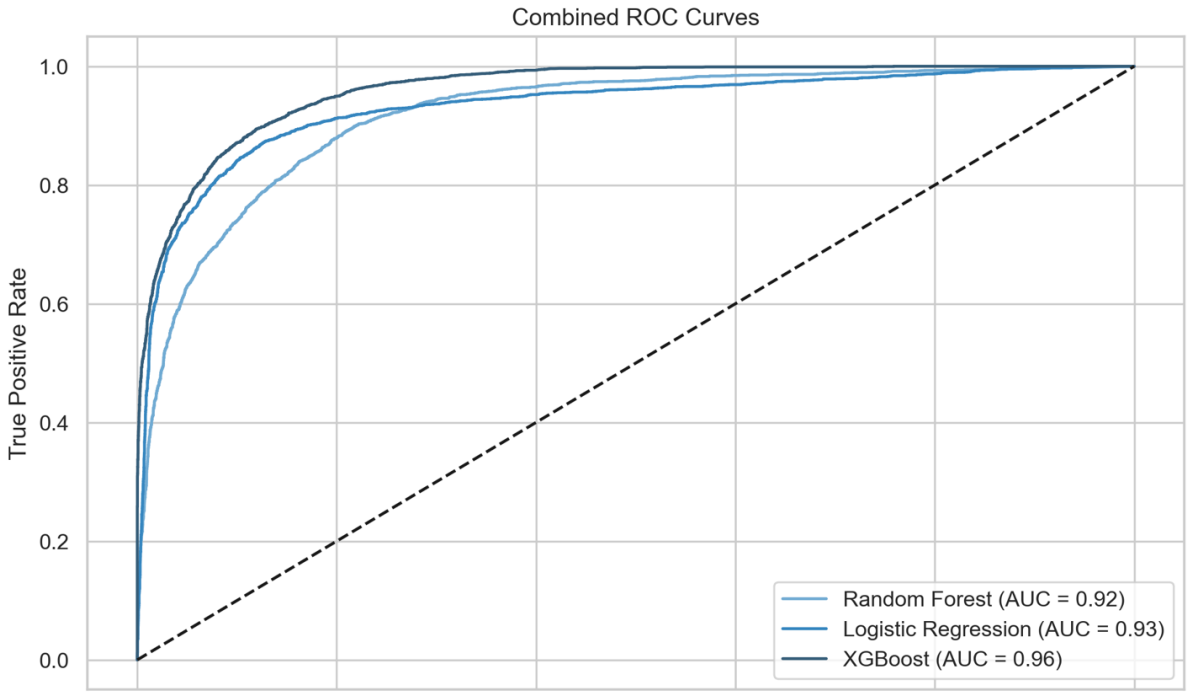

Figure 10 ROC curves

- Random Forest Classifier (Light blue line, AUC = 0.92): The ROC curve of the Random Forest model, represented by the blue line, exhibits a robust AUC of 0.92. This score demonstrates the model's ability to effectively differentiate between successful and unsuccessful startup classes. The curve closely follows the left-hand and top borders of the ROC space, indicating a high true positive rate and a low false positive rate.
- Logistic regression (Blue line, AUC = 0.93): The Logistic Regression model, represented by the blue line, presents a slightly better AUC score of 0.93 compared to Random Forest. This suggests that the Logistic Regression model has a slightly higher true positive rate for various threshold settings, indicating a slightly better measure of separability between the classes.
- XGBoost (Dark blue line, AUC = 0.96): Our Extreme Gradient boosting algorithm achieved an AUC of 0.96, as indicated by the dark blue line, indicating superior performance compared to the other two models. Its curve is closest to the top-left corner, indicating a higher true positive rate for a given false positive rate.

4.2.3 Precision-Recall curves

Precision-Recall curves demonstrate the trade-off between precision and recall, which is crucial for understanding the balance each model strikes between sensitivity and specificity (Raschka

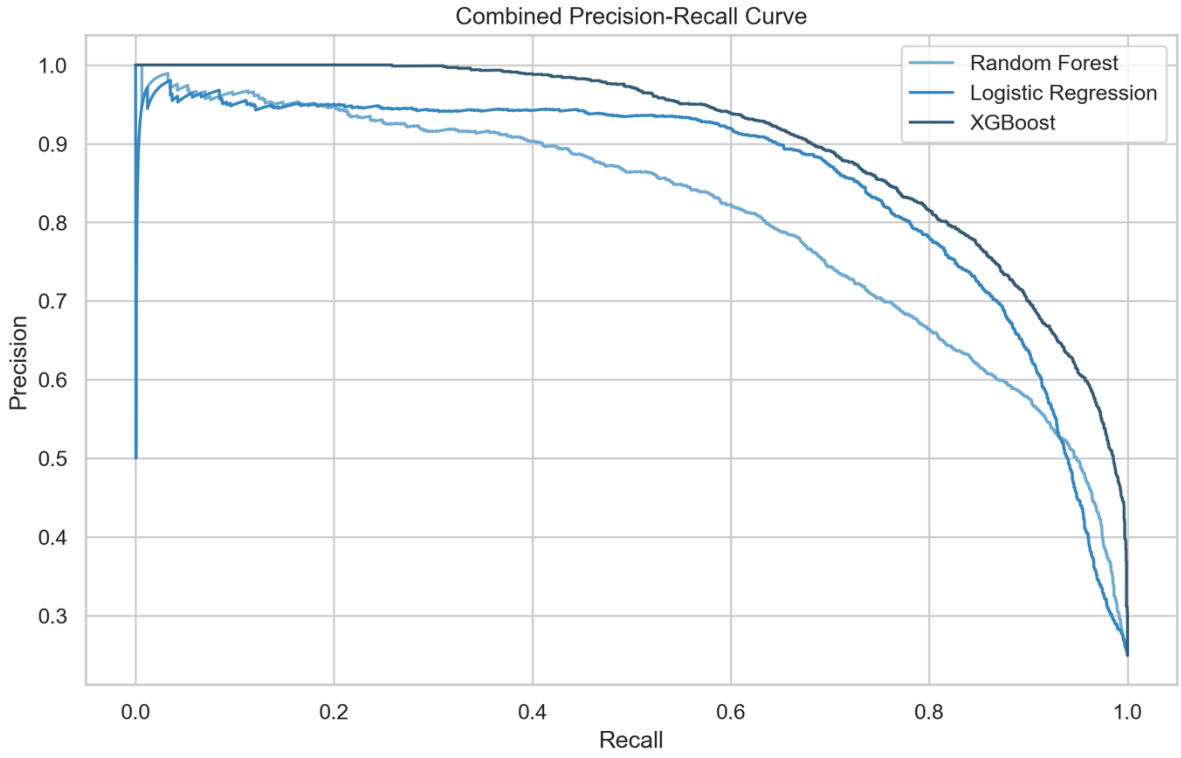

Figure 11 Precision-Recall Curves

When analyzing these curves, it is noteworthy that the XGBoost model (represented by the dark blue line) maintains a higher precision over a large range of recall values compared to the other models. This suggests that XGBoost is more capable of correctly identifying successful startups (precision) while also capturing a large proportion of the actual successful startups in the dataset (recall).

