Norwegian School of Economics

Bergen, Spring, 2023

Founder Success in Norwegian Startups: A Machine Learning Approach

A study on the use of machine learning and personality traits to predict startup performance from a pre-seed perspective

Alexander Hogstad Wik & Håkon Otterlei

Supervisor: Sondre Nedreås Hølleland

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible − through the approval of this thesis − for the theories and methods used, or results and conclusions drawn in this work.

Abstract

This thesis aims to investigate founder characteristics in the Norwegian startup ecosystem and if machine learning can help venture capital firm identity successful founders at a startup's earliest stages, when information is greatly limited. The authors collected and refined data from multiple sources, resulting in a unique dataset of 1918 tech-driven, scalable startups and 2700 unique founders. Especially outstanding in the dataset is the inclusion of personality traits estimated though the use of artificial intelligence.

Four supervised machine learning models were employed to classify the founders into two created success categories, low success, and high success. The two tree-based methods, Extreme Gradient Boosting and Random Forest performed best considering the evaluation metrics, resulting in a classification accuracy of over 62%, while Logistic Regression and K-Nearest Neighbours did not follow far behind. The thesis finds significant evidence that the Number of Founders of a company and the personality trait Conscientiousness are strong predictors of success in the Norwegian startup landscape. Both of our findings showcase a positive correlation with startup performance, meaning entrepreneurs who inherits high Conscientiousness and are part of founding teams are more likely to succeed as entrepreneurs in Norway.

The research has two use cases. One, to narrow the research gap on founders in Norwegian startups, and two, motivate venture capital firms in Norway to adapt and implement machine learning models to help with decision-making, despite the challenges of limited data. The authors encourage others to continue research on this area, such as investigating the validity of personality traits obtained through artificial intelligence and broadening and expanding the research to other companies in Norway and other Scandinavian countries.

The thesis recognizes the potential ethical considerations that arise when collecting public data on private individuals. The weaknesses of this research are also discussed, which include the chosen data structure and biases in the data.

Acknowledgements

This thesis is the concluding achievement of our master's degree in Business and Administration, undertaken with a major in Business Analytics, at the Norwegian School of Economics.

We wish to express our heartfelt gratitude to our supervisor, Sondre Nedreås Hølleland, for always being available on short notice and his invaluable insights throughout the semester.

We also acknowledge the generosity of the Norwegian School of Economics in providing sponsorship for our access to Humantic AI and Forvalt. The invaluable data and resources derived from these platforms significantly enriched our research, contributing substantially to the depth and quality of this thesis.

Our appreciation extends to Kjetil Holmefjord, whose enlightening interviews shed invaluable light on the dynamics of the venture capital industry in Norway.

Lastly, we wish to extend our sincere thanks to our friends and family. Their constant support and encouragement through our study at the Norwegian School of Economics, and in the process of crafting this thesis, were a true source of strength and motivation.

Contents

FIGURES

TABLES

EQUATIONS

1 Introduction

In the constantly evolving business landscape, startups act as powerful engines of innovation and economic progress (Szarek & Piecuch, 2018).While startups are defined in many ways, we will adhere to the elegant definition by Forbes (2022) that startups are businesses that want to disrupt industries and change the world. Under this purview, startups are more than just newly established companies, and possess traits of often being highly scalable and technology driven. Motivated by their innovative founders, startups aspire to create unique products and services that people want, in the hope of success. However, the path is filled with obstacles, and despite their potential to bring significant change, the majority do not survive their first year (Albertsen et al., 2021).

Venture Capital (VC) firms play a vital role in this high-risk landscape, providing essential funding to promising startups. In addition, many of these firms share their extensive network and expert knowledge to assist startups during their challenging early stages (Skjelsbæk, 2023). VC firms make strategic investments in these promising businesses, with an aim of securing significant future returns. Given this, VC firms dedicate a substantial amount of time and effort in analysing and predicting the potential success of these startups.

Within these dynamics, the prominence of Artificial Intelligence (AI) in venture capital decision-making is experiencing rapid growth. According to the global research firm Gartner (2021), AI is predicted to be involved in more than 75% of venture capital investment decisions by 2025, a considerable increase from the modest 5% in 2021. This forecast illustrates the ongoing transformation the venture capital industry is currently navigating.

The independent Norwegian startup-focused newspaper Shifter recently published an article (Winther, 2023), underscoring the urgent need for Norwegian VC firms to embrace this shift. The article highlights the emerging divide between “old school” investors who rely predominantly on traditional methods such as network referrals and inbound deal flow, and “new school” investors who leverage big data, large language models, and AI to guide their decision-making process and enhance their understanding of what makes a business succeed in the future. With the race to first discover outstanding founders and exceptional investment opportunities, those who fail to board the AI train risk being left behind (Winther, 2023).

This needed shift towards AI integration is further supported by Bonelli (2022) who found that firms that adopt AI become better at identifying startups that survive and secure follow-on funding. However, Bonelli also found that these firms were less likely to invest in breakthrough companies as AI models that are trained on historical data can struggle to identify truly disruptive companies. This highlights the need for a balanced approach, where AI augments rather than replaced human judgement, aiding investors in making informed decisions and bridging the information gap between founders and investors.

From our engagements with VC firms in Norway, it is evident that a substantial number have not yet embarked on this vital transition, particularly those firms that specialize in early-stage investments (pre-seed). Despite the digital era we live in, obtaining comprehensive data on early-stage companies remains a challenge, with publicly available databases lacking in providing information at both the company and team level. This often leads venture capital firms to rely on their traditional methods, placing significant emphasis on human judgement in their evaluation of the proposed business model and the capabilities of the founding team.

Another reoccurring theme from our engagement with VC firms and personalities in Norway, was genuine interest in understanding who founds companies in Norway and who goes on to succeed. Due to the lacking existing research on the area, we found stakeholders in the Norwegian startup ecosystem were left to speculation and a tended to be biased towards own experiences to answer these questions.

The gap in the research literature, combined with the contrasting scenario of technological advancements on one side and its limited usage on the other, underlines the industry's call for transformation. In this evolving landscape abundant with unexplored possibilities, our curiosity was sparked, inspiring us to formulate our research question:

"How can machine learning be utilized to predict success of early-stage Norwegian startups and to identify founder characteristics that contribute to success".

Our exploration of this question forms the foundation of this thesis. The thesis is structured as follows. Firstly, the relevant literature and theoretical foundation will be presented in the Literature Review. Secondly, we will summarize how the dataset was collected and constructed. Following, we will then define startup success and produce a success metric for each startup company. An exploratory data analysis will then be presented to get familiar with the data and locate possible patterns and trends. The next part of the thesis is a theoretical introduction to the machine learning methods applied. A section explaining our feature engineering applied to prepare data for analysis follows before we delve into the results. The next step is to discuss these results regarding our findings, possible use cases, limitations, and ethical consideration, before we end the thesis with a conclusion.

2 Literature Review

The purpose of the literature review is firstly, to explain key theoretical frameworks that underpin this thesis. Secondly, it aims to highlight past research conducted on the subject and articulate how our study can enhance the existing body of knowledge.

2.1 Personality Framework: The Five-Factor Model

As the use of founder personalities plays a crucial role in answering our research question, it is necessary to present the theoretical framework surrounding these personality traits.

The Five-Factor Model (FFM), also known as the Big Five personality traits model is among the most well-known and regarded models for personality traits within psychology (Widiger, 2017). The use of five factors to describe personality was first introduced by Tupes and Christal (1961) who identified the recurrence of five factors in their analysis of personality ratings across eight distinct samples, a consistency unheard of in the field of personality traits thus far (McCrae & John, 1992). However, these findings did not reach academic popularity until the 1980s when multiple researchers from different fields and countries began reaching the same consensus (John & Srivastava, 1999). The popularity of the model amongst the academic field has since grown exponentially and John, Naumann and Soto (2008) went as far as to state "After decades of research, the field has now achieved an initial consensus on a general taxonomy of personality traits, the "Big Five" ".

The Five-Factor model is organized as a hierarchical taxonomy, where each of the five factors can be broken down into more specific facets or sub-traits (McCrae & John, 1992). Since the framework was developed by multiple researchers over a period of time, there are some discrepancies regarding labels and definitions (John & Srivastava, 1999). However, the OCEAN framework is considered to be the most applied in recent years, and we will therefore present this as defined by John and Srivastava (1999).

The OCEAN acronym is represented by the five factors Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Openness to experience, as opposed to closed-mindedness, encapsulates the range, depth, uniqueness, and complexity of an individual's mental and experiential life. Conscientiousness describes to what degree the individual inherits impulses that enable goal-oriented behaviours. It is associated with traits such as thoughtfulness before action, ability to delay gratification, adherence to norms and rules, effective planning, organization, and prioritization of tasks. Extraversion refers to an energetic engagement with the social and material world, characterized by traits such as sociability, assertiveness, activity, and a positive emotionality. Agreeableness embodies a prosocial and communal orientation towards others, characterized by qualities such as altruism, empathy, trust, and humility. Finally, Neuroticism is the contrast to emotional stability and even-temperedness, encompassed by emotional states, such as feelings of anxiety, nervousness, sadness, and tension. Note, in this paper we apply Emotional Stability as opposed to Neuroticism, which is simply the scale reversed (Humantic AI, 2023b).

2.2 Historical Research on Characteristics of Successful Founders

The study of entrepreneurship has a long history but is regarded as relatively new as an academic field (Carlsson et al., 2013). It is just recently the interest in research within the field has started gaining momentum, with an exponential growth in number of articles since the start of the 1990's (Chandra, 2018). However, as the field is still in its infancy, some scholars argue the field remains highly fragmented and lacks maturity and convergence (Schildt et al., 2006). Carlsson et al. (2013) finds there to be several areas within entrepreneurship that is still lacking substantial research. From our literature review of similar research, we find that there seems to be some degree of divergence in results between papers. However, we still find there to be some common ground.

For this study we are focusing on those factors which are available at the earliest stages of investment opportunities (pre-seed). In the earliest stages of startup's life, information is often limited to the founding team and the startup ide. However, as business ideas can be comprehensive and difficult to quantify and compare, the focus of this study is mainly the factors which are stemming from the founding team. As for the founding team's relative importance on successful outcome, the literature seems to differ slightly. Sevilla-Bernardo et al. (2022) did a statistical analysis of 60 articles regarding this conundrum. They found the founding team to contribute 8.9% and the CEO decisions to contribute about 10.3% to the outcome of a startup. However, some individual researchers have put a greater emphasis on the influence of the founding team to a successful outcome, such as Gross (2015), who found the total influence of the founding team and their execution ability to be 32%.

Numerous renowned entrepreneurs, including Mark Zuckerberg, Bill Gates, and Steve Jobs, launched their businesses in their twenties, fostering the widespread perception that entrepreneurship is most successful when embarked upon at a young age (Blumberg, 2021). Levesque and Minniti (2006) fuels the perception that firm creation is a "young man’s game", and presents findings that younger people aged 25-35, are more likely to become entrepreneurs. Azoulay et al. (2020) found that in contrast to popular belief, the most successful entrepreneurs are not young, but middle aged. They found the mean age of the 0.1 % fastest growing American startup companies to be 45 years. Furthermore, another finding from their analysis was that prior experience in a specific industry predicts much greater success rates. They reason their findings with that those advantages accumulating with age like human-, social- and financial capital, are overwhelming factors that young entrepreneurs may possess, like originality and energy.

Miloud et al. (2012) performed a study analysing 184 rounds of early-stage venture capital investments. They found that startups consisting of founder teams were likely to be valued higher than those of solo-founders. However, new research by Greenberg and Mollick (2018) finds that the common perception amongst investors that founder teams produce more successful startups is ungrounded. They present a potential gap in the literature regarding comparison of solo-founders versus larger founder teams, where, where it has been taken for granted that teams perform better. Contrastingly, by analysing a dataset of crowdfunded companied, they find that solo founders survive longer than teams, generate more revenue than founder pairs and do not perform significantly differently than large teams.

Westhead and Wright (1998) found in their renowned and highly cited work "Novice, portfolio, and serial founders: are they different?", that serial entrepreneurs did have significantly different characteristics from first-time founders regarding background, work experience, reasons for starting businesses, attitudes to entrepreneurship, and funding sources. However, they found there to be no significant difference in terms of the performance of their businesses. They argue that serial entrepreneurs are not a homogenous group and the number of founded companies does not predict success in itself. Another study originating from Harvard Business School by Gompers et al. (2010) found there to be a significant difference in performance between entrepreneurs who have succeeded before and the entrepreneurs who had failed in their previous venture, arguing that success breeds success.

Multiple studies performed over recent years have proven that there still exists a perception amongst the general public and professional investors that male entrepreneurs are likely to be more successful than female entrepreneurs (Thébaud, 2015) (Shane et al., 2015) (Brooks et al., 2014). However, as the findings of the field study of Kanze et al. (2018) proposes, the entrepreneurial gender gap, can be partly explained by a bias stemming from investors where females are asked prevention-focused questions while males are asked promotion-focused questions. A recent report by Innovation Norway (2019) supports these findings with statistics showing that about 45% of people interested in starting businesses are women, but they only make up 30.3% of actual entrepreneurs. Among businesses that survive past five years, only 19% are led by women, and in companies that generate a wage-competitive income, this proportion is lower at 16%. The percentage of women-led businesses with over 10% annual revenue growth is even lower, at 10%. They conclude their report with that there exists some major obstacles to female entrepreneurship, stemming from private investors, but also partly the government.

Among the most renowned and cited works on the Big Five framework and entrepreneurs are the meta-analytical review done by Zhao and Seibert (2006) and another meta-analytic review by the same group with the addition of G. T. Lumpkin four years later (Zhao et al., 2010). The first study was considered to be the first to compile empirical findings focused on the relationship between the five core personality dimensions and entrepreneurial characteristics.

A total of 23 studies were included, all comparing entrepreneurs to managers in terms of the Big Five personality traits. Their findings concluded that when comparing entrepreneurs to managers, entrepreneurs scored significantly higher on emotional stability, openness, and conscientiousness. Further, they found that entrepreneurs scored lower on agreeableness and that there was no significant difference in score for extraversion. The highest difference in scoring was for conscientiousness.

In their later study, another meta-analytical review was performed, this time studying the relationship between personality and entrepreneurial intentions and performance (Zhao et al., 2010). This time, 66 independent samples from 60 distinct studies were combined to reflect a total of 15,423 people in the meta-analysis. The results concluded that all traits, except for agreeableness, were positively linked to entrepreneurial performance, with the strongest effect found for openness, followed by conscientiousness and emotional stability. However, when investigating closer different types of business performance, such as growth and profitability, a more nuanced picture emerges. Conscientiousness and openness were both positively related to firm growth, but not necessarily to profitability. Emotional stability was positively related to both firm growth and profitability. Extraversion had a positive effect on overall performance, but specific performance types could not be detected. They argue that with an overall effect size of 0.31 (multiple R) for all five traits from their multiple regression that these are significant factors in determining entrepreneurial outcomes. However, these relationships are complex and could potentially be influenced by other factors.

Studies of job performance have revealed that the Big Five personality traits can be applied to significantly predict important outcomes in the workplace (Barrick & Mount, 1991) (Mount et al., 1998). Conscientiousness has been shown to be the greatest predictor of superior job performance across different sectors and dimensions, while other traits have been shown to predict performance in more specific areas. For example, Agreeableness and Neuroticism influence performance in roles that require group work, while Extraversion is associated with success in sales and managerial roles (John & Srivastava, 1999).

2.3 Bridging the Gap: The Unique Contributions of This Thesis

We propose that our study fills two quite unique gaps in the field of entrepreneurial study.

Firstly, as far as our research concludes, there seems to be very little research done on the Norwegian venture marked (Albertsen et al., 2021). Especially when considering our specific route of analysing the success of companies at a very early stage by only including founding information. As far as we can gather, the few studies done analysing Norwegian founder success seems to have had a qualitative approach, like the recent master thesis by Rosvold and Rosvold (2022). However, these findings do not surprise us, as it aligns with the dominating "old-school" approach of investing and negligence of AI opportunities in the Norwegian VC sector (Winther, 2023).We believe our thesis can shed some light on which early-stage detectable success factors are prominent in the Norwegian startup landscape and how these results compare with studies done in other countries and continents. Furthermore, we want to prove the applicability of machine learning methods for investment purposes in the Norwegian VC sector.

Secondly, our approach of gathering personality data is somewhat controversial and innovative. Traditional psychological entrepreneurship research typically relies heavily on questionnaires distributed in field studies (Leutner et al., 2014), as a result many research papers are based on a relatively small number of observations as it is limited by the number of participants in the field study. We find that many of these studies therefore lack the depth in numbers needed for certain machine learning methods and are therefore either qualitative or carried out by presenting descriptive statistics and doing statistical significance tests. This thesis on the other hand has collected estimated personality traits from ground-breaking artificial intelligence. We believe this paper can act as a starting point for further entrepreneurial research using machine learning methods on larger datasets gathered by applying cutting-edge AI technology.

3 Data Collection and Construction of Dataset

This section of the analysis, while brief, encompasses a significant percentage of the working hours completed to perform this study, as our comprehensive dataset is a product of extensive research, manual tasks, and data cleansing. Firstly, we will outline the various sources used for data collection, before discussing the complexities associated with our final merged dataset.

3.1 Data Collection

The absence of comprehensive databases encompassing the Norwegian Startup Ecosystem introduces significant obstacles to research and analysis. The Menon report for 2022 found that 1886 startups were established in 2021 in the Oslo Region. The number is reduced to 275 when considering only those having a particularly high growth potential (Albertsen et al., 2022). However, various sites and organizations collect data on these companies in a manner that aligns with their unique objectives and goals. The Brønnøysund Register Center documents all companies in Norway, but a filter for newly established firms encompasses all types of companies, not solely the ones fitting the defined definition of being scalable and technological-driven. Likewise, the publicly accessible database provided by Innovation Norway also include a wider range of companies and lacks comprehensive information about these firms. Data collected from Norwegian Venture Capital firms tend to correspond closely with the previously described company characteristics but are often confined to a smaller subset of startups suitable for their potential investments. Furthermore, person-specific data on founders are often limited or completely absent in most publicly available databases.