The logistic regression model (depicted by the blue line) starts with a relatively low precision compared to my other models for the same recall values. This may indicate a higher false positive rate initially, but the curve's gentler slope as recall increases suggests that its precision degrades at a slower rate than random forest classifier. This model demonstrates robustness across a balanced range of precision and recall values, making it a reliable choice when both metrics hold equal importance.

The Random Forest classifier (represented by the light blue line) also demonstrates strong performance, especially in the moderate recall range. Its curve begins higher on the precision axis, indicating that it initially identifies a high proportion of true positives out of all positive predictions. However, when recall increases, precision tends to decrease, which is a common behavior that indicates a trade-off between identifying as many successes as possible (recall) and maintaining accuracy in those identifications (precision).

4.3 Feature importance analysis

Feature importance in ML models provides insight into the relative significance of each feature in making predictions (Greewell (2018)). For tree-based models such as Random Forest and XGBoost, standard feature importance is derived from the model's construction. Specifically, it is derived from how much each feature contributes to splitting the data and reducing the model's entropy or Gini impurity. Essentially, the tree-building algorithms consider features that result in the most significant information gain as more important. This approach aids in comprehending which features have the greatest impact on the model's predictive outcomes (Zien (2009)).

For logistic regression, the approach is different due to the nature of the model. It estimates probabilities using the logistic function instead of splitting the data. The resulting coefficients can be interpreted as the change in the log-odds of the outcome for a one-unit change in the feature. To enhance comprehensibility and logical structure, coefficients are often transformed into odds ratios. These ratios indicate the factor by which the odds of success increase (for a ratio greater than 1) or decrease (for a ratio less than 1) with a one-unit increase in the feature. This approach

provides a more tangible interpretation of the model's coefficients in terms of risk or likelihood, which can be particularly useful when explaining model decisions in business terms (Aguilera, Escabias and Valderrama (2006)).

The odds ratio is a useful tool for indicating the effect size of each feature on the predicted outcome. This is particularly true for models like logistic regression, where the relationship between the feature and outcome is represented by coefficients in a linear fashion (Aguilera, Escabias and Valderrama (2006)).

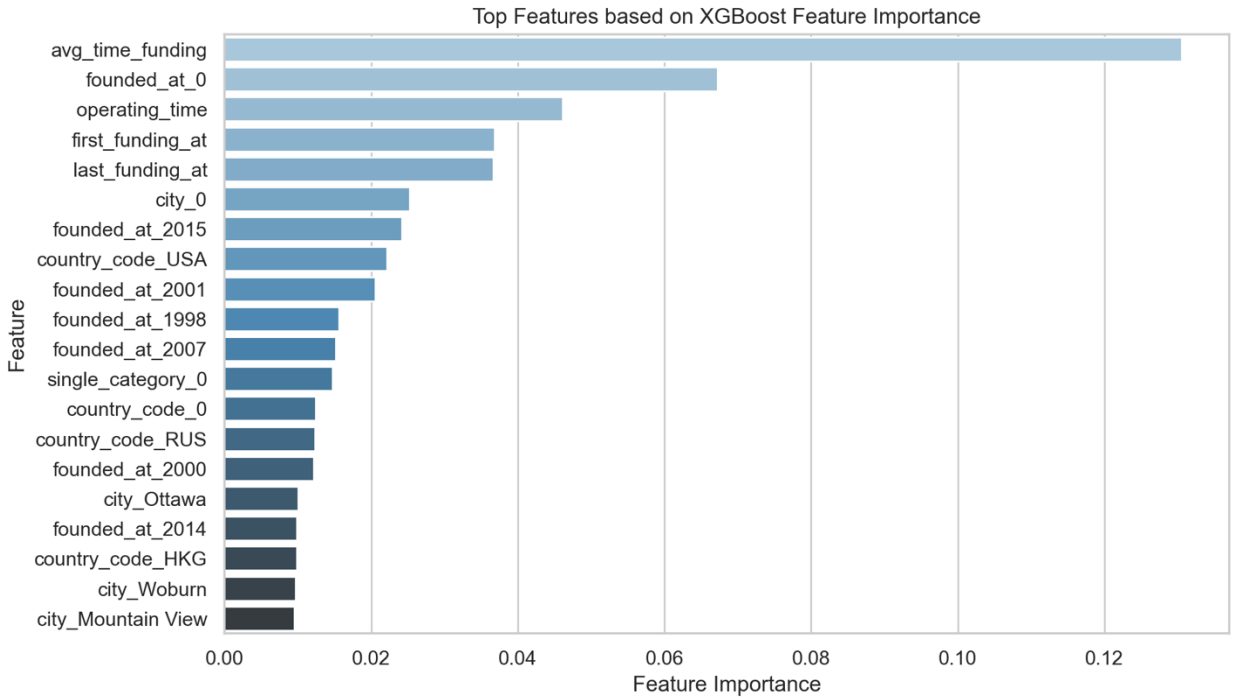

Figure 12 XGBoost - Feature importance

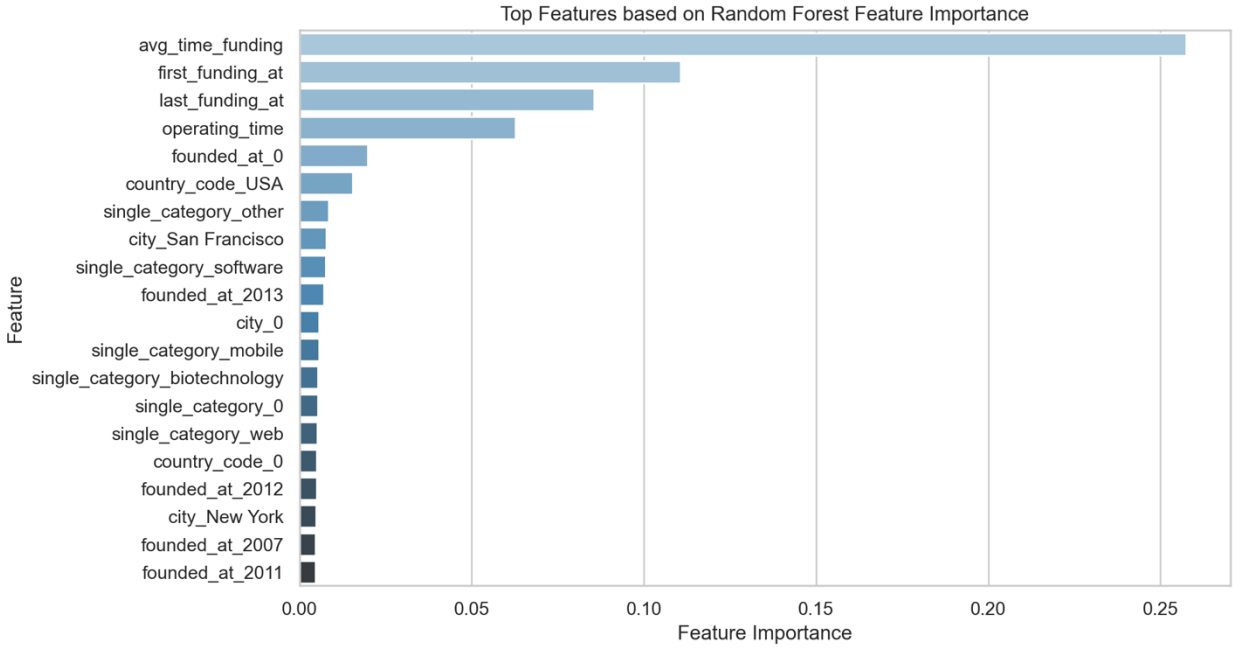

Figure 13 Random Forest Classifier - Feature importance

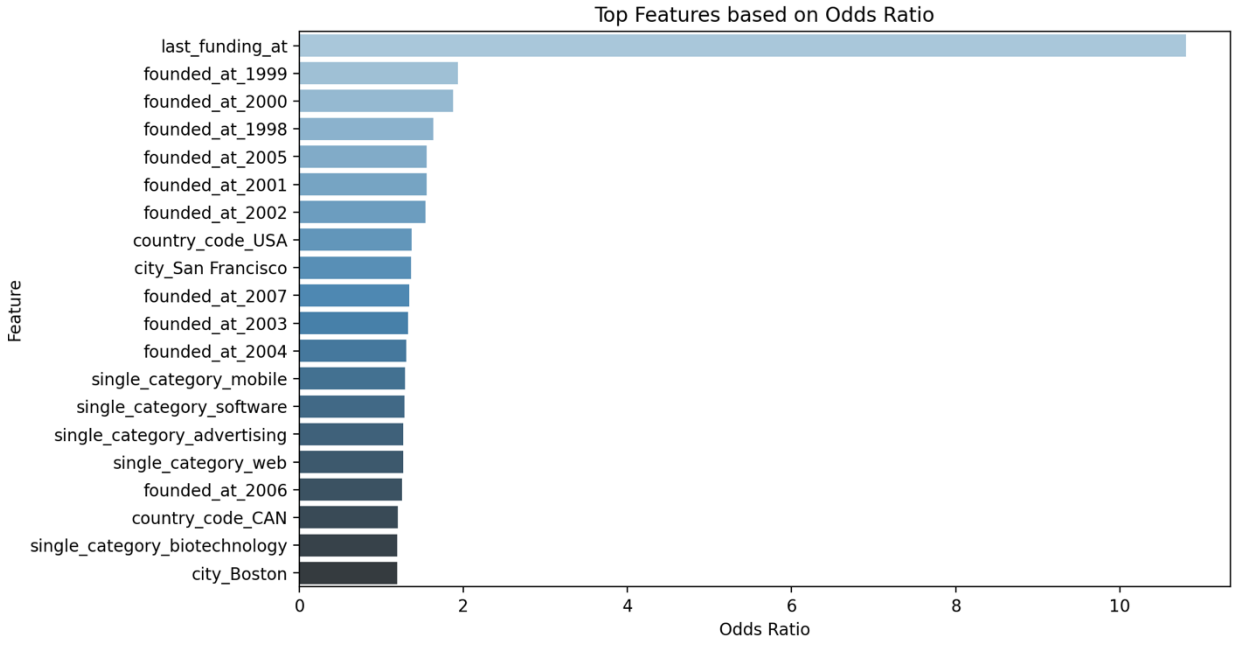

Figure 14 Logistic regression – Top features (Odds ratio)

When analyzing the importance of features in the Random Forest, XGBoost, and Logistic Regression models, a set of common predictors consistently emerges as significant *(Fig. 12, 13, 14)*. These predictors include avg_time_funding (average time elapsed between funding rounds), last_funding_at, and operating_time, highlighting the crucial role of funding timelines in determining startup viability. Additionally, geographic indicators such as country_code_USA and

city_San Francisco persistently appear, affirming the influence of location on startup success across all models.

Sector-specific categories, particularly single_category_software and single_category_biotechnology, are prominent, reflecting the high growth potential in these industries. The year of founding also features as a notable predictor, with various vintages such as founded_at_1998, founded_at_2000, and founded_at_2007 suggesting that temporal context carries weight in startup outcomes.

4.4 Interpretation of results

The interpretation of outcomes derived from ML models poses a distinct challenge, particularly when aiming to extract actionable insights for the VC industry. As emphasized by Goodman and Flaxman understanding the inner workings of ML predictors is paramount (Goodman and Flaxman (2017)). Without such understanding, it becomes difficult to effectively scrutinize and regulate these models to prevent the inadvertent encoding of discriminatory practices against minority groups. In addition, the inability to decipher the underlying mechanisms hamper our capacity to learn from instances where the model does errors. In this line, the interpretability is a critical feature within the analytical framework of ML generating informed recommendations crucial for decision-making processes. Numerous studies have been published to date on the topic of interpretability in relation to ML outputs. It is essential to maintain objectivity throughout such analyses, avoiding subjective evaluations, and opting instead for clear, objective, and value-neutral language. Finally, a delicate balance must be found out between interpretability and predictive performance of the ML model when assessing its efficacy. However, a continuous effort towards achieving the interpretability should not come at the expense of predictive accuracy, nor should it compromise the quality and integrity of the whole analytical process (Leblanc (2023)).

This section analyzes these findings, examining the effectiveness of each model in the context of evaluating startups and making investment decisions.

4.4.1 Effectiveness in predicting startup outcomes

The analysis of the ML models shows a high level of predictive ability, with the ROC-AUC curves confirming that all models are strong classifiers. The XGBoost model stands out as the best performer with an AUC score of 0.96, indicating its exceptional ability to distinguish between

successful and unsuccessful startups. The high accuracy rate of this model, which exceeds 90%, highlights its ability to manage the challenging class imbalance that is inherent in startup data (note: there will always be by far more unsuccessful companies than successful ones, the class imbalance is a very important concept when applying ML to startups)

XGBoost is further distinguished by its ability to maintain high precision while also achieving varying levels of recall, as demonstrated by the precision-recall curves. This balance is particularly advantageous in the VC sector, where false positives can have significant financial consequences. A model that can handle complex and non-linear relationships within the data can aid VCsts in screening potential investments, improving the quality and reliability of their decisions. The high AUC scores for all models indicate that the methodical approach to feature engineering and model training has been effective, resulting in models with robust predictive capabilities.