To overcome these limitations, we chose to collect data from multiple sources to produce our final dataset for our analysis. The subsequent section details each resource used in the creation of the final dataset, encompassing data at both the company and founder level.

3.11. Crunchbase

Crunchbase (www.crunchbase.com) is a globally acknowledged database platform offering detailed data on private and public companies worldwide. It provides a robust coverage of capital intensive and technology-driven startups, making it particularly appealing to venture capital firms and other potential investors (Liang & Yuan, 2016). Various sources rank Crunchbase as one of the best platforms on the market based on its extensive coverage of startups and high accuracy of key variables such as funding and location, as well as being a more cost-effective option relative to its competitors (Retterath, 2020). Using a Crunchbase Pro subscription, we identified 2768 Norwegian startups between 2010 and 2023. The result of exporting this data was a comprehensive set of 101 variables. Despite the overall reliability of the Crunchbase data, their data collection process is highly automated, employing web scraping and crowdsourced information (Startup Heatmap, 2023). This results in an extensive number of missing values and sometimes inaccurate and differing variable values. For example, the company names in some instances are not updated, exhibit variation in format, and occasionally language discrepancies. It is also important to note that this data is biased towards successful companies, as the percentage of bankrupt companies are relatively low.

3.1.2 Brønnøysund Register Centre

All Norwegian companies are registered through the Brønnøysund Register Centre (www.brreg.no). The Central Coordinating Register (Enhetsregisteret) is a publicly accessible database that includes basic information about companies, such as unique organizational numbers, official names, organizational forms, founding dates, and industrial codes (Brønnøysundregistrene, 2022). This dataset served as a reliable source for verifying some of the variables from Crunchbase and to match the Crunchbase data with the correct company name and organization number.

3.1.3 Proff & Forvalt

Proff (www.proff.no), operated by Proff AS, is a prominent search platform for businesses and industries in Norway, offering extensive business data to assist in decision-making, risk management, and business expansion efforts (Proff, 2023). A premium segment of the Proff service is Forvalt (www.forvalt.no), which, in partnership with Statistics Norway, Experian, and the Brønnøysund Register Centre, offers access to additional comprehensive, annual financial data and employee information (Proff Forvalt, 2023b). We were fortunate to have access to Forvalt through NHH, which enabled us to extract 2469 rows of detailed financial data related to the companies listed in Crunchbase, using their respective organization numbers. While many of the annual financial variables may not be relevant or available as predictors for startup success near the launch date, they played a vital role in characterizing each company's success. Moreover, the variables were helpful in excluding companies that did not align with our startup definition.

3.1.4 Manual Research and Data Collection

Both Proff and Forvalt were instrumental in the manual research required for matching and validating observations, as well as for gathering new data points. For example, the automated process of aligning company names from Crunchbase with the Central Coordinating Register in R only managed to identify half of the companies, and even those found were occasionally matched to the wrong entities. As a result, the only reliable solution was to manually research each company, verify if it was suited for our analysis, and collect the correct organization number. From these collected organization numbers, we could import the Forvalt dataset and further merge datasets.

Furthermore, we had to repeat this time-consuming process of manual searching to obtain the LinkedIn profiles and Proff employee profiles of the founders. This task was even more challenging, as we had to search for each founder individually for every company, given the limited founder names included in the Crunchbase dataset. This was driven by our commitment to incorporate detailed founder-specific data into our analysis.

In total, months of throughout research and investigation were devoted to acquiring the necessary information to ensure the validity and relevance of our analysis. However, during this labour-intensive process we were able to gain a substantial familiarity with the companies in our database and increasing our domain knowledge, which would help us in the analysis.

3.1.5 Humantic AI

Humantic AI (www.humantic.ai) is an artificial intelligence driven platform that specializes in deriving insights on personality and behaviour from text inputs. While the exact details of their algorithm are not publicly disclosed, the platform explains that they utilize machine learning, computational psychometrics, psycholinguistics, and social psychology to predict personality traits with a claimed accuracy of 80-100%, assuming adequate text input (Humantic AI, 2022). Despite scepticism from independent researchers like Rhea et al. (2022) about the stated accuracy, the platform is gaining popularity and is used by Fortune 500 companies such as PayPal, Caterpillar, and Cognizant (Humantic AI, 2023a). We are grateful for the sponsorship from NHH that facilitated our use of this innovative platform. Leveraging their API, we were able to make a script in Python to input the collected founder's LinkedIn profiles and extract estimated personality traits and basic founder details, such as education and work history. Particularly interesting to our analysis was the inclusion of the Big Five framework. Further discussions on the obtained traits, the validity of these estimations, and ethical considerations will be presented later.

Given that information such as birth year, gender, and the number of business roles are usually not made public on LinkedIn profiles, we turned to the collected Proff employee profiles, extracting this publicly available information by scraping the links using Python.

More details on the individual datasets are given in Appendix A.5-A.9.

3.2 Cleaning and Merging of Data Sources

The company-specific data collected was consolidated by merging the Crunchbase dataset and the Forvalt dataset. The result was a dataset composed of 2469 rows and 351 variables of raw data with unrefined information and a lot of missing values. Driven by our commitment to ensure data accuracy, we embarked on a comprehensive pre-processing phase to refine and optimize the dataset for future use. This process involved tasks such as reclassifying variables, discarding variables of low relevance or repetitive information, and various cleaning procedures of different variables. Moreover, we flagged or immediately deleted companies we suspected did not meet our startup definition, warranted further investigation, or had insufficient financial history to confidently determine their level of success. The detailed data cleansing process resulted in a narrowed down dataset of 1918 companies, a significant reduction from the original 2768 companies extracted from Crunchbase. This refined group of companies are largely made up of scalable startups in the technology sector and consequently creates the foundation for the analysis.

The founder-specific data obtained from Humantic AI and scraped data from Proff, encompassing 2925 founders and 54 variables, also required extensive cleaning. Many of the variables were encapsulated within text strings in json format, and numerous algorithms were implemented to create new variables and extract specific details.

3.3 Structure of Complete Dataset

In the process of merging the company and founder data, we opted for a structure where each row of the dataset represents an individual founder. This resulted in 2825 rows corresponding to 1918 unique companies. The restructuring also leads to a minor skew in the distribution of the target variable ("Target_Success"). Initially created to be perfectly balanced, the target now displays a 51.6/48.4 ratio at the founder-level, favouring success.

This decision was primarily motivated by our aim to focus on the founders and their individual contributions to start-up success. This approach capitalizes on the detailed founder-specific data we collected and considers the hypothesis that different founders could substantially impact a company's success in their own unique ways (Wasserman, 2012). However, this approach is not without challenges.

Firstly, our data structure results in dependencies between rows (founders) that are associated with the same company. This poses a challenge to the assumption of independence between observations, which is fundamental to conventional machine learning models. While advanced techniques, such as hierarchical or mixed-effects models, can address these interdependencies, they extend beyond the scope of standard classification models and introduce an additional layer of complexity (Gelman & Hill, 2006).

Alongside the issues of interdependencies, attributing company-level success, a collective effort, to individual founders, presents another challenge. This could overemphasize the significant of individual traits. For instance, a three-founder team's collective success is divided among each founder's traits in our data structure. This might imply that each founder's traits hold as much significance as those of a single-founder team, potentially distorting our understanding of the individual traits' impact on startup success.

To address these challenges, one might consider implementing a weighting mechanism (Hashemi & Karimi, 2018). For instance, in a company with three founders, each could be weighted at 0.33 to reflect their contribution. However, in practice, this approach introduced complexities that detracted from our objective of identifying broad patterns among founders, complicating interpretation of model outputs and risking new biases.

An additional complexity in our data structure arises from the wide time span (2010-2022) covered by the startups. Ideally, to ensure optimal comparison, startups founded within the same year should be considered. However, due to the limited number of observations, this was not achievable.

Despite these challenges, our chosen data structure aligns with our primary goal of uncovering general founder-specific characteristics that contribute to success. While it necessitates careful consideration in analysis and interpretation of results, it provides a meaningful framework for investigating the specific influences founders have on startup success.

4 Defining Startup Success

The consolidation of all the datasets left us with many variables useful for prediction, but no target variable that define the companies' level of success. Motivated by our goal of making the dataset as useful as possible and to simulate reality, we decided avoid shortcuts, and began a journey of creating a precise success variable that reflects our definition of success. Our reasoning behind leaving this research out of the literature review and present it in a separate chapter was due to the necessity of introducing our data sources beforehand to ensure understanding of our context.

4.1 Background Research

Regarding the characterization of a successful startup, the literature seems to somehow differ on what is considered to be the best metrics (Santisteban & Mauricio, 2017). However, it is apparent that the most used analysis methods would be some form of a two-factor analysis of certain characteristics. Most prevalent were survival analysis looking to define differences between survivors and non-survivors (Gartner et al., 1999). As stated by Statistics Norway SSB (2022), the survival rate of Norwegian newly established enterprises after five years is 26.5%, which would create a basis for a survival analysis. However, both due to the lack of information regarding most newly established enterprises in Norway and the lack of bankrupt companies in the Crunchbase dataset, a survival analysis is unfit for our purposes.

A second approach often applied when analysing the success of new ventures would be their ability to reach IPO or become acquired (Ünal, 2019). Reaching IPO-status is regarded as a staple of success for startups in USA, as it is regarded as a lucrative exit (Taulli, 2012). However, in Norway this could be considered a harsh requirement for success, as only 17 companies went public in 2022, furthermore only three of these companies went public at the main marketplace of Oslo Børs (Pareto Securities, 2022). As to the use of acquisitions as a factor for success, firstly there are a majority of missing datapoints on this variable from the Crunchbase data, secondly this distinction would be unfair towards companies in our database who have performed exceptionally but has yet to be acquired.

With these considerations in mind, it becomes apparent that we have no conventional method of differentiating our companies into pre-existing groupings. The created dataset consists mostly of companies who already to some extents have achieved success, as they are survivors and have been picked up by Crunchbase. Our method forward would therefore be to define the degree success by virtue of how they perform within the fields that the literature acknowledges as success, and as a supplement the answers provided by our interviews during background research.

4.1.1 Startup Success – Theoretical Foundation

Firstly, from a venture capitalist perspective growth in turnover is essential for an investment to be considered lucrative (Gartner et al., 1999). Our background interviews of Norwegian venture capital firms supported this claim. There seems to be consensus that growth is the most important factor. As most companies fail to survive beyond the first year (SSB, 2022), the successful companies would have to compensate the bankruptcies. One of the interviewees stated that they only considered one in about twenty investments to be truly successful. Venture capital provides capital to companies that might otherwise have difficulty attracting financing. There exists an information asymmetry between entrepreneurs and investors, which along with the highly dynamic markets entrepreneurs often operate in provides a high-risk investment. However along with the potential for rapid growth, this creates a high-risk high-reward assessment (Gompers & Lerner, 2001).

A second and somewhat obvious success factor is the size of the company itself at the time of our analysis. As the company has managed to obtain substantial size through its business activities, it would therefore be rational to resonate that the company's operations have been successful. Murphy et al. (1996), finds size to be one of the most frequently used dimensions for performance in the field of entrepreneurship research, with sales level being the largest contributor.

Another success factor highly regarded among the literature is the business's ability to create jobs. March-Chordà (2004) defines success by the number of jobs the company creates. The number of employees provides an indicator of the size of the company, and is among the most commonly used metrics for this purpose (Nassar et al., 2014). Davila et al. (2003) found that employee growth is positively related to changes in valuation, rationalizing their findings with: "The growth in number of employees may indicate that the business model of the firm is successful, and this success is spurring growth". They suggest that headcount growth therefore can be a relevant measure of success when public information is limited. Followingly, both previous growth in employment and the current number of employees are relevant success factors for investment and value creation purposes. In addition to company level financial growth, Decker et al. (2014) states that entrepreneurship and especially rapidly growing startups are essential for job creation in the United States on the macro level.

Good financial health is regarded as a staple of performance analysis as going into bankruptcy and business failure are likely to arise if health metrics are not considered (Rafiei et al., 2011). The use of financial ratios to measure firm performance is a widely renown and applied method. Delen et al. (2013) performed an explanatory factor analysis of financial ratios and found that liquidity, asset-structure, and asset/equity turnover rate alone explained more than 30% of the variance in performance. The data from Forvalt provides us with all these ratios but for some of the companies there are unfortunately missing values. However, conveniently Forvalt calculates a bankruptcy score for all companies based on the availability of data. The score is based on the above-mentioned financial ratios, but also records of credit and payment default (Proff Forvalt, 2023a).

Lastly, the inclusion of some sort of measure of the company's ability to produce financial gain would also be necessary. Murphy et al. (1996) listed profit and efficiency of investment as the two most used metrics of performance next to growth in the field of entrepreneurship research. From a venture capitalists' perspective, the primary motivation for investing is considered to be financial return (Metrick & Yasuda, 2021). As many of our companies are smaller in size and still early in development, specific data on each investor's return on investment is non-obtainable as this information may still be unconfirmed or private. However, firm specific profitability calculations can be used as an indicator.

4.2 Defining a Score for Startup Success

As a result of our research, it becomes apparent that more than one variable would be necessary to give a fair score of startup success, as there does not seem to be academic consensus around a single defining variable of performance. Our method going forward will therefore be to combine the variables defined by our research into a weighted score.

From our theoretical foundation we summarize our findings of success into four distinct categories:

1. Growth
2. Size
3. Profitability
4. Financial health

From our dataset we computed the following variables to best represent the dimensions:

Dimension	Variable	Derived from
Growth	1. Growth in turnover	"Sum_Driftsinnt_year_x" Year 1 to year 11
	2.Growth in employees	"Aarsverk_year_x" Year 1 to year 11
Size	3.Turnover	"Sum_Driftsinnt_year_x" Year 1 to year 11
	4.Number of employees	"Aarsverk_year_x" Year 1 to year 11
Profitability	5.Annual result / Turnover	"Aarsresultat_year_x" year 1 to year 11 & "Sum_Driftsinnt_year_x" Year 1 to year 11
Financial health	6.Forvalt bankruptcy score	Score_forvalt

Table 1: Variables to define the level of success.

Summarized in Table 1, we have a total of six variables of which we want to include in our weighted score. The calculation of these six scores was comprehensive, and a lot of important considerations were taken for each variable. However, the entire process of this calculation is quite long and complex, so we elected therefore to include it in the Appendix A.1, and instead provide a short summary inside the thesis.

Our methodology consisted of making specific calculations for each variable before we applied min-max normalization to give each variable a score in the range 0-100 (Al Shalabi et al., 2006). The growth of employees and growth in turnover were calculated using percentage growth calculated between accounting years. Then these calculations were combined into a single metric by aggregating them with weights so that growth closer to the firm's infancy was valued more. In the opposite fashion, the size variables used the same accounting metrics, but instead weighted values closer to their current status more so that their current size would define their success and not the size at the infancy of the startup. The profitability score was calculated by dividing results by turnover, also here the more recent years were weighed more. Lastly, the financial health score was simply calculated by averaging the bankruptcy likelihood indicator Forvalt Score as provided by Forvalt. For some of the variables, modifications were applied to reduce outliers before normalizing.

4.2.1 Weighing of Variables

Before combining the six scores into a single success score, we want to weigh them according to importance towards our definition of startup success. This was done by firstly applying the theoretical insight from our research and secondly by trial and error to see how different weights performed towards our analysis.

First and foremost, it became apparent from our research and interviews that future growth was the most important factor amongst venture capitalists when considering a possible investment. The possibility for expansive growth is essentially what differentiate a common newly founded company, like your local hairdresser, from a startup (Graham, 2012). The combination of information asymmetry between investors and entrepreneurs, and a high failure rate provides a high-risk scenario for investors (Sahlman, 2010), which in turn depends on a high growth rate among the successful companies for venture capital investment to be sustainable. Due to the different business models amongst the startups requiring different employment numbers, we find growth in turnover to be more reliable as a growth measure and chose to weigh it somewhat higher. Growth metrics are then accountable for 50 percent of the success score, with 30 percent stemming from turnover and 20 percent from employment growth.

As for the size variables we elect to weigh these relatively high at 30% (10% employee, and 20% turnover). As one can rationalize the size of the company represents the company's ability to perform their business activity on a higher scale.

Lastly, we argue that profitability and financial health are relatively less important when measuring startups' success and weigh these a total of 20% (10% each). Firstly, because one can assume it takes some time before a company becomes profitable due to entry barriers and costs. Secondly, newly founded firms are by nature risky ventures and will therefore possess less advantageous financial health scores. Furthermore, Oda Norway, a home delivery grocery service, which is acknowledged as one of the most successful Norwegian startups in the last decade (NTB, 2023), is among the worst performing companies in financial health due to a lasting struggle of establishing themselves amongst the highly competitive Norwegian grocery market. If we were to weigh the financial health score higher in the final score, we would risk losing companies regarded as successful amongst Norwegian venture capitalists.

Relative weights towards success score

Dimension	**Variable**	**Weight**
Growth	Growth in turnover	30%
	Growth in employees	20%
Size	Turnover	20%
	Number of employees	10%
Profitability	Annual result / Turnover	10%
Financial health	Forvalt bankruptcy score	10%

Table 2: Weights of success score

Before calculating the score, we remove the bankrupt and dissolved companies (137 instances), as we do not want these companies to upset the scoring. We then add them back to the dataset after the fact with a score of zero. For the companies where some of the six scores were unavailable, we used the average of the remaining scores for that company as a placeholder. By computing the scores as described above and combining them according to the weights, we obtain our target variable score. Table 3 showcases how our scoring system correctly presents some of the highest regarded Norwegian startups over the last decade as the top 10 performers.

Top 10 performing companies

Company Name	Success Score
DR. DROPIN AS	899.53
CUTTERS AS	883.13
SOCO NORGE AS	834.56
EASEE AS	827.02
SOLVENCIA AS	814.94
GODTLEVERTGRUPPEN AS	801.98
CLAVE CONSULTING AS	801.40
VENI AS	792.22
ODA NORWAY AS	779.83
STINGRAY MARINE SOLUTIONS AS	777.56

Table 3: Top 10 performing companies

4.3 Classification of Score

Some major issues arose when considering applying the success score directly for machine learning. Firstly, all the bankruptcies and dissolved companies were given the scoring value of 0 due to their failure to survive. A significant portion of these companies achieved an above average score when included, due to lacking or incorrect accounting data, and we elected therefore to provide them with a score of 0 instead, as they are essentially absolute failures. Consequently, our dataset consists of 137 companies scoring zero, which in turn could damage the results of our analysis. Secondly, our success score does not represent a natural range or metric for each company. It is merely a score we have created to the best of our ability. Therefore, we cannot with certainty assert that there exist relationships between each datapoint.