4.4.2 Practical application in venture capital

These models can serve as advanced tools for VCs, potentially streamlining the sourcing, screening, due diligence, or even selection processes. By leveraging these models, VCs could allocate their time and resources more efficiently, focusing on startups that are algorithmically vetted to have a higher likelihood of success.

Incorporating ML into VCst's decision-making processes has the potential to reduce reliance on intuition and heuristic-based judgments, which are susceptible to biases. These models' data-driven nature can complement VCst's traditional expertise, offering a more holistic approach to investment decisions (Leblanc (2023)).

4.5 Limitations of the study

The ML models' empirical analysis provides valuable insights. However, it is crucial to recognize the research's constraints and limitations to fully appreciate the context and applicability of the findings.

The reliance on historical data up to the year 2015 presents a significant limitation. The fast-paced and ever-evolving nature of the startup ecosystem means that patterns and trends can shift dramatically, rendering past data less predictive of future success. To improve predictive accuracy, it is recommended to utilize real-time data enhancement techniques like web scraping. This research's static dataset could not capture such dynamism, and although the performance metrics of the

trained models were solid, the statis nature of the dataset disregards the practical applicability of the empirical findings.

Moreover, the data used in this study was primarily focused on the business aspects of startups, without delving into the intricacies of the founding team's characteristics. This omission introduces a bias toward quantifiable business metrics, potentially overlooking the qualitative, human elements that play a critical role in a startup's success. The lack of data specific to the founders, such as their psychological profiles or team dynamics, highlights a significant gap. VCsts often heavily consider the team's caliber when making investment decisions, especially in the very early-stage companies.

One key question that arose during the empirical analysis is how to measure the effectiveness of algorithmic decision-making compared to human intuition in the VC domain. A theoretical analysis of ML's predictive power is insufficient. To truly validate the efficacy of an algorithmic approach, a longitudinal study is necessary. This study would compare the algorithm's investment selections to those of human VCs over an extended period, evaluating the returns generated by each.

The study's design does not allow for direct, real-world comparative analysis. Therefore, the conclusions drawn remain largely hypothetical. This research suggests the potential of ML as a tool for enhancing VC decisions. However, it does not provide empirical evidence of its superiority over traditional human-led investment strategies. Acknowledging these limitations is crucial, as they contextualize the findings within the scope of academic research rather than providing practical investment advice.

The research is limited by its reliance on historical data from 1998 to 2015. This dataset provides a substantial foundation, but it restricts the study to past patterns and trends, which may not be indicative of current or future startup successes. The dynamic and rapidly evolving nature of the startup ecosystem necessitates continual model updates and validations against emerging trends.

Another significant limitation is the absence of dynamic data enhancement, such as web scraping, which has been shown to substantially improve model performance (Sharchilev, Roizner, Rumyantsev, Ozornin, Serdyukov and de Rijke (2018)). The reliance on a static dataset from Kaggle, originally sourced from Crunchbase, further compounds this issue. Crunchbase data often includes information not available at the time of investment decisions, challenging the practical application of models trained on this data (Retterath (2020)).

The dataset's sole focus on business attributes introduces a bias, overlooking key aspects of a startup company, such as founder's profile or team dynamics. This missing data leads to negligence of the human element, which is critical in VC decision-making.

The interpretability of ML models poses a considerable challenge. Despite achieving accuracy and robust metrics, the underlying mechanisms of these models often remain opaque, making meaningful interpretations difficult. This "black box" nature of some ML models limits our understanding of how specific predictions are made (Israel (2020)).

A crucial limitation is the challenge of comparing algorithmic performance to that of human decision-makers. To make an effective comparison, a prolonged time window for training, testing, and practical application of the model would be required, likely spanning a decade or more. This extended timeframe complicates direct comparisons and raises questions about the practical feasibility of such an approach in real-world scenarios.

These limitations emphasize the importance of incorporating dynamic, real-time data sources and enhancing model transparency and interpretability. They also underscore the significance of integrating human-centric data to fully capture the complexity of the VC decision-making process.

4.6 Summary of the outcomes

The study's empirical findings are promising, but they also highlight the complexity of applying ML algorithms to the VC decision-making process. The predictive models, including Random Forest, XGBoost, and Logistic Regression, have shown considerable potential to identify patterns that may lead to successful venture outcomes through rigorous analysis of startup data. ML models have the ability to analyze large datasets and uncover insights that may be missed by humans. However, it is important to contextualize the conclusions drawn from this analysis within the scope of its methodological confines. The study relies on historical, unvalidated data from Crunchbase and omits qualitative founder-related factors. Furthermore, the algorithmic comparison to human decision-making is theoretical in nature, which limits its applicability.

5 DISCUSSION

This study presents a comprehensive exploration of the integration of ML into VC. The initial chapters are theoretical, while the latter ones present a study aimed at fitting an ML model for evaluating the predictive efficacy of different ML methods. The work not only examines the potential of ML in VC but also empirically tests, analyzes, and validates chosen models using real venture data, albeit limited. The XGBoost model shows superior performance, with the highest accuracy and recall rates for successful startups. Although empirical analysis is important, it is crucial to note that it should be used to offer new insights into how ML can enhance traditional VC decision-making processes. The presented study aimed to bridge the gap between theoretical understanding and practical application of ML methods in the field of VC.

The VC industry represents a complex interplay between quantitative data analysis and qualitative judgments aimed at maximizing profitability (Amornsiripanitch, Gompers and Xuan (2019); Chung, Liu and Smith (2023); Kaplan and Strömberg (2003)). However, this endeavor also carries significant financial risks for investors, necessitating a strategy that concurrently enhances the likelihood of success while mitigating risks (Gompers, Gornall, Kaplan and Strebulaev (2021)). Various systems have been proposed for this purpose, but none are entirely unfailing in practice. Importantly, human intuition and experience are indispensable in this field (Colombo and Grilli (2010)). Therefore, the implementation of ML should not aim to replace human expertise but rather to augment it. The integration of computational methods into VC practices is driven by the remarkable advancements in AI (Kaushik (2024)). Motivations for making better decisions then encompass improving profits and supporting the development of new products and services.

Random Forest, XGBoost, and logistic regression models were used to analyze empirical data sourced from public datasets. These models have previously shown promising outcomes in multiple fields of human activities, including VC practice (Bai and Zhao (2021); Dezhkam and Manzuri (2023); Hu, Hu, Hu and He (2022); Sruthi (2024); Thirupathi, Alhanai and Ghassemi (2021); Zbikowski and Antosiuk (2021); Zhang, Yuan, Mahmoudi, Ji and Fang (2023)). In this study, their application was extended to big data, and their performance in predicting startup success was compared against known final outcomes. Consistent with prior research, each model provided unique insights into predictive capabilities within the VC context. However, there were also notable differences in outcomes that require further investigation into their underlying reasons. These differences may stem from the mathematical foundations of each method.

The Random Forest algorithm functions as an ensemble of decision-making teams in ML (Hu, Hu, Hu and He (2022); Sruthi (2024)). It aggregates the outputs of numerous individual models or trees to enhance prediction accuracy. Random Forest algorithms construct a robust and precise overall model by leveraging both classification and regression tasks simultaneously (Sruthi (2024)). They can effectively handle both continuous (in regression) and categorical (in classification) variables. However, to effectively utilize these methods, a profound understanding of the impact of various hyperparameters in Random Forest is necessary (Ahmed, Raza, Hussain, Aldweesh, Omar, Khan, Eldin and Nadim (2023)). Numerous studies have explored the application of Random Forest modeling in VC contexts, with many highlighting its advantages in estimating startup success (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019); Hu, Hu, Hu and He (2022); Li (2020); Ross (2021)). This analysis showed that Random Forest had slightly lower performance compared to regression models and XGBoost. However, the differences were not significant enough to exclude Random Forest from the decision-support methods suitable for VC. Furthermore, emerging trends involve exploring hybrid ML systems that combine the strengths of multiple ML methods to estimate the potential success of various startups (Ali (2019)).

ML regression models have become popular in VC due to their exceptional performance in predictive analytics, which enables the forecasting of trends and outcomes (Aguilera, Escabias and Valderrama (2006); Aygüneş (2019); Zbikowski and Antosiuk (2021)). These models analyze the relationship between the target variable (dependent variable) and a set of independent variables. Like all ML methods, regression models require appropriate training to comprehend the relationship between the dataset's independent variables and the target outcome. Various types of regression are used in machine learning, including linear, logistic, ridge, lasso, polynomial, and Bayesian linear regression (Maulud (2020)). The choice of regression depends on the nature of the data and the specific task at hand. In this study, we confirm the effectiveness of logistic regression in estimating startup success or failure, ranking it as the second most effective after XGBoost. Other researchers have also investigated the potential of regression models to enhance VC decision-making (Aygüneş (2019); Zbikowski and Antosiuk (2021)). They have primarily highlighted the advantages of regression models in forecasting the future success of evaluated startups. In addition, further research is needed to explore the potential benefit of other types of regression ML in improving prediction accuracy (Ali (2019); Dellermann (2017); Dellermann, Lipusch, Ebel and Leimeister (2019); Hu, Hu, Hu and He (2022)).

XGBoost is a machine learning algorithm that uses gradient-boosted decision trees to improve data understanding and decision-making (Thirupathi, Alhanai and Ghassemi (2021); Zbikowski and Antosiuk (2021); Zhang, Yuan, Mahmoudi, Ji and Fang (2023)). It has high accuracy and ROC-

AUC for predictions, and its rapid computation speed saves time. Data scientists frequently use XGBoost to optimize other machine learning models (Meng, Yang, Qian and Zhang (2021); Wang, Kuo and Chen (2022)). XGBoost has the potential to be advantageous in the VC realm especially due to its ability to handle tasks associated with regression, classification, and ranking simultaneously. Despite that several studies have explored the contributions of XGBoost to improving precision accuracy in various fields (Matloob, Aftab, Ahmad, Khan, Fatima, Iqbal, Alruwaili and Elmitwally (2021); Niazkar, Menapace, Brentan, Piraei, Jimenez, Dhawan and Righetti (2024)), there is limited evidence regarding its support for VC decision-making. Although our study has acknowledged limitations, it found that XGBoost outperformed Random Forest and logistic regression in evaluating a large number of startups. The theoretical potential of XGBoost ML methods is evident, suggesting that it could offer valuable insights for VC practices. However, further extensive research is needed to solidify evidence of its benefits in the realm of economics. Additionally, there is potential to investigate hybrid machine learning techniques, such as the fusion of neural networks and XGBoost, to reveal novel pathways for analysis (Ali (2019); Dellermann, Lipusch, Ebel and Leimeister (2019); Wang, Kuo and Chen (2022)).

Consistent with previous research, this study provides an empirical analysis of how ML can improve conventional VC decision-making processes. ML tools provide VC practitioners with valuable insights and decision support by analyzing available data with unparalleled speed and accuracy. The potential for ML in VC is evident. Recently, a startup investment decision support application employing VC scorecards based on ML approaches has been proposed and tested (Bai and Zhao (2021)). Other methods have been proposed to improve the accuracy of event occurrence (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019); Dezhkam and Manzuri (2023); Tuli (2023)). However, addressing inherent limitations is necessary for the full realization of this concept. This presents both a challenge and an opportunity for future research. It is important to acknowledge that chance still plays a role in outcomes, despite the robustness and speed of new ML methods (Kläs and Vollmer (2018)).

6 CONCLUSION

This concluding chapter summarizes the insights and findings from the preceding chapters of this thesis. Chapter 2 provides an overview of VC theory and literature, exploring the dynamics of VC and introducing ML. It investigates the potential synergy between ML and VC, particularly in the context of enhancing VC decision-making, and reviews existing research in the field. Chapter 3 outlines the methodology used for data collection and model building, which sets the stage for empirical analysis. Chapter 4 presents the results of applying ML models to VC data, providing a practical perspective on the integration of ML in VC. The discussion aims to synthesize the discussed elements, examine the strengths and limitations of integrating ML with traditional VC practices, and suggest avenues for future research.

The study's integration of ML into VC highlights advancements for the industry. The empirical results indicate that models such as XGBoost can be effectively used to make data-driven decisions in VC. This suggests a potential shift from traditional, intuition-based strategies to more analytical approaches. The models demonstrated an improved ability to predict the success of a startup, which could enhance the accuracy and efficiency of VC investments by adopting these ML techniques.

This research contributes to the intersection of ML and VC. It validates the practicality of ML models in real-world VC settings and provides a robust framework for future research. The study emphasizes the significance of feature selection and model tuning in predicting startup success, offering valuable insights for further exploration in this field.

This study has practical implications for venture capitalists as it demonstrates the feasibility of applying machine learning to qualitative VC data. This paves the way for a broader shift towards more intelligent, data-informed VC processes, enhancing the accuracy of venture funding decisions and supporting the allocation of capital to potentially successful startups. Improved funding strategies can have far-reaching effects on the economy and society. Appropriately funded startups that are more likely to succeed can contribute to job creation and generate positive societal impacts. This research has the potential to transform VC practices, resulting in more efficient and impactful investment landscapes.

In conclusion, these findings address a crucial gap in both practical and academic realms. They provide venture capitalists with innovative tools for investment decision-making and establish a precedent for academic research into the synergy between machine learning and venture capital.

7 FUTURE DEVELOPMENTS

The exploration conducted in this thesis uncovers several promising avenues for future research that merge the intricate realms of VC and ML.

One potential area of advancement lies in the refinement of ML models specifically tailored to VC tasks (Shi, Eremina and Long (2024)). This could include the development of novel algorithms, ensemble methods, or other deep learning techniques to improve predictive accuracy and decision-making in VC investments. Tailoring ML models to specific VC challenges and opportunities is critical in this endeavor. In addition, the exploration of hybrid ML methods and decision-making tools holds great promise and warrants further investigation (Ali (2019)).