However, by applying classification and dividing the companies into groups based on ranges of performance we can avoid these complications while still having a clear distinction between performance of companies. We elected to go for a uniform 50/50 split dividing between the 50% best performing companies and the 50% worst performing companies. For more insight we also elected to create a split of three uniform groups to use for descriptive analysis. The decision behind going for uniform groups was because we have no natural or distinct boundaries to divide based on. We will still apply the success score in the descriptive analysis however, but for the machine learning methods the binary performance groups are applied.

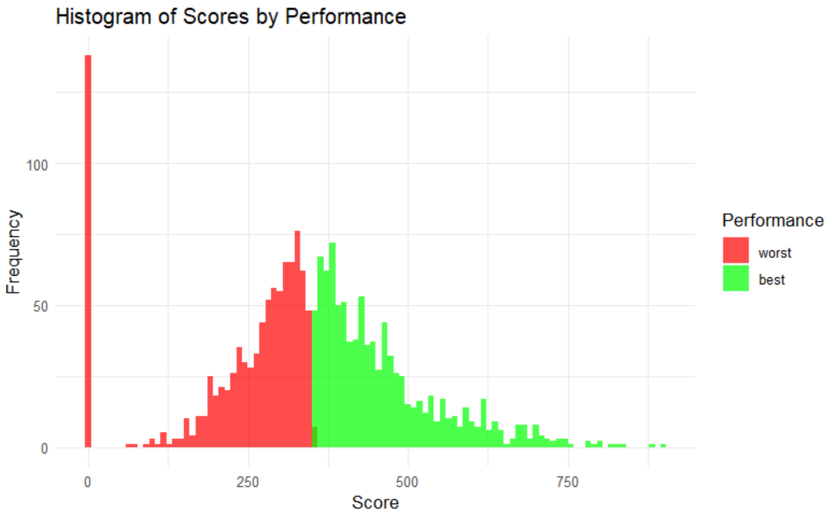

Figure 1: Two groups (K-means clustering)

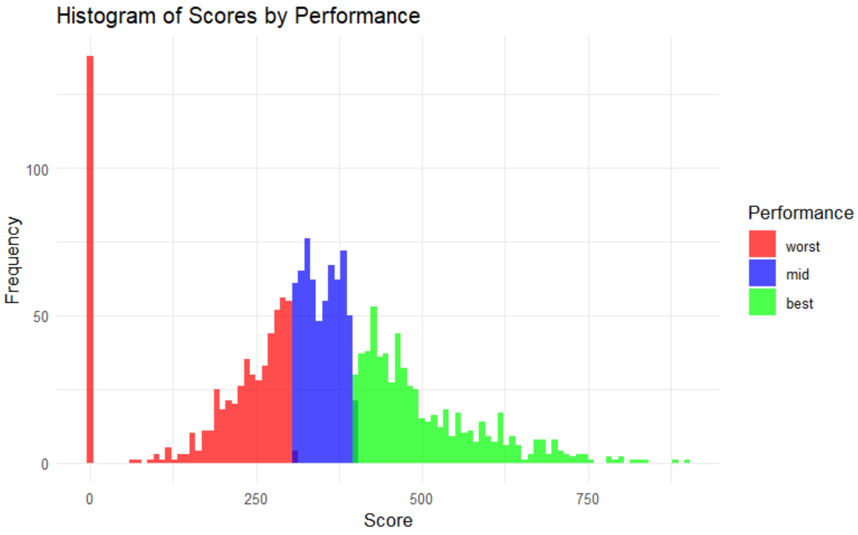

Figure 2: Three groups (K-means clustering)

4 Exploratory Data Analysis

The creation of our founder dataset was a labour-intensive task lasting about three months. As to our current knowledge, it is unique in its existence as no such founder dataset exists publicly or has publicly existed in Norway. The results stemming from our exploratory data analysis can therefore provide unseen insight into the workings of the Norwegian venture sector. A substantial part of this report will therefore be the presentation of these results. However, the focus will be towards observing patterns and not proving significance with statistical tests, as this task is left for the machine learning models. The purpose of this chapter is also to get familiar with the data for the purpose of feature engineering.

4.1 Company Characteristics

A deeper analysis on the company characteristics offers key insights into (1) the general characteristics of our companies and their distributions, (2) the companies' relative success in relation to our success score and classifications, and lastly (3) whether our unique Norwegian datasets aligns with the trends observed in previous research conducted on data from other nations.

4.1.1 NACE Groups

NACE Group distribution

NACE Group	Count	Percentage
J - IT and communication	1084	56.52
M - Professional, scientific, and technical activities	429	22.37
G - Wholesale and retail trade and repair of motor vehicles	96	5.01
N - Administrative and support service activities	91	4.74
C - Industry	52	2.71
K - Financial and insurance activities	47	2.45
L - Real estate activities	25	1.30
P - Education	13	0.68
H - Transport and storage	12	0.63
A - Agriculture, forestry, and fishing	11	0.57
F - Construction	11	0.57
R - Arts, entertainment, and recreation	11	0.57
Other grouping	10	0.52
Q - Human health and social work activities	10	0.52
Undisclosed or insufficiently disclosed activity	9	0.47

NACE Group	Count	Percentage
D - Electricity, gas, steam, and hot water supply	7	0.36

Table 4: Distribution of NACE groups

NACE is the classification of economic activity in the European Union (SSB, 2023). As expected, tech companies are predominantly represented in our dataset. Group J and M are most represented and can be divided into the subgroups as displayed in Table 3 and 4.

Top 10 NACE subgroups in "J - IT and communication"

NACE subgroup	count
Programming Services	614
IT Consulting Services	308
Operation of Web Portals	45
Data Processing, Data Storage and Related Services	34
Publishing of Other Software	17
Production and Publishing of Music and Sound Recordings	13
Other Services Related to Information Technology	10
Management and Operation of IT Systems	9
Publishing of Video Game Software	7
Production of Film, Video and Television Programs	6

Table 5: Top 10 NACE subgroups in category J

Top 10 NACE subgroups in "M - Professional, scientific and technical activities"

NACE subgroup	count
Other Technical Consulting Services	92
Business Consulting and Other Administrative Consultation	89
Other Research and Development Work in Natural Sciences and Engineering	65
Advertising Agencies	49
Industrial Design, Product Design, and Other Technical Design Services	34
Head Office Services	25
Research and Development in Biotechnology	18
Other Professional, Scientific and Technical Activities Not Elsewhere Classified	14
Graphic and Visual Communication Design	13
PR and Communication Services	8

Table 6: Top 10 NACE subgroups in category M

The two major NACE groups are dominated by IT-services and consulting firms. On one hand, this can lead to a skewed dataset of a specific sector of the Norwegian business landscape, and thereby to some degree delegitimize the findings of this report. However, we argue the overrepresentation of tech companies is simply a consequence of a tech trends in startups over the last decades (FasterCapital, 2023), combined with the fact that tech and consultancy firms more often inherit capabilities for major scalability (Werth & Greff, 2018) (The World Financial Review, 2023), and are therefore more likely to be discovered by investors and posted on spaces like Crunchbase.

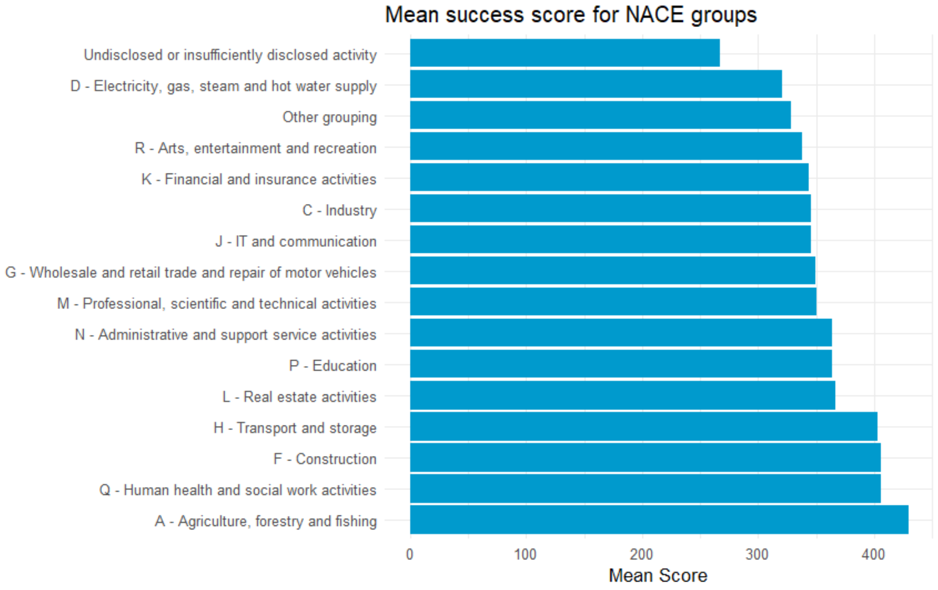

Figure 3: Mean success score for NACE groups

Figure 3 illustrates that some NACE groups, on average, appear to outperform others based on our success score. Nevertheless, this difference might be caused by the insufficient number of observations within these groups, as the most frequent groups seems to have similar scores. Still, we do observe a pattern among the most successful groups, which mainly represent industries requiring large investments to enter the market. Give that our definition of success incorporates the size of the company, this could explain why these sectors have higher average scores.

4.1.2 Founding Year

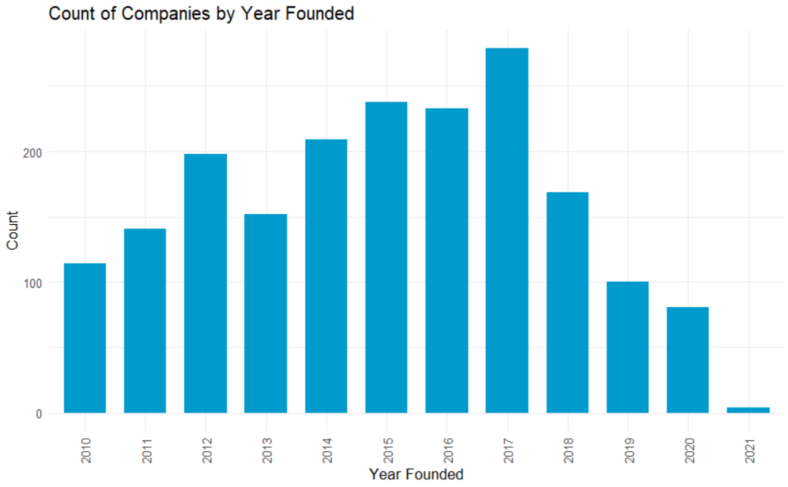

Figure 4: Bar plot of companies by year founded

We observe there to be a clear trend in companies founded per year, it seems to be rising until it reaches a maximum in 2017 before it again decreases. There could be some trending in when people decide to start companies, yet we believe it is more likely to do with Crunchbase's detection strategy. Our reasoning being that Crunchbase routinely removes companies that have gone bankrupt, and therefore fewer companies will remain from the earlier years. Among the newest companies, one could argue that Crunchbase is lacking information as many of these companies are too small or insignificant for the company to be added to the database. If these assumptions are true, this is somewhat unfortunate as our data is slightly biased consequently.

4.1.3 Number of Founders

The European Startup Monitor noted in 2016 that the average number of founders in Europe was 2.5 (Kollmann et al., 2016), while others report a slightly lower estimation around 2. While there might be cultural and industry differences, we should expect a somewhat similar range in Norway.

Average number of founders manually found (rows)	Average number of estimated founders (“Num_Founders_c”)	True average number of founders per company
1.47	1.82	2-2.5

Table 7: Average number of founders

Table 7 shows that the average number of founders we were able to identify is significantly lower than the true average. This is not surprising, given our experience of the difficulties of identifying certain founders. During our research, we recognized the pattern that unsuccessful founders, like those involved in a bankrupt company, did not advertise themselves to the same degree online compared to successful founders. This can make the data slightly biased as it would contain more successful founders. On the other hand, we did not have the resources to identify more than three founders per company. If a company with a high number of founders are generally more successful, the lack of identification of these founders could create a bias in the opposite direction.

The number of estimated founders contains the number of manually collected founders per company or the number of founders variable from Crunchbase if this number was above three and available. Figure 5 showcases a small trend in increased success scores among companies with more founders with the mean performance rising from 343 at one founder to around 370 for companies with three or more founders. Thereby supporting the perception amongst investors that founder teams perform better than solo-founders (Miloud et al., 2012).

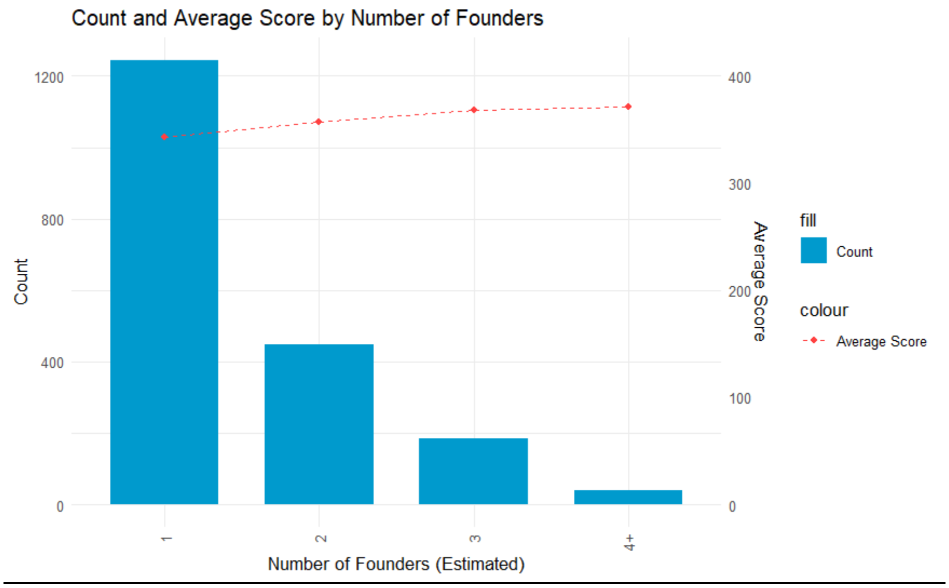

Figure 5: Bar plot of number of estimated founders

4.1.4 Location

Figure 6 showcases the distribution of counties the company is operating in and the average success score. Oslo is overrepresented, which is expected as this is the major business destination of Norway (Albertsen et al., 2021). The top performers, Oslo, Rogaland, and Nordland distinguish themselves from the others by a small margin. The top performer being Rogaland is likely due to the city of Stavanger's status as the petroleum capital of Norway, with many companies connected to the industry (Gjerde, 2023). As for Nordland, this is the county with the fewest observations, making the score somewhat less reliable and could be due to a fluke. However, it could be due to large investment industries like fishery and fish farming located along the county's long coastline (NHO, 2023). As for Oslo, the high average performance is likely a result of many of the major companies' headquarters being in Oslo, contributing to the high average score.

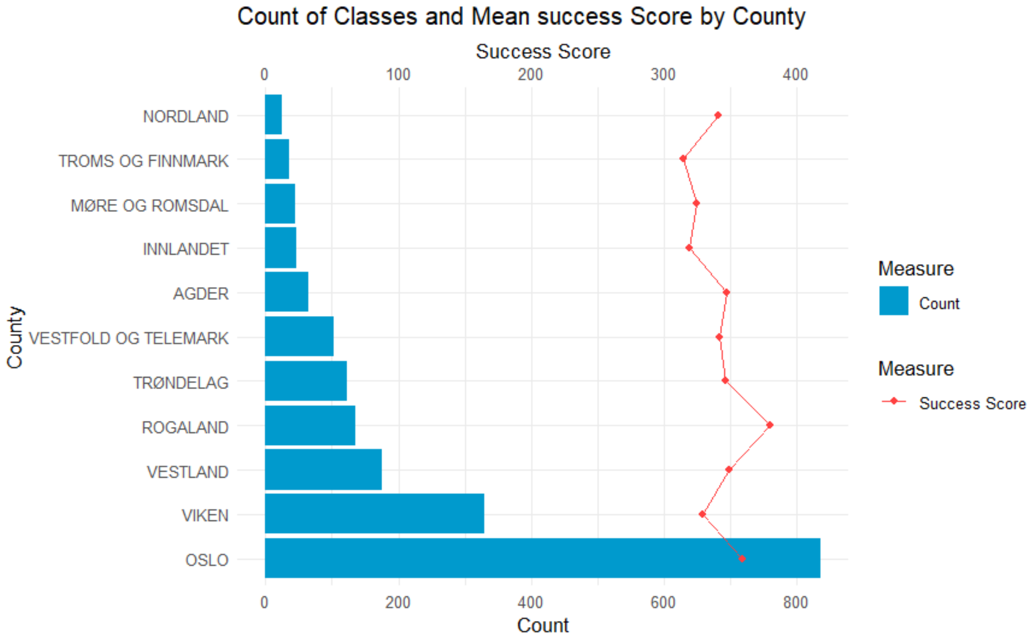

Figure 6: Bar plot of the county the company are located in

4.2 Founder Characteristics

Studying the founder data gives us valuable insights into what results we might expect from our machine learning models. Our goal is to outline key characteristics of the founders, understand how these characteristics correlate with our success measures, and if these trends are supported by previous research.

4.2.1 Number of Founded Companies

Our final 1918 companies were linked to 2825 founders, whereas 2700 were unique founders. This means that several founders have started more than one company in our dataset.