Currently, there is a lack of comparative analyses of different VC models. Future studies could examine the actual returns generated by human led VC funds, augmented VCs, and algorithmic VCs, providing invaluable insights into their effectiveness and efficiency, particularly in terms of investment returns and decision-making processes (Blohm, Antretter, Sirén, Grichnik and Wincent (2022); Dellermann, Calma, Lipusch, Weber, Weigel and Ebel (2019)).

Another promising area of research is to examine the impact of ML at different stages of the VC process (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019); Park and Tzabbar (2016); Torres (2022)). This includes analyzing its role in deal sourcing, screening, selection, and post-investment advisory to elucidate how ML influences the effectiveness of each stage.

Furthermore, it is crucial to examine the reception of algorithmically selected startups among founders. This includes exploring whether founders prefer VCst that offer more than just capital, such as strategic advice and networking opportunities, and identifying any biases toward or against algorithm driven VCst (Amornsiripanitch, Gompers and Xuan (2019); Harris, Jenkinson, Kaplan and Stucke (2023); Retterath (2020)).

In addition, future research could focus on feature engineering and data fusion to improve the quality and relevance of input data for ML models in the VC space (Dong (2018)).

The interpretability of ML models is paramount for VC practitioners to better understand model predictions, including identifying influential variables and mitigating biases and discrimination associated with investment processes (Gilpin, Bau, Yuan, Bajwa, Specter and Kagal (2018); Leblanc (2023); Ross (2021)).

In addition, there is an urgent need for research in risk management and uncertainty quantification, which should be incorporated into ML-based VC models to improve their effectiveness (Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019); Kläs and Vollmer (2018)).

Exploring the use of synthetic data in AI represents a new avenue for predictive modeling in VC (Lee (2024); Lu (2024)). Future research could explore how AI-generated synthetic data can better predict market trends and cycles, especially in the dynamic and unpredictable realm of startups.

Ethical considerations are also critical, not only for VC domains, but for responsible AI development as a whole (Glücksman (2020); Liu (2017); Yang, Fang, Wang and Su (2023)). Research in this area should focus on the ethical aspects of VC decision-making, transparency, fairness, accountability, or privacy considerations.

Finally, longitudinal studies are essential to assess the long-term viability and impact of ML-based VC investments on profitability, innovation, and socio-economic outcomes (Petty and Gruber (2011); Petty, Gruber and Harhoff (2023)). Such studies could provide valuable insights into the effectiveness and sustainability of ML-based VC.

By prioritizing research in these areas, the intersection of ML and VC can continue to evolve and transform investment practices, fostering more informed, efficient, and impactful decision making in the VC space.

8 REFERENCES

Aguilera, A. M., M. Escabias, and M. J. Valderrama, 2006, Using principal components for estimating logistic regression with high-dimensional multicollinear data, *Computational Statistics & Data Analysis* 50, 1905-1924.

Ahmed, S., B. Raza, L. Hussain, A. Aldweesh, A. Omar, M. S. Khan, E. T. Eldin, and M. A. Nadim, 2023, The deep learning resnet101 and ensemble xgboost algorithm with hyperparameters optimization accurately predict the lung cancer, *Applied Artificial Intelligence* 37.

Ali, H.S., Morteza B, Siamak N, Saman MP, 2019, Developing a hybrid intelligent system for optimizing syndicated venture capital portfolios, *Journal of intelligent and fuzzy systems* 6483-6497.

Amornsiripanitch, N., P. A. Gompers, and Y. H. Xuan, 2019, More than money: Venture capitalists on boards, *Journal of Law Economics & Organization* 35, 513-543.

Ang, Y.Q., Chia A, Saghafian S, 2022, Using machine learning to demystify startups funding, post-money valuation, and success, in V. Babich, Birge JR, Hillary G, ed.: Innovative Technology at the Interface of Finance and Operations (Springer).

Antretter, T., Blohm I, Siren Ch, Grichnik D, Malmstrom M, Wincent J, 2020, Do algorithms make better - and fairer - investments than angel investors?, (Harvard Business Review, Harvard Business Review).

Antretter, T., C. Siren, D. Grichnik, and J. Wincent, 2020, Should business angels diversify their investment portfolios to achieve higher performance? The role of knowledge access through co-investment networks, *Journal of Business Venturing* 35.

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

Aygüneş, G., 2019, Modelling the venture capital investments via logistic regression method, International conference of numerical analysis and applied mathematics ICNAAM (AIP Publishing, Rhodes, Greece).

Bai, S., and Y. J. Zhao, 2021, Startup investment decision support: Application of venture capital scorecards using machine learning approaches, *Systems* 9.

Blohm, I., T. Antretter, C. Sirén, D. Grichnik, and J. Wincent, 2022, It's a peoples game, isn't it?! A comparison between the investment returns of business angels and machine learning algorithms, *Entrepreneurship Theory and Practice* 46, 1054-1091.

Budach, L., Feuerpfeil, M, Ihde, N, Nathansen, A, Noack, N, Patzlaff, H, Naumann, F, Harmouch, H, 2022, The effects of data quality on machine learning performance, *arXiv:2207.14529*.

Cao, L., von Ehrenheim V, Krakowski S, Li, X, Lutz, A, 2023, Using deep learning to find the next unicorn: A practical synthesis on optimization target, feature selection, data split and evaluation strategy, *IJCAI-2023 Workshop of Multimodal AI For Financial Forecasting (MuFFin)* 11.

Caruso, C., Enriquez F, Oshotse A, Pradeep G, 2015, Algorithmic venture capital: Predicting valuation step-up multiple in venture backed companies through deep learning techniques, (Stanford University, Stanford University).

Colombo, M. G., and L. Grilli, 2010, On growth drivers of high-tech start-ups: Exploring the role of founders' human capital and venture capital, *Journal of Business Venturing* 25, 610-626.

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

Crawford, G. C., H. Aguinis, B. Lichtenstein, P. Davidsson, and B. McKelvey, 2015, Power law distributions in entrepreneurship: Implications for theory and research, *Journal of Business Venturing* 30, 696-713.

Davidsson, P., 2008. *The entrepreneurship research challenge* (Edward Elgar Publishing Limited, Cheltenham, UK).

Davila, A., G. Foster, and M. Gupta, 2003, Venture capital financing and the growth of startup firms, *Journal of Business Venturing* 18, 689-708.

Dellermann, D., A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, 2019, The future of human-ai collaboration: A taxonomy of design knowledge for hybrid intelligence systems, *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences* 274-283.

Dellermann, D., Lipusch N, Ebel P, Popp KM, Leimeister JM, 2017, Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method, International Conference on Information Systems (ICIS) (Seoul, South Korea.).