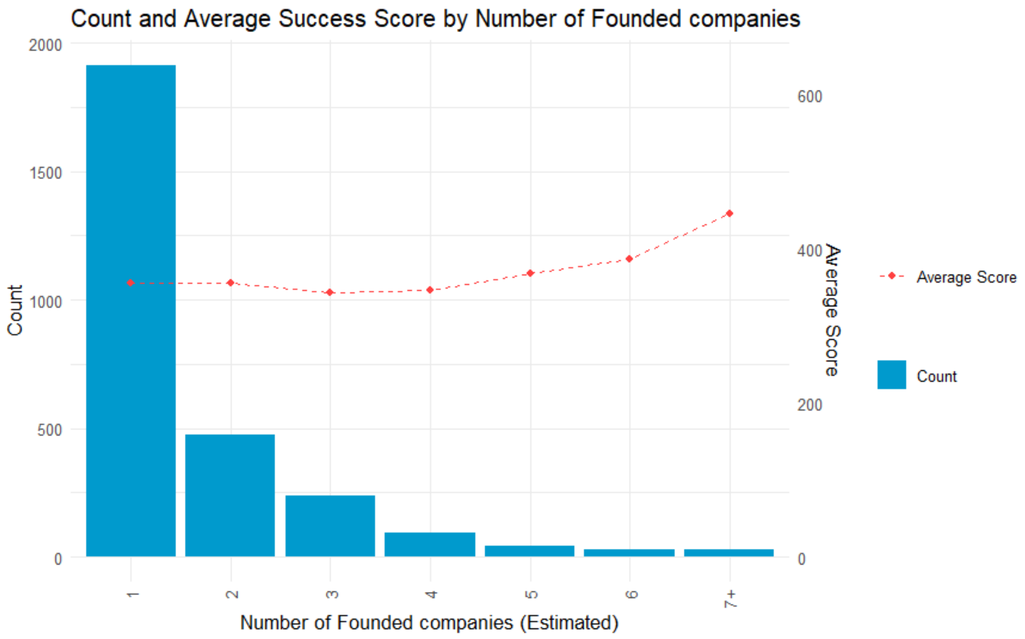

Figure 7: Bar plot of number of founded companies per founder.

Figure 7 showcases that our data distribution seem to align with common belief, that serial entrepreneurs outperform first timers, but contradict the study by (Westhead & Wright, 1998), who found no significant difference in business performance between novice and serial entrepreneurs. However, these findings can to some degree be supported by Gompers et al. (2010) who found that previous successful entrepreneurs were likely to succeed again.

4.2.2 Founder Demographics

Table 8 provides an overview of key characteristics by gender, offering insights into which demographics are likely to start a company.

Demographic Summary

	Male	Female	Total
Count	2560	265	2825
Percent	90.62	9.38	100.00
Mean Success Score	357.62	335.88	355.58
Performance Group worst 1/3 (%)	31.80	36.98	32.28
Performance Group mid 1/3 (%)	32.54	32.08	32.50
Performance Group best 1/3 (%)	35.66	30.94	35.22
Mean age (At startup)	38.20	38.02	38.18
Mean experience in years (At startup)	11.83	10.88	11.74
Mean number of jobs (Today)	8.50	9.12	8.56
Mean number of schools (Today)	2.51	3.45	2.60
Education: Bachelor's (%)	25.63	21.31	25.20
Education: Master's (%)	42.50	47.95	43.04
Education: Doctorate (%)	4.26	6.56	4.49
Education: None or missing data (%)	27.60	24.18	27.27
Mean founded companies	1.63	1.57	1.62
Mean number co-founders	2.07	1.99	2.07
Mean LinkedIn followers (Today)	1437.36	1527.00	1445.77

Table 8: Summary statistics of founder demographics

The average Norwegian entrepreneur is a 38-year-old male who has accumulated 11-12 years of experience before starting his company. He is likely to have some form of higher education, at least a bachelor's degree, but more likely a master's degree. He is likely to start the company alongside someone else, and most likely just one partner. The somewhat high average of LinkedIn followers of 1446 could indicate that the average Norwegian entrepreneur is active and engaged on social media and follows news and trends in his business sector.

We observe that women are severely underrepresented in our dataset, with females only being 9.4% of the total founder population. They perform slightly worse than their male counterparts, with a lower average success score and a relative overrepresentation in the worst 1/3 of startups.

In line with the report from Innovation Norway (2019), we anticipated these findings, yet our dataset appears to have an even smaller representation of females. This discrepancy could be because of the pronounced presence of tech startups, which traditionally attract fewer women (Sven Blumberg, 2023).

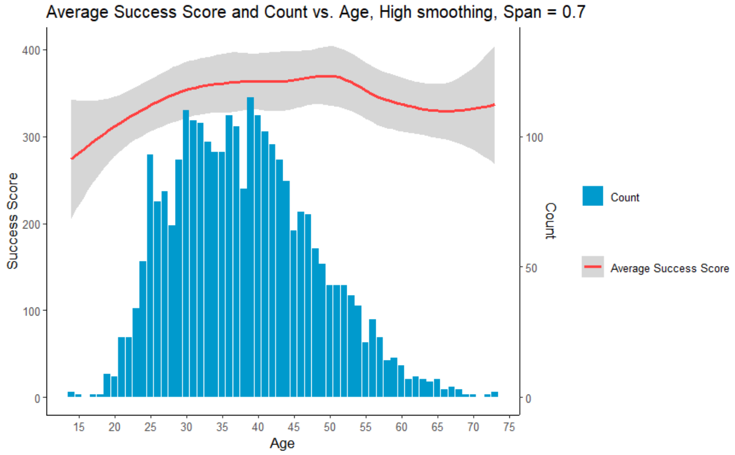

Figure 8: Average success vs age with loesss (span of 0.7)

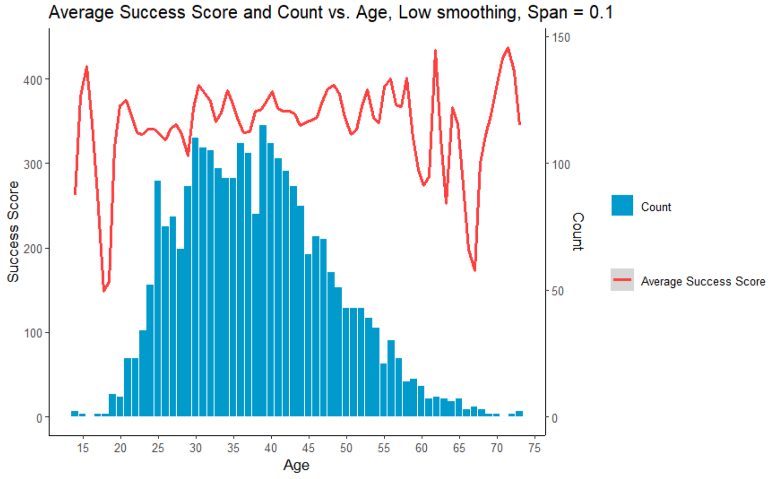

Figure 9: Average success vs age with loess (span of 0.1)

By, applying the Loess smoothing method (Appendix A.2), the Figures 8 and 9 are produced. In figure 8, we can observe the distribution of age and average success score. It is apparent people in their thirties and early forties are the most occurring entrepreneurs. We observe a trend of middle-aged entrepreneurs (40-50) creating the most successful companies, as predicted by

Azoulay et al. (2020). However, by reducing the smoothing metric substantially to 0.1, a more complex and noisy relationship occur in Figure 9. A trend of the youngest entrepreneurs performing worse is still somewhat apparent, but after the age of 30 it is difficult to argue that any likely pattern exist.

4.2.3 Personality Data

The personality data consists of scores for each of the Big Five traits with a range of 0 to 10. The scores are calculated to be normally distributed around the mean of 5, representing the distribution of the total population.

A Critical Overview of Humantic's Personality Estimates

Before diving deeper into the personality traits, we aim to critically examine the estimates provided by Humantic AI. As this platform and its methods applied are somewhat ground-breaking and operating in untested waters at an academic level, it is appropriate to exercise caution before using these traits blindly.

Roughly 15% of the founders, which equates to 412 individuals, had LinkedIn profiles that did not provide sufficient information to exceed the 40% confidence threshold set by Humantic AI. Consequently, they have missing values for all personality traits. Low confidence scores are likely to result in high variation and inaccurate estimates that can reduced validity of our findings and should therefore be carefully considered. For those surpassing the required threshold, the average confidence score is 85.42%.

As a validation test, we applied factor analysis (Appendix A.3), and got the result as listed in Table 9.

	MR1	MR2	MR3	MR4	MR5
Cumulative Var	0.33	0.59	0.65	0.65	0.65
Factor Loadings					
Openness	0.70	-0.62	-0.09	0.08	0.00
Conscientiousness	-0.19	0.67	0.13	0.01	0.00
Extraversion	-0.47	0.28	0.35	0.01	0.00
Agreeableness	-0.46	0.57	0.33	-0.09	0.00
Emotional Stability	0.83	-0.24	-0.20	0.01	0.00

Table 9: Factor analysis of Big Five traits

The results are somewhat alarming, as each trait is supposed to be independent, and we would therefore expect the traits to dominate different factors to a higher degree. Furthermore, all variance explained by this model is from the first three factors, and nearly all variance by the first two. To investigate further we calculate a correlation matrix as seen in Table 10.

	Openness	Conscientiousness	Extraversion	Agreeableness	Emotional Stability
Openness	1.00	-0.56	-0.53	-0.71	0.75
Conscientiousness	-0.56	1.00	0.32	0.51	-0.34
Extraversion	-0.53	0.32	1.00	0.48	-0.52
Agreeableness	-0.71	0.51	0.48	1.00	-0.58
Emotional Stability	0.75	-0.34	-0.52	-0.58	1

Table 10: Pearson correlation of Big Five traits

We observe that multiple of the traits seem to have strong correlations between each other. This could indicate that the AI algorithm is biased and inaccurate. However, when taken into consideration the bias of our population, being that they are all entrepreneurs, there are some reasons to believe our data is correct after all. The correlations does somewhat align with the findings of Zhao and Seibert (2006), who found that when comparing entrepreneurs to managers, entrepreneurs scored significantly higher on emotional stability, openness, and conscientiousness. Further, they found that entrepreneurs scored lower on agreeableness and that there was no significant difference in score for extraversion. These findings can explain the two highest correlations, as emotional stability is highly correlated with openness, and agreeableness is highly negatively correlated with openness. Furthermore, extraversion is the trait that appears to be among the least correlated to the other variables. However, we observe that among the lower correlated variables there are some unexpected tendencies, especially conscientiousness, that correlates somewhat opposite of what we were expecting. Still, keeping in mind that correlations below 0.5 are usually considered weak, and only two correlations for conscientiousness barely make it above this threshold, there is no reason to discard the validity of the data on this basis.

There are reasons to believe that the "true" distribution of our data is unknown, as there appears to be no large personality databases or research papers regarding Norwegian entrepreneurs. However, on a general basis, Norwegians tend to score lower on extraversion and openness, and higher on agreeableness (Løset & von Soest, 2023). Furthermore, it has been shown that there are major inconsistencies between distribution of scores between men and women as well (Sharma et al., 2022), which further separates our data from the general population as it is above 90% males.

To summarize, the data fails to pass validity tests for Big Five analysis. However, this could be due to our biased data dominated by male, Norwegian entrepreneurs. Furthermore, the largest correlations between the traits seem to line up with what one could expect from a dataset of entrepreneurs according to earlier research. We can therefore neither prove validity nor reject the validity of the personality data. We must therefore keep these possible shortcomings in mind when presenting the data and analysis results.

Personality: Descriptive Overview

By reviewing the summary statistics in Table 11 and the density plot in Figure 10, we get an overview of how the personalities of the founders are distributed.

Trait	Mean	Median	SD
Openness	6.57	6.55	1.18
Conscientiousness	5.79	5.82	1.32
Extraversion	5.42	5.45	1.25
Agreeableness	6.06	6.08	1.52
Emotional Stability	6.80	6.82	1.59

Table 11: Summary of distributions of Big Five traits

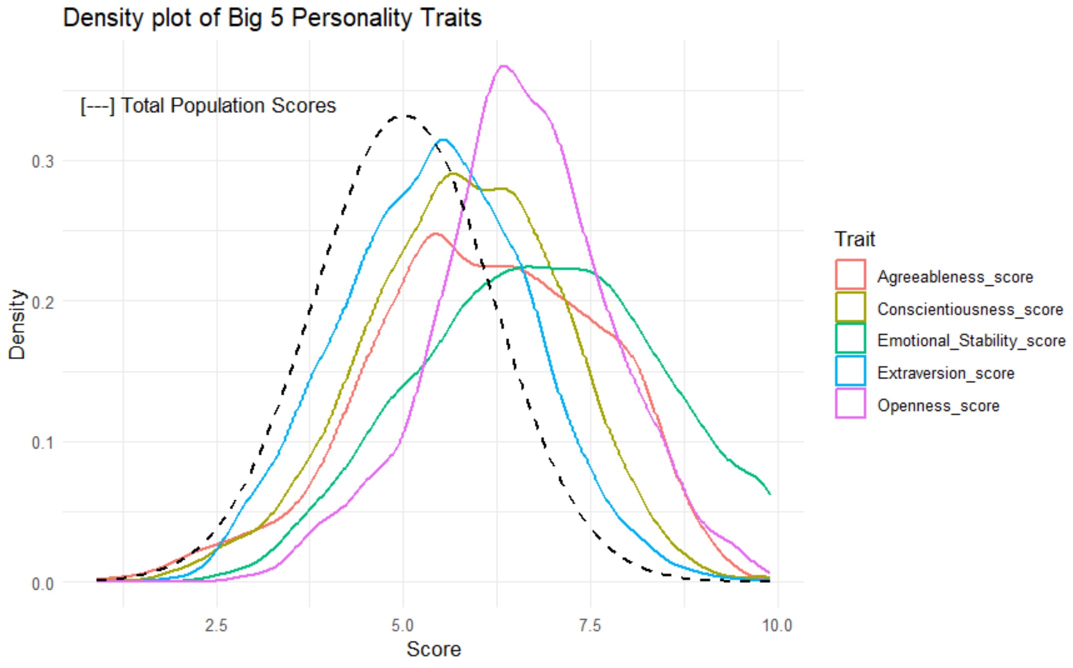

Figure 10: Density plot of Big Five traits

The overall scores seem to be skewed towards the right of what we would expect from the total population. Especially, openness and emotional stability emerge as important traits among entrepreneurs in our dataset. Conscientiousness also scores above average, while extraversion is the closest to the population mean. These results falls in line with the meta-analysis of Zhao and Seibert (2006), however the above average agreeableness score is contrary to their findings. We notice all scores resemble normal distributions to some degree, which makes them eligible for using parametric statistics.

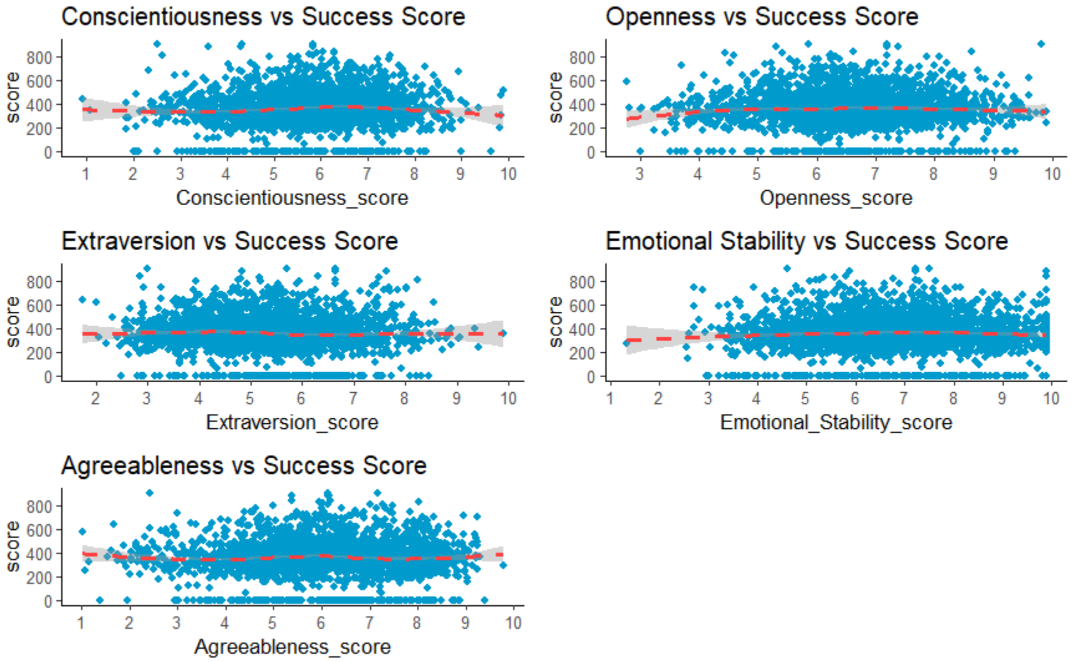

Figure 11: Scatter plots of Big Five traits vs Success Score

Looking at Figure 11, we observe that there is a lot of noise in the scatterplots of personality traits vs the success score, and it is hard to interpret any trends visually from the loess smoothed line. However, there might still be complex trends undiscoverable by the naked eye. There might be combinations of variables that increase success, like older males with high extraversion could be an example. This is where our machine learning methods may be especially helpful.

Personality Traits Performance Group Summary

Mean of Trait	**Worst Performers**	**Mid Performers**	**Best Performers**
Openness	6.54	6.59	6.57
Conscientiousness	5.72	5.75	5.87
Extraversion	5.46	5.43	5.36
Agreeableness	6.05	6.06	6.06
Emotional Stability	6.78	6.80	6.81

Table 12: Big Five traits vs three groups

In Table 12, we look at performance from another angle using the three predefined performance groups each containing 1/3 of the founders. Firstly, we observe that conscientiousness seems to increase with performance, and emotional stability also to some degree. This is consequent with earlier research (Zhao et al., 2010). However there seems to be no substantial trends in agreeableness and extraversion seems to negatively correlate with success score.

5 Methodology

5.1 Selection of methods

In machine learning a major distinction between models is whether the learning is supervised or not. Supervised learning models provide an associated response measurement for each observation, while unsupervised models no such response variable is provided (James et al., 2013, p.26). As our intention was to understand the underlying factors for success, we elected to focus on supervised learning models as our intention is to apply our success score as guiding variable. However, for another analysis where the relationship between the variables is of interest, we believe our dataset could provide useful insight into relationships between firms and startups in Norway using unsupervised methods.

Given that our target variable is categorized by levels of success, we must employ methods capable of handling classification. Firstly, we will apply Logistic Regression as our benchmark model as it is easy to interpret and is considered to be the most applied classifier in real-life applications (Yang & Loog, 2018). Furthermore, we will apply Extreme Gradient Boosting, Random Forest, and K-Nearest Neighbours as supplementary models. The rationale behind our selection of models is to explore models that demonstrate different trade-offs between accuracy and interpretability. Generally, more sophisticated models often yield high performance on complex datasets but at the expense of interpretability (Abdullah et al., 2021).

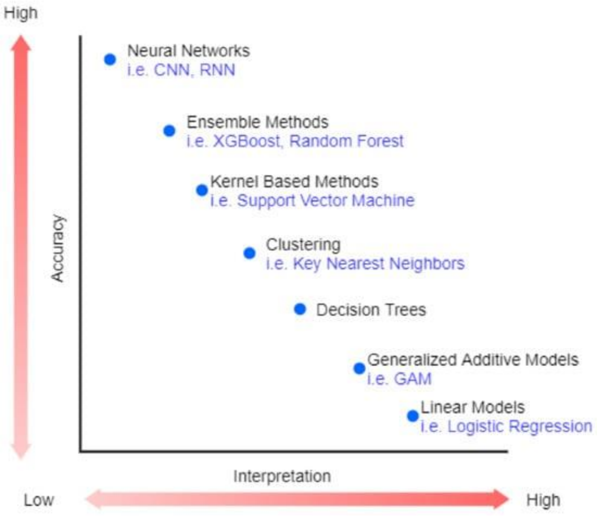

Figure 12: Accuracy vs interpretability (Abdullah et al., 2021)

5.2 Logistic Regression

Logistic Regression, a type of Generalized Linear Model (GLM), presumes a linear relationship between the dependent and independent variables. This model's objective is to yield a probability outcome that lies within the range of 0 to 1, implying that outcomes closer to 0 predict the first outcome, while those closer to 1 predict the second (James et al., 2013, p.134). To ensure this, the underlying function of the model must be designed to produce a prediction within this 0 to 1 range for any input. One such function is the logistic function, where each coefficient is represented by β_p.

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

Equation 1: Logistic function

The model ensures linearity using log odds, which is essentially the logistic function reorganized and by taking the logarithm of both sides. For each incremental increase in X, the log odds increase with β_p. Consequently, for each one-unit increase in X, the odds of the given outcome occurring gets multiplied by e^{β} (James et al., 2013, p.135).

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Equation 2: Log odds function

The estimation of the regression coefficients is done through the method of maximum likelihood, an iterative process that adjusts the coefficients until the likelihood function is maximized. The machine learning task is to find the estimates for $\hat{\beta}_0$ and $\hat{\beta}_p$ that maximizes this objective (James et al., 2013, p.135).

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_1) \qquad \prod_{i':yi'=0} (1 - p(x_{i'}))$$

Equation 3: Likelihood function

The primary advantage of using logistic regression is the simplicity and interpretability of results. The coefficients provide useful insight regarding the importance and direction of each independent variable. There are also multiple useful tools to develop for understanding underlaying factors, like variance importance plots. Another advantage of logistic regression is that there is no normality assumption, which is advantageous when dealing with classification problems (Healy, 2006).

However, the assumption of linearity can limit the model's prediction ability, given that real-world data can stray far away from linearity. Furthermore, logistic regression could have problems handling outliers (Stoltzfus, 2011).

5.3 Tree-based Methods

Tree-based methods are based on the use of decision trees, specifically classification trees in our case. Decision trees take on the task of segmenting the predictor space into smaller regions based on how the observations compare to the independent variables (James et al., 2013, p.327). By doing this multiple times, each split divides the data into subgroups until a certain stopping criterion is met and a prediction is provided, thereby creating a tree like structure with branches and nodes.

The most defining predictors will be closer to the root of the tree, as all observations are divided by thus criteria. The final prediction is made based on what is the most commonly occurring class in the terminal node. Pruning is applied to reduce the size of the tree to reduce noise, as larger trees are prone to overfitting (James et al., 2013, p.338). The splits and the pruning process are calculated through machine learning methods differing on the applications of the decision tree.

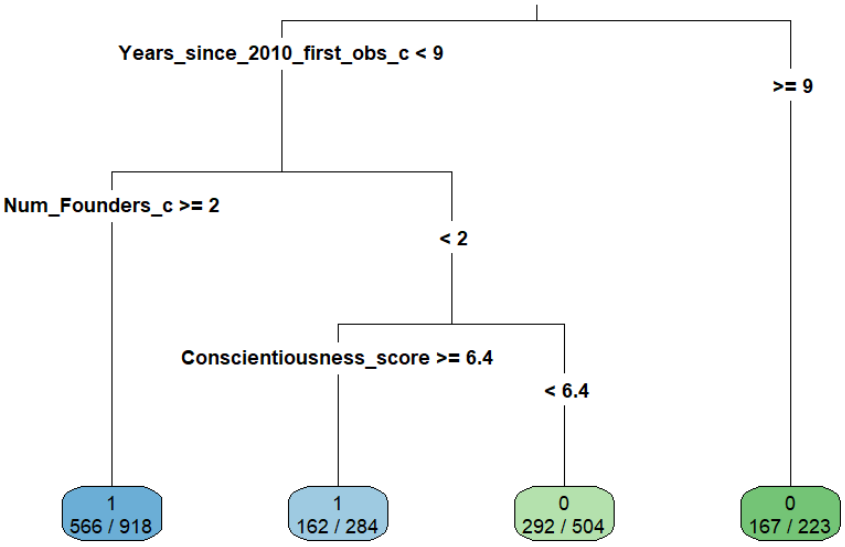

Figure 13: Decision tree example

The presentation makes decision trees easily explainable, as the simplicity makes it interpretable for people even with limited statistical experience. It is easy to follow the logic as one can argue it resembles human decision making. However, decision trees tend to perform worse than other classification approaches. As there are distinctive splits dividing the data, small changes in data can therefore have major consequences on the structure of the tree (Bengio et al., 2010). Due to the high variance, decision trees are considered non-robust. However, by applying multiple trees in a single model, we can increase the robustness of the model. This is the concept in extreme gradient boosting and random forest.

5.3.1 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble learning method, meaning it combines many simple "weak learners" to construct the model itself (James et al., 2013, p.340). In our case the "weak learners" are decision trees. XGBoost is a scalable and efficient application of the gradient boosting network by (Friedman, 2001) (Friedman et al., 2000). Boosting works by growing each tree subsequently, applying the information gained from the previous tree (James et al., 2013, p.345). Machine learning is applied to minimize the objective function, which consists of a loss function and a regularization term (Chen & Guestrin, 2016).

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Equation 4: XGB objective function

The loss function (l) will differ depending on the format of data, in our case as we have a classification problem, it will apply the logistic loss function. The logistic loss function punishes incorrect classifications by considering the likelihood associated with each classification (Saha, 2018). The objective is to find the sets of splits that optimizes this function. The negative gradient is applied as a guidance for the construction of the next tree in the ensemble, using gradient descent optimization (Chen & Guestrin, 2016). Ruder (2016) defines gradient descent as "a way to minimize an objective function $J(\theta)$ parameterized by a model's parameters $\theta \in R^d$ by updating the parameters in the opposite direction of the gradient of the objective function $\nabla_\theta J(\theta)$ w.r.t. to the parameters". The regularization term (Ω) is applied as a pruning method to penalize complex trees that may lead to overfitting.

To ensure our XGBoost results are optimized we perform hyperparameter tuning. Hyperparameters are used to configure the machine learning model itself or to minimize the loss function. The tuning of hyperparameters is considered a key-part of building machine learning models, especially tree-based models like XGBoost and random forest (Yang & Shami, 2020). Hyperparameter Optimalization (HPO) is applied to automate the tuning process. While there are different techniques to perform HPO, we decided to perform grid search which is "a decision-theoretic approach that exhaustively searches the optimal configuration in a fixed domain of hyper- parameters" (Yang & Shami, 2020) .

Hyperparameters for XGBoost

Hyperparameter	Tuned value	Description
Stop_iter	5	" The number of iterations without improvement before stopping. "
Trees	1000	" An integer for the number of trees contained in the ensemble. "
Min_n	3	" An integer for the minimum number of data points in a node that is required for the node to be split further. "
Tree_depth	3	"An integer for the maximum depth of the tree (i.e. number of splits)"
Learn_rate	0.0	" A number for the rate at which the boosting algorithm adapts from iteration-to-iteration."
Loss_reduction	1.4	"A number for the reduction in the loss function required to split further."
Sample_size	0.8	" The number of iterations without improvement before stopping."

Table 13: Hyperparameters for XGBoost

All descriptions are directly quoted from the Parsnip Tidymodels documentation on boosted trees (Tidymodels, 2023a).

XGBoost has many advantages over other machine learning methods and is considered one of the best performing and most popular systems available (Nielsen, 2016). It has great scalability and can handle large datasets and it is known to be both computationally fast and efficient (Chen & Guestrin, 2016). It also has great flexibility with many parameters to define custom optimization. Furthermore, by applying regularization, the learning is set up to actively avoid overfitting. Still, XGBoost supreme pattern-recognition ability makes it prone to overfitting even with the built in regularization as it may find patterns in noise. Another disadvantage is its complexity, which can make it hard to interpret.

5.3.2 Random Forest

Random forest is another ensemble method that uses decision trees as "weak learners". In random forests, bagging is applied to create an ensemble of trees. Bagging is the aggregation of multiple decision trees created from a number of bootstrapped samples to reduce variance (James et al., 2013, p.340). This method involves resampling the data with replacement to estimate the sampling distribution. Equation 5 illustrates the bagging formula where B is the different bootstrapped training data sets, while $\hat{f}^{*b}(x)$ is the trained model for each bootstrap (b).

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

Equation 5: Random Forest bagging formula

However, random forest has another feature that separates the method from standard bagging methods, which is the concept of random feature selection (James et al., 2013, p.344). In essence, this means decorrelating the ensemble of trees by forcing the model to only be able to select a random set of the predictors available. The reasoning behind being that a large majority of the trees likely would have chosen the same predictor at the root of the tree if allowed, and thereby construct an ensemble of highly similar trees. By forcing the machine learning method to consider different sets of predictors, we can reduce variance more due to a less correlated forest (James et al., 2013, p.345).

In bagging each tree is usually grown deep, and therefore no pruning methods are required (James et al., 2013, p.341). For the splits of branches, we applied the Gini index, as it is the default and standard method applied in the "Ranger" package for R which we used to implement our model (Wright & Ziegler, 2015). The Gini index measures how often a randomly selected observation would be incorrectly labelled given it was randomly labelled by the distribution of the labels in the subset. Meaning if most observations in the subset is the same, the Gini index is small, and vice versa large if the observations have no clear majority of class. The machine learning algorithm will summarize a weighted sum of the Gini impurities and try to minimize this metric.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Equation 6: Gini function

In formula 6, $\hat{p}_{mk}$ is the proportion of training observations in the mth region, so that $\hat{p}_{mk}(1 - \hat{p}_{mk})$ calculates the probability of a random sample being wrongly labelled for the kth class. The summation formula calculates probability for all K classes.

As with XGBoost, we perform hyperparameter tuning to optimize performance. However, due to the lack of pruning and sequential learning the available parameters for tuning are fewer.

Hyperparameters for Random Forest

Hyperparameter	Tuned value	Description
Mtry	4	" An integer for the minimum number of data points in a node that is required for the node to be split further. "
Trees	750	" An integer for the number of trees contained in the ensemble. "
Min_n	1	" An integer for the minimum number of data points in a node that is required for the node to be split further. "

Table 14: Hyperparameters for Random Forest

All descriptions are directly quoted from the Parsnip Tidymodels documentation on boosted trees (Tidymodels, 2023a).

The advantages and disadvantages of Random Forest are similar to XGBoost, being superior performance accompanied by high complexity and somewhat hard to interpret. It makes no assumptions about the distribution input data and can handle data with missing values and outliers. Random Forests also comes with a high degree of flexibility during implementation (Hengl et al., 2018).

5.4 K-Nearest Neighbours

KNN is an instance-based learning method, meaning it applies memory-based learning or what we consider “lazy-learning”. Essentially, this means that it does not build a general model from the training data, but rather compares new observations to those already in the training data (Ostfeld & Salomons, 2005).

KNN's algorithm operates by identifying the 'K' closest neighbours to a new, unseen observation denoted as x_0. These identified neighbours are collectively referred to as N_0. The measure of closeness is often determined using the Euclidean distance, a calculation of a straight line between two points in a multidimensional space. To predict the class 'j' for this new observation, KNN calculates the conditional probability, which is determined by the proportion of points in N_0 that belong to class 'j'. It then classifies the new observation x_0 to the class with the highest probability (James et al., 2013, p.39).

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \epsilon N_0} I(y_i = j)$$

Equation 7: KNN function

When deploying the KNN algorithm, the selection of K (the number of nearest neighbours to consider) serves as a critical hyperparameter that can drastically influence the performance of the model. A small value of K leads to a model with low bias and high variance, making it highly flexible but prone to overfitting. On the other hand, a large K results in a less flexible model with high bias and low variance, which might lead to underfitting as it may overlook patterns in the observations (James et al., 2013, p.39). The optimal choice of K should balance the model’s flexibility and stability, reducing the risk of overfitting while still allowing the model to capture patterns in the data.

Hyperparameters for KNN

Hyperparameter	**Tuned value**	**Description**
Neighbours (k)	19	" A single integer for the number of neighbours to consider."

Table 15: Hyperparameters for KNN

All descriptions are directly quoted from the Parsnip Tidymodels documentation on K-nearest neighbours (Tidymodels, 2023b).

The advantage of KNN lies in the model's simple nature. The model does not make any assumptions about the underlying distribution and can be robust to noise if the number of neighbours is chosen optimally. However, KNN is computationally expensive, sensitive to outliers and the scale of the data, tends to struggle with high-dimensional data, and are not suitable for imbalanced datasets (Jain, 2022). Careful pre-processing of the data is therefore often required for the model to be utilized to its fullest potential.

5.5 Evaluating Models

5.5.1 Understanding Decision Thresholds

Decision thresholds in machine learning dictate the point of separation between classes in a classification problem (Iguazio, 2023). The default treshold is normally set to 0.5, meaning that if the model predicts that the probability of an observation is above this treshold, it will classify this observation as a success. However, in situations where the dataset is imbalanced or the cost assoiciated with misclassifying a successful company as failure is very costly, we might lower the treshold to predict more succesess (Esposito et al., 2021). This decision influences the counts of correct and incorrect classification of the classes, and subsequently the calculated performance metrics.

To gain a comprehensive understanding of the perfermance of the implemented machine learning models, we will analyse them from various angles using multiple evaluation metrics. It is important to consider the stengths and weaknesses of each model, and how the model perfermance aligns with the intended objectives. The analysis of these metrics will help to make informed decisions regarding model selection and deployment.

5.5.2 Confusion Matrix

Confusion matix is valuable perfermance measurment that visualize the classification results. It highlights the amount of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the classification results (Vujović, 2021). Confusion matrices are useful to grasp the results of the classifcation and are instrumental in measuring sensitivity, precision, specificity, accuracy, and ROC curves.

		True Class	
		Negative	Postive
Predicted Class	Negative	True Negative	False Negative
	Postive	False Postive	True Postive

Table 16: Confusion matrix example

5.5.3 Prediction Accuracy

Accuracy is a commonly used evaluation metric that quantifies the overall performance of a classification problem as the proportion of correct predictions out of all predictions (Gupta et al., 2022). Of interest is the test accuracy, which reflects the model's performance on unseen data and can give an indication of how the model will perform in the real world. Conversely, training accuracy reveals how the model fares on the training data, indicating possible overfitting. However, this straightforward metric does not provide class-specific performance insights and can be misleading for imbalanced datasets. If the dataset is imbalanced, incorporating the Cohen's Kappa statistic to adjust for random chance is appropriate (Vujović, 2021). However, for balanced datasets, accuracy is an informative measure.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 8: Accuracy

5.5.4 Precision, Sensitivity & F1-score

Precision is the proportion of true positives among the total predicted positives (Vujović, 2021). A high precision score indicates that the model is highly accurate when it predicts success, but does not account for the succeses that the model missed (false negatives).

$$Precision = \frac{TP}{TP + FP}$$

Equation 9: Precision

Sensitivity, also known as recall or true positve rate, measures the proportion of true postives among the actual positive cases (Vujović, 2021). A high sensitivy score implies that the model catches a large proportion of the successes, but does not account for the succeses that were incorrectly classified (false positives).

$$Sensitivity = \frac{TP}{TP + FN}$$

Equation 10: Sensitivity

The F1 score is the harmonic mean of both precision and sensitivty, assigning equal importance to both metrics in the evaluation proccess (Vujović, 2021). However, if correctly predicting one class are of more importance, the weighted F1-score, or the ROC curve can provide additional insight.

$$F1\ Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$

Equation 11: F1 score

5.5.5 ROC Curve

The Receiver Operating Characteristics (ROC) curve is a graphical representation that showcases the trade-off between the true postive rate (sensitivity) and the false positive rate for various decision tresholds (Vujović, 2021) .Sensitivity, or the true positive rate, refers to the proportion of actual positive cases that are correctly identified, whereas the false positive rate represents the proportion of actual negative cases that are incorrectly indentifed as positive (Vujović, 2021).