Dellermann, D., N. Lipusch, P. Ebel, and J. M. Leimeister, 2019, Design principles for a hybrid intelligence decision support system for business model validation, *Electronic Markets* 29, 423-441.

Desarbo, W., I. C. Macmillan, and D. L. Day, 1987, Criteria for corporate venturing - importance assigned by managers, *Journal of Business Venturing* 2, 329-350.

Dezhkam, A., and M. T. Manzuri, 2023, Forecasting stock market for an efficient portfolio by combining xgboost and hilbert-huang transform, *Engineering Applications of Artificial Intelligence* 118.

Dong, G., Liu H, 2018. *Feature engineering for machine learning and data analytics* (CRC Press; Taylor & Francis group; A Chapman & Hall Book).

Dorogush, A.V., Ershov V, Gulin A, 2018, Catboost: Gradient boosting with categorical features support, *arXiv:1810.11363*.

Emerson, S, Kennedy R, O'Shea L, O'Brien J, 2019, Trends and applications of machine learning in quantitative finance, 8th International Conference on Economics and Finance Research (ICEFR 2019) (Lyon, France).

Ferrati, F., and M. Muffatto, 2021, Entrepreneurial finance: Emerging approaches using machine learning and big data, *Foundations and Trends in Entrepreneurship* 17, 232-329.

Ferrati, F., and M. Muffatto, 2021, Reviewing equity investors' funding criteria: A comprehensive classification and research agenda, *Venture Capital* 23, 157-178.

Fried, V. H., and R. D. Hisrich, 1994, Toward a model of venture capital-investment decision-making, *Financial Management* 23, 28-37.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, 2018, Explaining explanations: An overview of interpretability of machine learning, *2018 Ieee 5th International Conference on Data Science and Advanced Analytics (Dsaa)* 80-89.

Glücksman, S., 2020, Entrepreneurial experiences from venture capital funding: Exploring two-sided information asymmetry, *Venture Capital* 22, 331-354.

Gompers, P. A., W. Gornall, S. N. Kaplan, and I. A. Strebulaev, 2020, How do venture capitalists make decisions?, *Journal of Financial Economics* 135, 169-190.

Gompers, P., W. Gornall, S. N. Kaplan, and I. A. Strebulaev, 2021, How venture capitalists make decisions an inside look at an opaque process, *Harvard Business Review* 99, 70-+.

Goodman, B., and S. Flaxman, 2017, European union regulations on algorithmic decision making and a "right to explanation", *Ai Magazine* 38, 50-57.

Greenwood, J., P. F. Han, and J. M. Sánchez, 2022, Venture capital: A catalyst for innovation and growth, *Federal Reserve Bank of St Louis Review* 104, 120-130.

Greewell, B.M., Boehmke BC, McCarthy AJ, 2018, A simple and effective model-based variable importance measure., *CoRR abs* 1805.04755.

Gupta, N., 2023, Optimising data quality of a data warehouse using data purgation process, *International Journal of Data Mining Modelling and Management* 15, 102-131.

Gupta, N., S. Mujumdar, H. Patel, S. Masuda, N. Panwar, S. Bandyopadhyay, S. Mehta, S. Guttula, S. Afzal, R. S. Mittal, and V. Munigala, 2021, Data quality for machine learning tasks, *Kdd '21: Proceedings of the 27th Acm Sigkdd Conference on Knowledge Discovery & Data Mining* 4040-4041.

Harris, R. S., T. Jenkinson, S. N. Kaplan, and R. Stucke, 2023, Has persistence persisted in private equity? Evidence from buyout and venture capital funds, *Journal of Corporate Finance* 81.

Hsu, D. K., J. M. Haynie, S. A. Simmons, and A. McKelvie, 2014, What matters, matters differently: A conjoint analysis of the decision policies of angel and venture capital investors, *Venture Capital* 16, 1-25.

Hu, J. T., L. Y. Hu, M. Z. Hu, and Q. Z. He, 2022, Machine learning-based investigation on the impact of chinese venture capital institutions' performance: Evaluation factors of venture enterprises to venture capital institutions, *Systems* 10.

Chen, C. W., 2023, A feasibility discussion: Is ml suitable for predicting sustainable patterns in consumer product preferences?, *Sustainability* 15.

Chen, T. Q., and C. Guestrin, 2016, Xgboost: A scalable tree boosting system, *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 785-794.

Chung, Y. P., Y. Liu, and R. Smith, 2023, Venture capital exits and investments: The influences of market run-up, market timing, and media attention, *Quarterly Journal of Finance* 13.

Ionescu, S. A., and V. Diaconita, 2023, Transforming financial decision-making: The interplay of ai, cloud computing and advanced data management technologies, *International Journal of Computers Communications & Control* 18.

Israel, R., Kelly, BT, Moskowitz, TJ, 2020, Can machines 'learn' finance?, *Journal of Investment Management* 21.

Jaimovich, E., 2010, Adverse selection and entrepreneurship in a model of development, *Scandinavian Journal of Economics* 112, 77-100.

Jain, A., H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. S. Mittal, and V. Munigala, 2020, Overview and importance of data quality for machine

learning tasks, *Kdd '20: Proceedings of the 26th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining* 3561-3562.

Jordan, M. I., and T. M. Mitchell, 2015, Machine learning: Trends, perspectives, and prospects, *Science* 349, 255-260.

Kaplan, S. N., and J. Lerner, 2010, It ain't broke: The past, present, and future of venture capital, *Journal of Applied Corporate Finance* 22, 36-+.

Kaplan, S. N., and P. Strömberg, 2001, Venture capitalists as principals:: Contracting, screening, and monitoring, *American Economic Review* 91, 426-430.

Kaplan, S. N., and P. Strömberg, 2003, Financial contracting theory meets the real world:: An empirical analysis of venture capital contracts, *Review of Economic Studies* 70, 281-315.

Kaushik, P., 2024, The role of ai in venture capital: Evolution or replacement?, Techopedia.

Kläs, M., and A. M. Vollmer, 2018, Uncertainty in machine learning applications: A practice-driven classification of uncertainty, *Computer Safety, Reliability, and Security, Safecomp 2018* 11094, 431-438.

Köppen, M., D. H. Wolpert, and W. G. Macready, 2001, Remarks on a recent paper on the "no free lunch" theorems, *Ieee Transactions on Evolutionary Computation* 5, 295-296.

Leblanc, B., Germain P, 2023, Interpretability in machine learning: On the interplay with explainability, predictive performances and models, *arXiv:2311.11491v1*.

Lee, P. , 2024, Synthetic data and the future of ai, *110 Cornell Law Review; SSRN Electronic Journal* 52.

Li, J., 2020, Prediction of the success of startup companies based on support vector machine and random forset, WAIE 2020: 2nd International Workshop on Artificial Intelligence and Education.

Liu, H., and M. Cocea, 2017, Semi-random partitioning of data into training and test sets in granular computing context, *Granular Computing* 2, 357-386.