$$Specificity = \frac{TN}{TN + FP}$$

Equation 12: Specificity

$$False\ Positive\ Rate = 1 - Specificity = \frac{FP}{TN + FP}$$

Equation 13: False positive rate

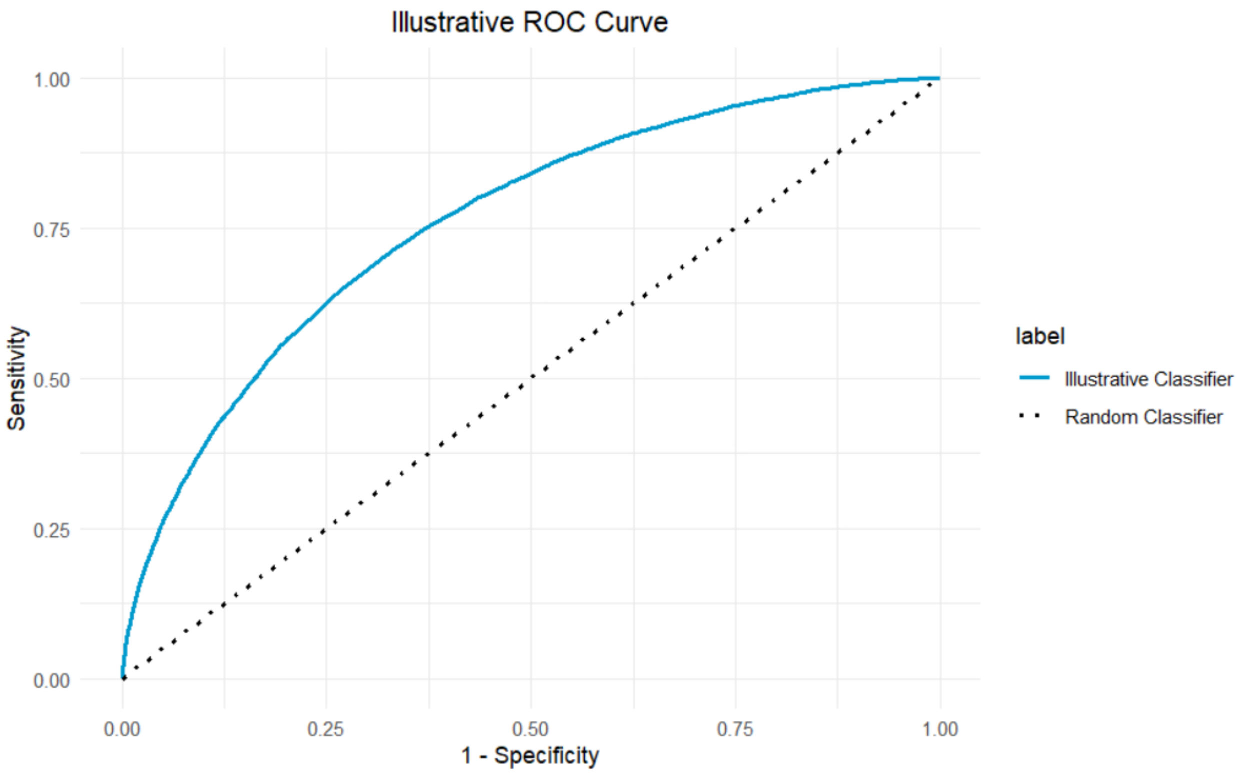

Figure 14: ROC curve example

Figure 14 illustrates the ROC curve. The curve reflects the balance between the sensitivity and the false positive rate (1-specificity) for various decision tresholds. The threshold starts from 1 at the bottom-left of the plot and decreases to 0 at the upper-right side of the plot. This is why the ROC curve is increasing. As the model becomes more sensitive and correctly indentifies a higher proportion of positive cases, it also becomes prone to falsely identifying negative cases as positive, increasing the false positive rate.

The Area Under the Curve (AUC) serves as a crucial measure of the model's discriminative power. It represents the model's ability to distinguish between successful and unsuccessful companies. A ROC curve that arches closer towards the top left corner signifies a higher AUC, implying that the model is superior in distinguishing between positive and negative class.

5.6 Feature Importance

Feature Importance refrers to the concept of assigning a score to input features based on how useful they are to predict the target variable. This can be useful for variable selection and for understanding the underlying decision-making process of the model. This is especially true for black-box models (XGBoost and Random Forest), as they do not provide any direct explanation for their predictions (Casalicchio et al., 2019).

Different models calculate the feature importance in different ways. In a Logistic Regression model, the magnitude of the coefficents are used, while the tree-based models determine feature importance based on the average decrease in impurity (or increase in information gain) brought by features used in trees (Rengasamy et al., 2022). KNN does not inherently provide feature importanc, but tecniques such as permutation feature importance, where each feature's value is randomly shuffled, and the decrease in model perfermance is mesured, can be used. However, one should note that feature importance measures are not always directly comparable across different modeles due their unique computation methods.

SHapley Additive exPlanations (SHAP), as introduced in a paper by Lundberg and Lee (2017), is another option for feature importance. This method is based on computing Shapley values, a concept from coaliitional game theory, and it provides a fair allocation of the prediction among the features. SHAP provides a unified measure of feature importance and offers more interpretability by attributing the change in the expected model prediction to each feature's contribution (Casalicchio et al., 2019).

6 Feature Engineering

Feature engineering, in the context of machine learning, broadly refers to the pre-processing step of creating and optimizing variables to enhance a model's predictive accuracy, efficiency, and interpretability (Dai et al., 2020). This process involves several interrelated tasks, each catering to a different aspect of data preparation for machine learning.

Given the extensive nature of our data collection phase, various feature engineering techniques were employed across different datasets and at distinct stages of the process. In this section, we aim to outline the most critical steps that were undertaken to ensure that our data was appropriately prepared and optimized for application in our machine learning models.

6.1 Handling Missing Values

The complete dataset, derived from platforms such as Crunchbase and LinkedIn, inherently contains numerous missing values and incomplete observations. Appropriate management of missing values across numerous and diverse variables is crucial to the integrity of predictive models (Palanivinayagam & Damaševičius, 2023). Consensus in the literature suggests that missing values should be either removed or imputed using various techniques. The choice between these two options, as well as the specific imputation technique employed, is intimately tied to the reason for the missingness (Emmanuel et al., 2021). Methodological decisions concerning the treatment of the missing data were primarily made by the degree of missingness, the relevance of the variables, and our gathered domain knowledge to determine the nature of the missing data.

A substantial subset of variables that exhibited significant missingness or were perceived to not be of high importance considering our domain knowledge of which predictors guide success were simply removed. This decision was driven by our belief that imputation of such variables could produce biased estimates and overfitting without significantly contributing to the performance of our classification problem. Additionally, we also removed 412 founders from the dataset that had no data on the personality traits because of thin LinkedIn profiles. We considered it infeasible to impute missing personality traits, as this would result in inaccurate values unsuitable for our analysis.

For variables considered to be of high importance, we adopted diverse imputation techniques depending on each variable's nature. For variables with moderate missingness, suspected to be missing not at random, we used algorithms based on other variables or our domain knowledge to impute simplified values. For example, when implementing a counting algorithm for the number of previously founded companies, many were not detected. We assumed that these missing values were due to titles not being recognized from our LinkedIn scraping, but that the true value is likely to be very low. Here, we imputed a value of 1 for the missing values as the founder at least had founded his own company and therefore could not have 0. This might distort the relative scale in instances where multiple founder titles were detected. In other cases, multiple sources were used to reduce the number of missing observations.

For other important variables such as the founder's age or the company specific variable of similar companies, we presumed the missingness to be random and utilized Predictive Mean Matching (PMM) for probabilistic missing data imputation. This imputation method replaces missing data with observed values in the dataset that have predicted values closes to the predicted value of the missing data (Buuren, 2023). We utilized both regression and polytomous logistic regression as the specified imputation model, depending on the variable type. Predictive Mean Matching is considered as a robust imputation method and are in many cases preferable over other methods because it preserves the original distribution of the data and maintains the relationships between variables (Akmam et al., 2019). However, the model renders careful use as it can cause problems of many duplicates if the missingness is extensive and does not work well if the sample size or the number of predictors is too small (Buuren, 2023).

Addressing missing values in our dataset was a complex task. In ensuring the reliability of our predictive models, we balanced the need to retain as much information as possible against the risk of introducing bias through imputation. During the process, we leveraged the advanced

XGBoost model that is proficient in handling missing values and discerning patterns within them (Rusdah & Murfi, 2020). After comparing various models with different levels of imputation and deletion, we concluded that a modest decrease in test accuracy of about 1-2% was an acceptable trade-off for a much simpler, more interpretable model that required fewer imputations to address the missing data. A table of the imputation methods used can be found in the Appendix.

6.2 Correlation Considerations

The statistical measure of correlation quantifies the degree to which two or more variables move in relation to each other and provide guidance on feature selection and in ensuring the robustness of machine learning models (Hall, 1999). In the context of predicting founder success, correlation between a particular feature and the target variable are typically valuable for prediction, hence their inclusion in our model. Nevertheless, we exercised caution in incorporating variables that correlate for reasons unrelated to inherent startup performance or variables that provide information not available at the founding date. Examples of such variables are annual financial data, the web traffic the last six months, and the number of followers a founder has on their LinkedIn profile today. These types of variables are directly or indirectly dependent of the target variable, and subsequently left out of our final set of predictors.

A high degree of correlation between features, a phenomenon known as multicollinearity, can be challenging as it may lead to overfitting and complicate the interpretability of features (Chan et al., 2022). Certain machine learning models, such as Extreme Gradient Boosting and Random Forest, are generally robust against multicollinearity issues, while simpler models like Logistic Regression are not (Sahani & Ghosh, 2021). However, despite some models being resistant to problems associated with inter-correlated features, it can limit the interpretability of those features individually. For instance, a model might overemphasize one of the correlated variables, which can obscure the significance of the other individual feature. We avoided these potential issues by eliminating many highly correlated variables, acknowledging the trade-off between model accuracy and interpretability. For example, the calculated variable of a founder’s years of experience at the founded date was highly correlated with the age of the founder and was therefore excluded for interpretability purposes.

Nevertheless, some correlated variables are still present in our final set of predictors. We opted not to remove any of the correlated Big Five traits. This decision was predicated on the unique insights each trait could offer into startup success.

However, we did strive to limit the number of correlated personality traits where it was reasonable. We excluded 11 behavioural traits that were somewhat related to the Big Five framework, both in theory and in our data.

Moreover, the final stage of the feature selection process was predominantly iterative, involving repeated testing and adjustment of the machine learning models, and systematically eliminating correlated variables. This approach made the application of principal component analysis and other dimension reduction techniques less relevant for us.

6.3 Creation and Transformation of Variables

New variables can be derived from existing ones to capture more specific and nuanced information. This can be particularly useful when the raw data would not capture the underlying patterns (Verdonck et al., 2021). For example, we processed the lists of schools attended by a founder and the list of jobs they held to calculate the total unique number of schools, jobs, and previous founder titles. Other variables were transformed to reflect the information when the company was founded, such as the founder's age.

Nominal categorical variables were refined and reduced in the number of categories. For instance, the NACE categories were initially extremely detailed categorized by Brønnøysund Register Centre and spanned over 197 distinct categories. Leveraging data from Statistics Norway (SSB, 2023), we grouped these into 21 broader categories, and then further condensing them to 7 groups by accommodating all low-frequency observations into an "other" category. Similar categorization refinements were also applied to the county of both company and founder, and the country of the founder. Reducing the number of categories to fewer, more meaningful categories can potentially reduce overfitting and make the model easier to understand and interpret (Verdonck et al., 2021).

6.4 Outliers

Outliers can considerably affect the outcomes of data analysis, especially in linear models that are vulnerable to extreme values (James et al., 2013, p.97). They can distort patterns and lead to misleading interpretations. Outliers were treated with a methodical approach backed by our in-depth domain knowledge.

We thoroughly examined each outlier, determining its validity and cause. If we identified an outlier as a result of an error, we opted for its removal or modification, where suitable. However, when outliers represented true data points, we primarily retained them, as they contribute valuable information. This approach is particularly justifiable for models that demonstrate resilience to outliers. For distanced based models, such as Logistic Regression and K-nearest neighbours, we removed most extreme outliers. In all instances, we made these decisions carefully, balancing the need for accurate representation with the desire for robust and reliable models.

6.5 Encoding Categorical Variables

The transformation of categorical variables into numerical values is a fundamental requirement for machine learning algorithms, a process known as encoding (James et al., 2013, p.83). One-Hot Encoding (OHE) was employed for all nominal categories, which do not bear a particular order of importance. OHE works by allocating a binary value to every category depending on its presence or absence in a given row.

Nonetheless, using OHE can inadvertently result in the "curse of dimensionality," a scenario where high-dimensional data can hamper computational efficiency and obscure pattern recognition (James et al., 2013, p.107). While alternative encoding strategies, such as Frequency Encoding or Target Mean Encoding, will mitigate this issue, their impact was negligible in our case due to our careful limitation of the number of categories within the variables.

Additionally, in the context of linear models, which are not inherently equipped to handle multicollinearity, we strategically omitted one category from each set of dummy variables. This approach effectively helped us to sidestep the “dummy variable trap”, ensuring the robustness of our model against potential data-induced biases (Fernandes, 2022).

6.6 Feature Scaling

Scaling ensures that the differing scales or units across features are on the same scale, which can avoid bias in the model’s learning process (Ozsahin et al., 2022). The two most common scaling methods are standardization and normalization. We utilized standardization, demonstrated in Equation 14.

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

Equation 14: Standardization

Standardization transforms the data to have a mean of zero and a standard deviation of one. This process is important for machine learning models that rely on distance between data points to make predictions, such as logistic regression or K-nearest neighbours. If the variables are not scaled, variables with higher values will have higher impact on the predictions (Peterson, 2009). Additionally, tree-based methods can also benefit from standardization, as it could improve the stability and speed of learning.

6.7 Pre-processed Dataset

The careful feature engineering process leaves us with a pre-processed dataset of 18 variables. Four variables are company-specific, while the rest are founder characteristics and personality traits. The variables are chosen to represent information available at the earliest stages of a startup.

Number of rows: 2413
Number of columns: 18
Number of NA: 0

Variable	Explanation	Type
[1] "County_c"	County where the company is located	categorical
[2] "Num_Founders_c"	Number of estimated founders	integer
[3] "Similar_Companies_c"	Number of similar companies	integer
[4] "Age_f"	Age of the founder when company was founded	integer
[5] "Num_roles_proff_f"	Number of active roles a founder is involved in	integer
[6] "County_8_f"	County where the founder is located	categorical
[7] "Country_Norway_f"	If the founder is living in Norway	categorical
[8] "Num_Schools_f"	Number of schools a founder has attended	integer
[9] "Num_Jobs_f"	Number of jobs the founder has had	integer
[10] "Num_Founder_Titles_Work_f"	Estimated number of previously founded companies	integer
[11] "Openness_score"	Openness score 1-10	numeric
[12] "Extraversion_score"	Extraversion score 1-10	numeric
[13] "Emotional_Stability_score"	Emotional Stability score 1-10	numeric
[14] "Agreeableness_score"	Agreeableness score 1-10	numeric
[15] "Conscientiousness_score"	Conscientiousness score 1-10	numeric
[16] "Gender_f"	Gender of founder	categorical
[17] "NACE_group_7_c"	Main NACE industry of the company	categorical
[18] "Success_Target"	Target, high success=1, low success = 0	categorical

Table 17: Pre-processed dataset

7 Results

In this section we will present the results of our machine learning methods as described in our methodology section. Firstly, we will present the models' performance for our evaluation metrics, then we will investigate the results of each model and analyse our findings. Afterwards, we will compare the results of each model and make a case for the model we find to be the greatest performer.

7.1 Performance Evaluation

To apply our evaluation methods, we divide our data into a training and a test set. The split is done by dividing 80% to the training set and 20% to the test set using our target variable as stratification variable. Stratification is a proactive statistical method ensuring that distinct sub-populations are adequately represented within the final selection and treatment groups (The World Bank, 2023). The resulting training set consists of 1929 founders while the test set consists of 484 founders. The training set will be applied to train the machine learning models without any insight into the test set, afterwards the test set will be applied to calculate our evaluation metrics. The separation of test and training data is important to predict previously unseen data and can simulate real-life applications (James et al., 2013).

7.1.1 Prediction Accuracy

Accuracy			
Logistic Regression	XGBoost	Random Forest	KNN
0.616	0.624	0.622	0.591

Table 18: Accuracy results

Among the models, Random Forest and XGBoost demonstrate the highest test accuracy, with respective values of 62.4% and 62.2%. Given the relatively balanced target variable, accuracy is a meaningful measure of model performance. Logistic Regression, despite being a simpler model than its tree-based counterparts, is closely competing with an accuracy of 61.6%.

All models significantly surpass the baseline accuracy of 51.6% achieved by always predicting success. Considering the multitude of factors beyond the founding team influencing startup success (Sevilla-Bernardo et al., 2022), the models' accuracies are acceptable. However, accuracy alone does not encapsulate the full performance spectrum of these models.

7.1.2 Confusion Matrix

Considering the confusion matrix for each model can provide valuable insights into how the models classify founders into the two success categories at the standard decision threshold of 0.5. Out of 484 founders in the test set, XGBoost correctly identifies 190 founders as highly successful, with only 60 misclassified as low success. This indicates that XGBoost is particularly proficient at identifying successful founders, delivering a sensitivity of 76%. However, it is important to note that this also results in a substantial number of false positives, leading to lower precision of 60.9% and, as such, is outperformed by both Random Forest and Logistic Regression on this metric.

	Predictions							
	Logistic Regression		XGBoost		Random Forest		KNN	
Truth	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
False	126	108	112	122	130	104	124	110
True	78	172	60	190	79	171	88	162

Table 19: Confusion matrix results

	Logistic Regression	XGBoost	Random Forest	KNN
Precision	0.614	0.609	0.622	0.595
Sensitivity	0.688	0.760	0.684	0.648
F1-Score	0.649	0.676	0.651	0.621

Table 20: Precision, sensitivity, F1-score results

7.1.3 ROC and ROC AUC

The ROC curve is a robust metric for model evaluation, showcasing the model's classification proficiency across all decision thresholds. Considering the Area Under the ROC Curve (AUC) it is the Random Forest model that stands out. The higher score shows that Random Forest exhibit greater classification abilities across all thresholds. This implies that by careful selection of the decision threshold, the Random Forest model may surpass XGBoost in sensitivity. Logistic Regression have ROC close to XGBoost, and once again, it is worth noting that KNN falls behind in comparison to the other models.