Liu, L., 2017, Research on the tripartite cooperation of the venture capital institutions, business incubators and start-ups and ethical risk from the perspective of game theory, *Proceedings of the 2017 2nd International Seminar on Education Innovation and Economic Management (Seiem 2017)* 156, 419-422.

Lu, Y., Shen M, Wang H, Wang X, van Rechem C, Wei W, 2024, Machine learning for synthetic data generation: A review, *arXiv:2302.04062* 1-18.

Matloob, F., S. Aftab, M. Ahmad, M. A. Khan, A. Fatima, M. Iqbal, W. M. Alruwaili, and N. S. Elmitwally, 2021, Software defect prediction using supervised machine learning techniques: A systematic literature review, *Intelligent Automation and Soft Computing* 29, 403-421.

Maulud, D., Abulazeez AM, 2020, A review on linear regression comprehensive in machine learning, *JASTT [Internet]* I, 140-47.

Meng, Y., N. H. Yang, Z. L. Qian, and G. Y. Zhang, 2021, What makes an online review more helpful: An interpretation framework using xgboost and shap values, *Journal of Theoretical and Applied Electronic Commerce Research* 16, 466-490.

Mitchell, T., 1997. *Machine learning* (McGraw Hill).

Nagar, Y., Malone TW, 2011, Making business predictions by combining human and machine intelligence in prediction markets, *Proceedings of the Thirty Second International Conference on Information Systems*.

Niazkar, M., A. Menapace, B. Brentan, R. Piraei, D. Jimenez, P. Dhawan, and M. Righetti, 2024, Applications of xgboost in water resources engineering: A systematic literature review (dec 2018-may 2023), *Environmental Modelling & Software* 174.

Park, H. D., and D. Tzabbar, 2016, Venture capital, ceos' sources of power, and innovation novelty at different life stages of a new venture, *Organization Science* 27, 336-353.

Paruchuri, H., 2019, Market segmentation, targeting, and positioning using machine learning, *Asian Journal of Applied Science and Engineering* 7-14.

Petty, J. S., and M. Gruber, 2011, "In pursuit of the real deal" a longitudinal study of vc decision making, *Journal of Business Venturing* 26, 172-188.

Petty, J. S., M. Gruber, and D. Harhoff, 2023, Maneuvering the odds: The dynamics of venture capital decision-making, *Strategic Entrepreneurship Journal* 17, 239-265.

Potanin, M., Chertok A, Zorin K, Shtabtsovsky C, 2023, Startup success prediction and vc portfolio simulation using crunchbase data, arXiv:2309.15552v1.

Raschka, S., Mirjalili, V., 2019, Python machine learning: Machine learning and deep learning with python, scikit-learn, and tensorflow 2. Expert insight., (Packt Publishing).

Retterath, A., 2020, Human versus computer: Benchmarking venture capitalists and machine learning algorithms for investment screening, *SSRN Electronic Journal*.

Retterath, A., 2022, Data-driven-vc, Data-Driven-VC.

Retterath, A., Braun R, 2020, Benchmarking venture capital databases, *SSRN Electronic Journal* 47.

Ross, G., Das SR, Scirco D, Raza H, 2021, Capitalvx: A machine learning model for startup selection and exit prediction, *SSRN Electronic Journal*.

Sandberg, W. R., and C. W. Hofer, 1987, Improving new venture performance - the role of strategy, industry structure, and the entrepreneur, *Journal of Business Venturing* 2, 5-28.

Sessions, V., Valtorta, MG, 2006, The effects of data quality on machine learning algorithms, MIT International Conference on Information Quality.

Sharchilev, B., M. Roizner, A. Rumyantsev, D. Ozornin, P. Serdyukov, and M. de Rijke, 2018, Web-based startup success prediction, *Cikm'18: Proceedings of the 27th Acm International Conference on Information and Knowledge Management* 2283-2291.

Shi, Y., E. Eremina, and W. Long, 2024, Machine learning models for early-stage investment decision making in startups, *Managerial and Decision Economics* 45, 1259-1279.

Sommer, S. C., C. H. Loch, and J. Dong, 2009, Managing complexity and unforeseeable uncertainty in startup companies: An empirical study, *Organization Science* 20, 118-133.

Sruthi, E.R., 2024, Understand random forest algorithms with examples (updated 2024), (Analytics Vidhya).

Stahl, R.H.A., 2021, Leveraging time-series signals for multi-stage startup success prediction, Department of Management, Technology, and Economics (ETH Zürich, Zürich).

Thirupathi, A. N., T. Alhanai, and M. M. Ghassemi, 2021, A machine learning approach to detect early signs of startup success, *Icaif 2021: The Second Acm International Conference on Ai in Finance*.

Torres, J. P., 2022, The venture capital feedback cycle: A critical review and future directions, *Entrepreneurship Research Journal* 12.

Tuli, M.S., Shreyas G, Dheemanth AN, Chand SR, Anupama YK, 2023, Smart start-up analyzer: Prediction model, analysis tool for venture capitals using machine learning, *International Journal of Scientific Research in Engineering and Management* 7.

Tyebjee, T. T., and A. V. Bruno, 1984, A model of venture capitalist investment activity, *Management Science* 30, 1051-1066.

Wang, C. C., P. H. Kuo, and G. Y. Chen, 2022, Machine learning prediction of turning precision using optimized xgboost model, *Applied Sciences-Basel* 12.

Wolpert, D. H., 2002, The supervised learning no-free-lunch theorems, *Soft Computing and Industry* 25-42.

Wolpert, D. H., 2023, The implications of the no-free-lunch theorems for meta-induction, *Journal for General Philosophy of Science* 54, 421-432.

Yacouby, R., Axman D, 2020, Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models, in S. Eger, Gao Y, Peyrard M, Zhao W, Hovy E, ed.: Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Association for Computational Linguistics).

Yang, Y. X., Y. L. Fang, N. X. Wang, and X. Su, 2023, Mitigating information asymmetry to acquire venture capital financing for digital startups in china: The role of weak and strong signals, *Information Systems Journal* 33, 1312-1342.

Zacharakis, A. L., and G. D. Meyer, 1998, A lack of insight: Do venture capitalists really understand their own decision process?, *Journal of Business Venturing* 13, 57-76.

Zbikowski, K., and P. Antosiuk, 2021, A machine learning, bias-free approach for predicting business success using crunchbase data, *Information Processing & Management* 58.

Zhang, J. S., J. F. Yuan, A. Mahmoudi, W. Y. Ji, and Q. S. Fang, 2023, A data-driven framework for conceptual cost estimation of infrastructure projects using xgboost and bayesian optimization, *Journal of Asian Architecture and Building Engineering*.

Zien, A., Krämer N, Sonnenburg S, Rätsch G, 2009, The feature importance ranking measure, *ECML/PKDD* 2, 694-709.