ROC AUC

Logistic Regression	XGBoost	Random Forest	KNN
0.642	0.648	0.667	0.600

Table 21: ROC AUC results

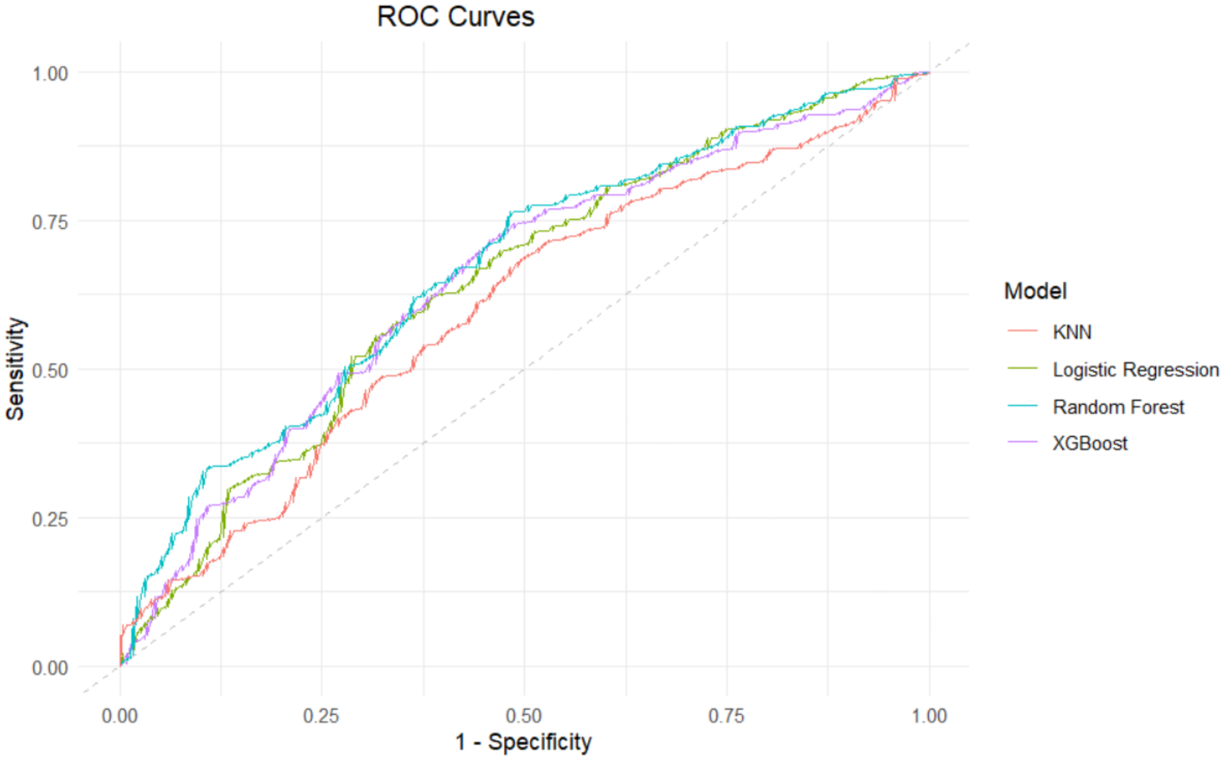

Figure 15: ROC curves for all models

7.1.4 Summarizing Results

Investigation of the performance metrics suggests that, for predictive purposes, the tree-based methods demonstrate superior performance. XGBoost marginally demonstrate the highest accuracy and showcases a high sensitivity at the set 0.5 decision threshold, while Random Forest exhibit a greater potential for adaptability to modifications with a higher ROC AUC. However, except for KNN, which falls a tad behind, the performance of the models is comparable. Despite being a simpler model, Logistic Regression should not be disregarded as it offers significant interpretability, a point we will delve into in the following section.

	Accuracy	ROC AUC	Precision	Sensitivity	F1-Score
Logistic Regression	0.616	0.642	0.614	0.688	0.649
XGBoost	0.624	0.648	0.609	0.760	0.676
Random Forest	0.622	0.667	0.622	0.684	0.651

Table 22: Summary of evaluation metrics

7.2 Feature Analysis

Interpretability of machine learning models refers to the ability to explain or present the model in understandable terms to a human (Doshi-Velez & Kim, 2017). For business case applications, it is crucial to understand how the model is making the predictions, thus, interpretability becomes a part of one of the metrics for choosing the appropriate model. Here, we will consider the feature importance of the models, which helps us to estimate how much each feature contribute to the model's prediction. KNN will not be considered due to its reliance on distance metrics rather than individual features, rending such analysis less informative and computationally expensive.

7.2.1 Logistic Regression

The simplistic nature of Logistic Regression assumes linear relationships and independent variables, which means that coefficients can be interpreted as indicators of feature importance. The coefficients of the predictors can be interpreted as the expected change in log odds of the outcome (success) per unit change in the predictor. Increasing the predictor by one-unit multiples the odds of success by e^{β} (odds ratio).

Feature	Standardized coefficient estimate	Unstandardized coefficient estimate	Unstandardized odds ratio	p-value
Num_Founders_c	0.178	0.179	1.196	0.0003
Conscientiousness_score	0.176	0.133	1.142	0.0099
Similar_Companies_c	0.122	0.017	1.017	0.0174
County_8_f_Rogaland	1.308	1.308	3.698	0.0439
Num_roles_proff	0.104	0.019	1.019	0.0445

Table 23: Top five most significant variables for logistic regression

We observe that only five features have a p-value under 5%, indicating their statistical significance. Higher values for these features correspond to increased odds of success. For instance, an increase in the number of founders from 1 to 2 multiplies the odds of success by 1.194 (the odds ratio), assuming all other variables are held constant. This indicates that each additional founder increases the odds of success by about 19.6%, a relative change. It is important to clarify that this influences the odds of success rather than the direct success rate.

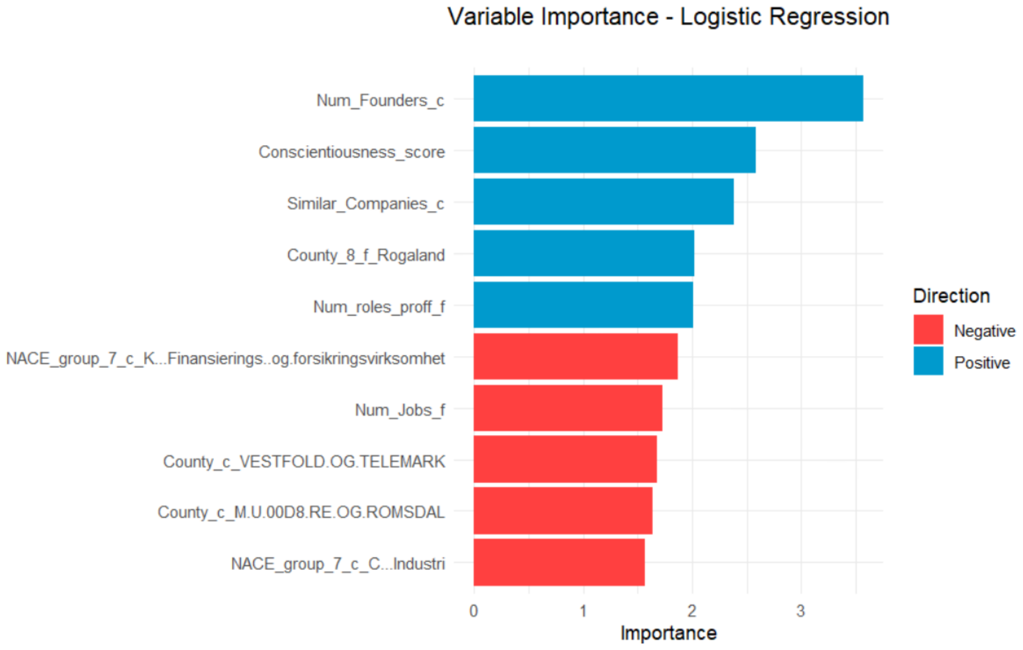

Figure 16: Variable importance for logistic regression

The variable importance plot displays the top 10 predictors in the Logistic Regression model. Interestingly the Big Five personality trait conscientiousness are ranked as the second most important feature in predicting success, right behind the total number of founders. The model detects a positive linear relationship between the odds of success and the conscientiousness score of a founder. Furthermore, we see that a higher number of similar companies increase the likelihood of success.

Skepticism towards the preference of a founder being from Rogaland is warranted. As described in the exploratory data analysis, this could simply be a consequence of a biased data set stemming from relatively few observations combined with oil and gas clustering in the area (Gjerde, 2023).

We observe that a higher Number of Proff roles predicts success, a possible indicator of more experience leading to success. On the other hand, unexpectedly, the model finds a negative relationship between the number of jobs of a founder and success, however, this is not as significant with higher p-value of 8.35 %.

The Linear Regression model allows for precise evaluation of model coefficients and p-values. We can assert with high confidence that the number of founders of a company and the founders' conscientiousness score contribute to success. However, several variables are insignificant or lack evidence of causal effect. One limitation of Logistic Regression is its linearity. Significant variables with non-linear relationships to the target variable might be overlooked by this model.

7.2.3 Tree-based Models

Interpreting feature importance in complex tree-based models like XGBoost and Random Forest can be challenging due to interaction between features. However, it can provide insightful understanding on how these models "think", though their individual feature importance measures are not directly comparable to one another. We include some discussion involving literature in this section, as a measure of identifying possible overfitting. We argue the inclusion of this discussion is necessary in this section rather in the later Discussion chapter, as we want to remove these possible false predictors from our presented findings.

XGBoost

The feature importance plot in Figure 17 reveals that XGBoost and Logistic Regression attribute importance to similar features in making predictions. Notably, we find that a higher count of founders, roles, similar companies, and a higher conscientiousness score contribute positively to the prediction of success. Moreover, age emerges as the second most influential feature and other personality traits are also prevalent amongst the most important predictors, hinting at the possibility that XGBoost could be detecting some non-linear relationships that Logistic Regression might overlook. Following, we will present some of these relationships and their perceived validity.

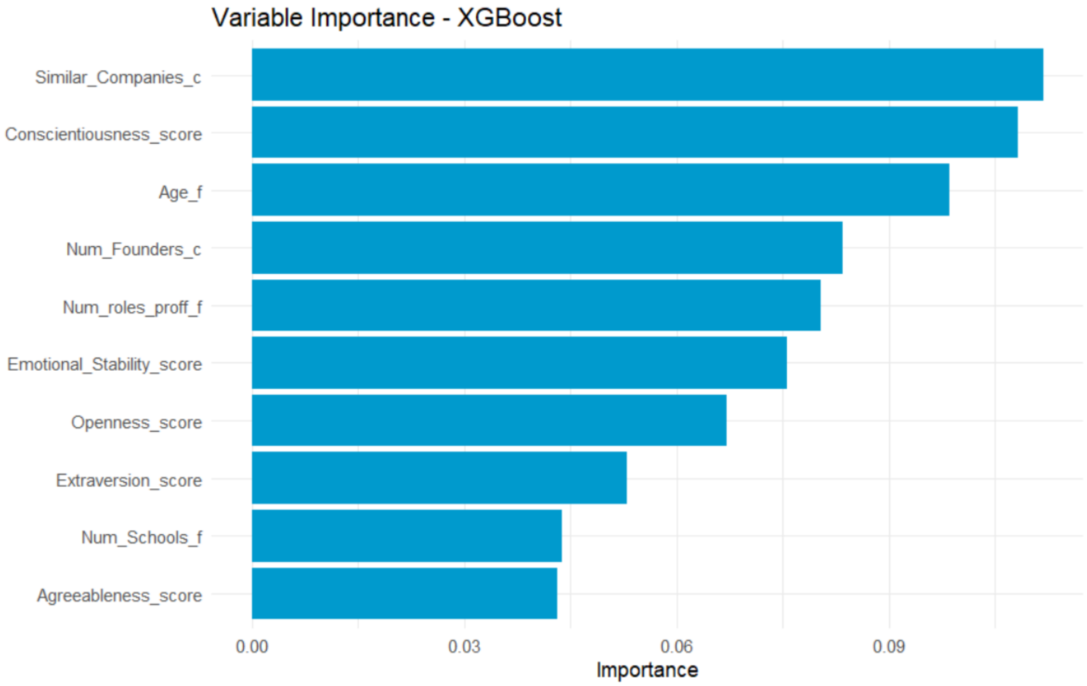

Figure 17: Variable importance for XGBoost

SHAP dependency plots provide a deeper look into the relationship between the features and their impact on model predictions by visualizing how the model output varies with feature values, enabling us to interpret how XGBoost perceives these features.

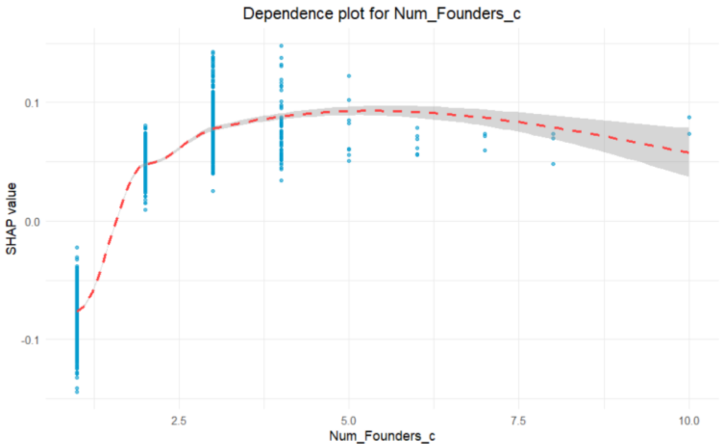

Figure 18: Dependency plot for number of founders (XGBoost)

For the fourth most impactful variable, the number of founders for a company, it is prevalent that XGBoost can identify non-linear patterns. The insightful finding here is that a single founder is clearly worse than multiple founders. However, we must acknowledge that only 2.5% of companies have more than three founders.

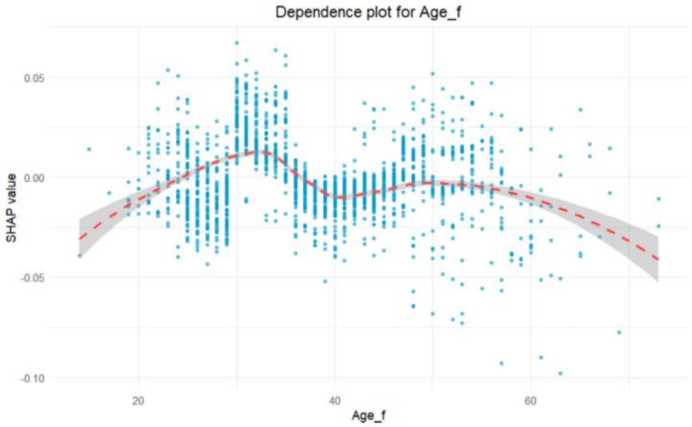

Figure 19: Dependency plot for age of founder (XGBoost)

We observe that XGBoost detects a non-linear pattern for age and places it highly as the third most important predictor. However, by plotting the dependency plot, a complex relationship occurs. As expected from our literature review and the descriptive analysis, entrepreneurs aged 25 and below appear less likely to succeed. But rather unexpectedly, the age group rising as the most likely to achieve success, would be those in their mid-thirties. As seen in Figure 9 in the descriptive analysis, the data appears noisy and complex, and the patterns detected by our XGBoost method is therefore prone to overfitting. We are therefore skeptical to our findings regarding age, especially since it contradicts established research concluding that middle-aged entrepreneurs are the most prosperous in terms of success (Azoulay et al., 2020)

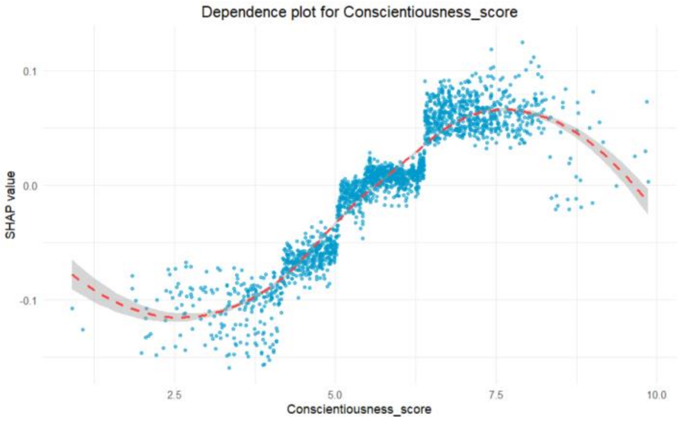

Figure 20: Dependency plot for conscientiousness (XGBoost)

For conscientiousness, besides the few observations at the high and low end of the score, a straight line could be well fitted. The linearity of this graph can help explain why Logistic Regression also picked up on this trait. This result further solidifies the findings of Zhao et al. (2010) and John and Srivastava (1999).

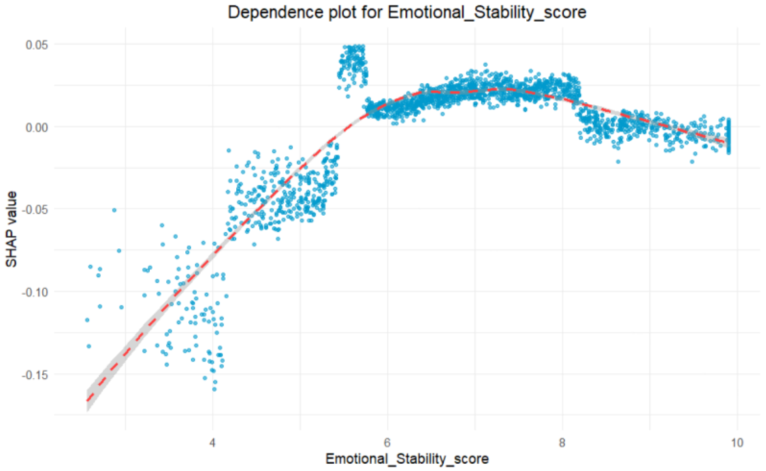

Figure 21: Dependency plot for emotional stability (XGBoost)

The SHAP dependency plot reveals clear non-linear patterns for emotional stability. For each prediction, an emotional stability score below 6 negatively affects the likelihood of success. Conversely, while high scores are generally beneficial, they start to slightly reduce the prediction of success for scores above 8. This observation suggests a nuanced relationship, where struggling to manage negative emotions can be detrimental to success, excessive emotional insensitivity may also be sub-optimal. This exact pattern is unprecedented in literature and must be viewed with caution. However, higher emotional stability as a predictor for entrepreneurial success aligns with the meta-analysis by Zhao et al. (2010).

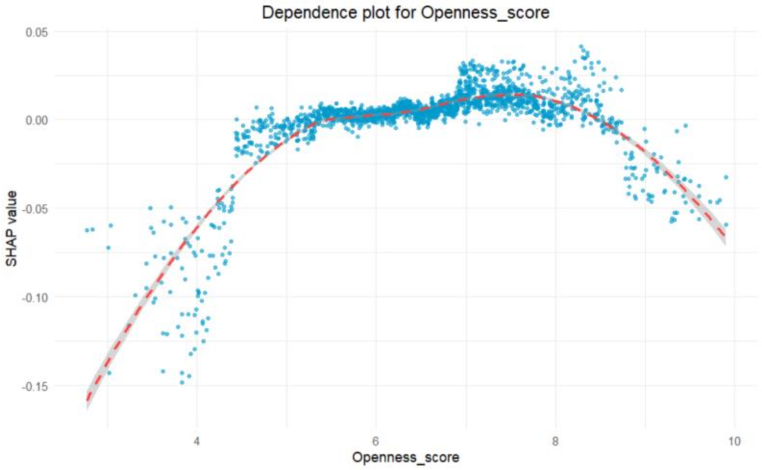

Figure 22: Dependency plot for openness (XGBoost)

Figure 22 illustrates XGBoost's interpretation of the personality trait openness as depicted in the dependency plot. From our literature review it became apparent that openness was the number one personality trait predictor for entrepreneurial success (Zhao et al., 2010). Our XGBoost model seems to disagree regarding this importance as it places it seventh. Still the pattern described mostly aligns with the literature as higher values of openness predicts higher success probability. However, as the openness exceeds a score of 8, we observe a negative trend, theorizing that people who are too open can have some disadvantages like being perceived as unpredictable or lacking focus. Still, as XGBoost finds this trait to be somewhat of less importance and there are no earlier studies backing this claim, we must approach these results with caution.

Random Forest

Random Forest attributes high importance to all the Big Five traits (Figure 23). While this is highly interesting, it is hard to interpret exactly why the other traits are deemed less important by Random Forest.

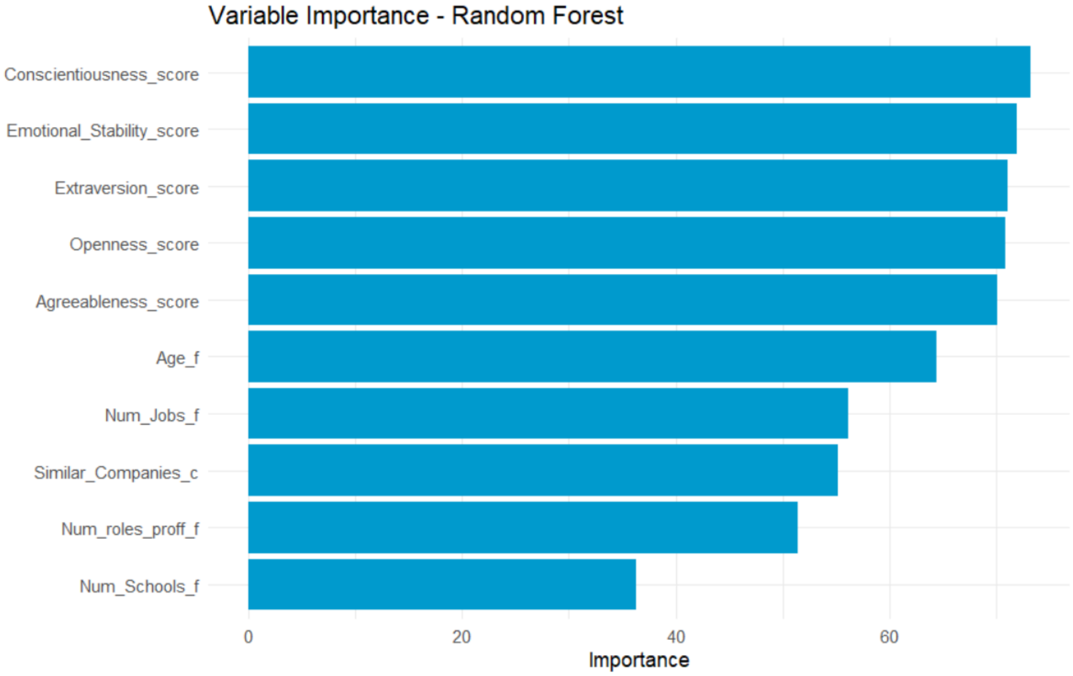

Figure 23: Variable importance for Random Forest

Inspection of the dependency plots reveals almost identical non-linear trends as observed in XGBoost. Consequently, we will only show extraversion and agreeableness.

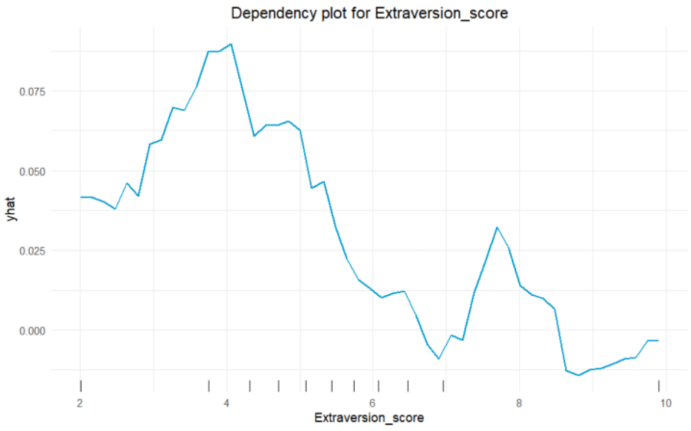

Figure 24: Dependency plot for extraversion (Random Forest)

Contrary to existing literature which often associates higher levels of extraversion with the success of founders (Zhao et al., 2010), our Random Forest model tends to favour lower extraversion scores for success. Furthermore, we observe peculiar non-linear patterns with a peak appearing around the scores of 7-8. While XGBoost finds a similar pattern, asserting this as a credible pattern is overly speculative, given the inconsistency with previous findings in the literature.

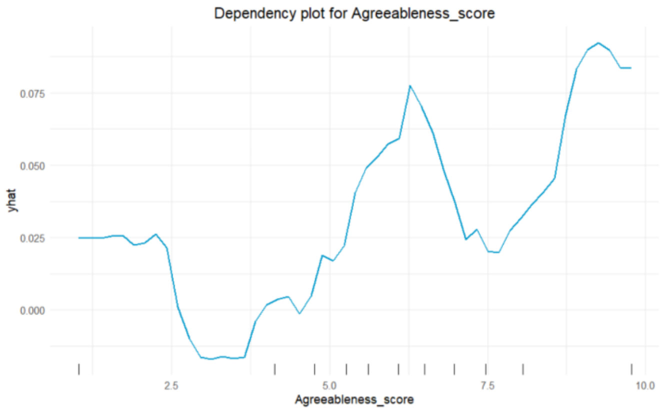

Figure 25: Dependency plot for agreeableness (Random Forest)

Likewise, the model hints that high agreeableness scores correlate with success, a proposition that greatly contradicts the finding of Zhao et al. (2010). Thus, the distinct patterns in this plot also call for considerable scepticism.

7.2.4 Summary of Feature Analysis

Table 24 highlights the five most important features for each model.

Feature Rank	**Logistic Regression**	**XGboost**	**Random Forest**
1	Number of Founders	Similar Companies	Conscientiousness
2	Conscientiousness	Conscientiousness	Emotional Stability
3	Similar Companies	Age	Extraversion
4	Founder from Rogaland	Number of Founders	Openness
5	Number of Roles Proff	Number of Roles Proff	Agreeableness

Table 24: Top five most important features for each model

Conscientiousness is deemed important by all three models. Furthermore, Logistic Regression and XGBoost seems to agree that the number of founders, the number of similar companies and the number of roles on Proff are important. XGBoost hints that there might exist non-linear patterns for the age of the founders, while Random Forest are all in on the Big Five traits.

However, diving deeper into the dependency plots for the tree-based models reveals clear non-linear patterns that are not supported by the literature nor logic sense. It is hard to ignore the obvious problem: XGBoost and Random Forest is overfitting to the personality traits. This is further supported by the fact that removal of these variables did not drastically reduce the model performance. This does not necessarily mean that these traits are not important at all and that no non-linear patterns exist, but that their effect is less impactful than what these models "think". This showcases the interpretability problems of these black box models. We might see the impact of too few observations in highly variable data.

7.3 Comparing Models

Choosing the best model will depend on the performance, interpretability, and the use case.

In the performance assessment of our selected machine learning models, there are a few factors to consider. At the set decision threshold of 0.5, XGBoost demonstrate an exceptional ability to identify successful founders (high sensitivity), but at the cost of more false positives (low specificity). From the perspective of venture capital firms, the potential benefits of accurately classifying founders as successful often outweighs the risk of mistakenly classify a successful founder as a failure, making high sensitivity-models more attractive. However, it is important to recognize that different firms may have varied risk tolerance and investment strategies, which could influence their readiness to accept this trade-off. However, Random Forest showed to outperform all models on ROC AUC. This implies that by careful selection of the decision threshold, the Random Forest model may surpass XGBoost in sensitivity, and exhibit greater potential for adaptability to modifications.

Considering the other side of the equation, the interpretability of the model, Logistic Regression is the clear winner. While the tree-based model can find complex non-linear patterns and interactions between features, they are hard to interpret and show signs of overfitting to the personality traits. The famous quote "all models are wrong, but some are useful" by George Box in 1976, holds true in this case as well. Despite being slightly behind the tree-based methods in classification performance, its interpretability and simplicity will make the Logistic Regression model more useful for the purpose of understanding the variables and reliable in real world applications. Logistic Regression is therefore our final model choice and will serve as the deciding tool when presenting findings.

8 Discussion

As we embark on the discussion of our findings, it's pivotal to reintroduce the research question that has guided this study:

"How can machine learning be utilized to predict success of early-stage Norwegian startups and to identify founder characteristics that contribute to success".

We will first answer the second part of the question, namely the founder characteristics we have found to contribute to success. The discussion will be conducted by firstly summarizing our findings, then interpreting our results and connecting our findings to previous research. The first part of the research question will then be discussed by presenting the possible implementations of our model and the utility we argue our model can provide for Norwegian VC companies and other stakeholders in the Norwegian startup landscape. We will also discuss how the limitations of our thesis may have affected our results, and how future research can leverage the knowledge gained from this thesis for further studies. Finally, we will delve into certain ethical considerations that were integral during the process of composing this master's thesis.

8.1 Summary of Findings

From our results, two predictors arise as the most prominent as characteristics of successful founders. Firstly, they tend to appear in founder teams rather than as solo founders. This finding was consistent with our best model for interpretability, logistic regression, and the tree-based model XGBoost. From the logistic regression, the coefficient p-value for number of founders is 0.0003, meaning the model predicts a positive correlation between number of founders and success with a very high significance. Secondly, we find successful founders tend to have higher Conscientiousness than their less successful counterparts. This result is reoccurring in the tree-based models and our Logistic regression model with a p-value of 0.01, which again proves significance beyond doubt.

Additionally, some promising indicators of success were also found. However, these findings are lacking in evidence for this thesis to acknowledge any significance. These were as follow: number of similar companies, number of roles Proff (Administrative roles in Norwegian companies), Extraversion, Emotional stability, and Age.

Number of similar companies has a coefficient p-value of 0.017 from the logistic regression, thereby claiming significance, indicating that a higher number of similar companies relates to a higher probability of success. This is not necessarily intuitive, as one should expect a higher number of similar companies means higher competition and less uniqueness in their product and services. On the other hand, it could be an indication of partitioning in a trending and profitable sector. However, this variable is calculated by Crunchbase's machine learning model to compare attributes like industries, company description, and financials (Crunchbase, 2022), which can result in biases towards more successful clusters of companies. Claiming any significance regarding predictability of success for this variable is therefore unfounded.

The number of roles Proff variable is also significant according to the regression output with a p-value of 0.0445, barely below the significant 0.05 threshold. An increased number of Proff roles of a founder predicting success may reflect their experience and expertise, which are linked to higher success (Azoulay et al., 2020). However, as this metric is simply an estimation of experience and is based on current data, this variable is unfortunately distorted by present-bias and can therefore not claim significance for prediction purposes.

The three variables Openness, Emotional stability and Age all aligns with earlier research to some degree, with higher Openness and Emotional stability being established predictors for entrepreneurial success (Zhao et al., 2010). Furthermore, Age has been proven as a predictor for success with middle-aged entrepreneurs arising as the top performers (Azoulay et al., 2020). However as observed in figure 19, 21, and 22 and explained in the result section, all these traits predict complex patterns unsupported by earlier research and due to the high variance and the relatively small dataset, the likelihood of overfitting cannot be ignored. The determining factor is that no significance is found for these variables in our most trusted model, the logistic regression, and as we have observed signs that our tree-based models are prone to overfitting, we are therefore reluctant to claim any significance.

However, all the above-mentioned promising indicators show signs of predictive ability, which can serve as an indication towards that abroad entrepreneurial literature can also be applicable in the Norwegian startup landscape for these variables. Regardless, further studies would be needed to prove any validity. To summarize, Conscientiousness and number of founders are the only variables which this thesis deems to significantly prove success amongst Norwegian entrepreneurs.

8.2 Interpreting and Contextualizing Findings within Existing Research

Our findings can be interpreted such that Conscientiousness is the greatest Big Five personality trait to apply as a predictor for success amongst Norwegian entrepreneurs. This is inconsistent with the meta-analysis of Zhao et al. (2010), as Conscientiousness is regarded as the second greatest predictor for entrepreneurial success just behind Openness. Contrastingly, Openness shows no significant importance in the determining model of logistic regression. Still the metanalysis by Zhao et al. (2010) does confirm our finding that Conscientiousness is a credible predictor for entrepreneurial success. Furthermore, the prediction of conscientiousness as the most important trait, is supported by research showing that this trait is superior in predicting high work performance across all dimensions compared to the other four traits (John & Srivastava, 1999). We can further interpret these results by that Norwegian entrepreneurs who possess a high Conscientiousness are more likely to succeed. Meaning that individuals showing personal characteristics like goal-oriented behavior, good organization and impulse control are more likely to succeed as entrepreneurs.

We can interpret our findings regarding number of founders such that it is beneficial to start a company alongside someone else in Norway, as this is more likely to produce a successful outcome. The number of founders as the most important predictor in the logistic regression model aligns with the current common perception amongst investors, that more founders generate more valuable companies (Miloud et al., 2012). This thesis's finding thereby contradicts the recent study done by Greenberg and Mollick (2018), where they found that solo-founders where as successful and possible more successful than founder teams.

A short mention of the lacking significance of certain variables in any of the models is necessary. Rather unexpectedly, no significance whatsoever was found for the variable Number of founded companies, suggesting that being a serial entrepreneur is not a reasonable predictor of success in the Norwegian Startup landscape. This is consistent with the study of Westhead and Wright (1998), who found there to be no significant difference in performance between novice and serial founders. Furthermore, we found no indication that gender was a significant factor in predicting success from a founder perspective. These findings support the claim that the bias existing towards female founders is ungrounded (Kanze et al., 2018).

8.3 Potential Impact and Utility of Our Thesis and Machine Learning Models

In this section, we will introduce two potential usage cases that could benefit and foster growth within the Norwegian startup ecosystem. Firstly, the thesis as a guidance tool for entrepreneurs, investors, educators, policymakers, and the academic community in the Norwegian startup landscape. Secondly, the usage of machine learning as a tool for early-stage founder assessment for venture capital firms.

8.3.1 Use case 1: Guidance Tool for Stakeholders in the Norwegian Startup Landscape

Our research and findings serve as a bridge to fill the literature gap of which founder characteristics and personality traits make up founders for scalable, tech-driven Norwegian startups and which that are associated with success. The information in this thesis provides valuable insights that can be used to further foster growth of Norway’s startup scene.

Existing and upcoming founders can increase their understanding of the predictors of success, guiding them in their strategic decisions like the hiring process, forming initial teams and be aware of their own characteristics and personality traits that both help and hurt them.

Venture capital firms and other investors can use apply insights from this thesis in due diligence processes. They can expand their knowledge and refine their investment strategies to be better aligned with recent findings. Furthermore, they can be motivated to consider personality traits to a greater extent and analyse these scientifically.

It can benefit educators, such as support organizations and incubators, to help potential founders reach their goals in more effective ways, by focusing on and enhancing important traits. Though personality is hard to change, one can improve behaviour and habits related to conscientiousness and emotional stability, such as clear goal planning and incorporating mediative practises.

Policy makers can use this thesis alongside existing literature to form policies and initiatives with a more nuanced understanding of what drives startup success in Norway, creating a more productive entrepreneurial ecosystem with more diversity. One example could be substituting programs like start-up labs and other social events for entrepreneur so that they can get familiar and cofound companies together and thereby increasing success as a founder team.

For the academic community, this thesis contributes to expanding the body of literature on founders and startups in Norway, enriches the collective understanding and motivate a clear path of further research. Especially, with the use of the Big Five model and with the use of new emerging AI technologies.

8.3.2 Use case 2: Implementation of ML for Early-Stage Founder Assessment

Our machine learning model presents a strategic tool that VC firms can leverage to assess the potential success of founders at the very earliest stages. The implementation could help the screening process, provide insightful decision-making data, and ultimately lead to higher investment returns.

For this implementation to be possible, VC firms would need to integrate the machine learning model with databases of startups and founders. These databases could be startups that they are already tracking or external databases. This integration could be achieved through an Application

Programming Interface (API), facilitating the automatic exchange of data between the model and the databases.

On a regular basis, depending on the firm's needs, the model would process the latest founder characteristics from LinkedIn, predicting which founders are likely to lead to successful startups.

The output would feed into a flagging system, identifying startups and founder teams worth focusing on. Traditionally, this process has been labour-intensive, often requiring investors to investigate large amounts of data and spend considerable time meeting with founders. The flagging system should not replace these tasks but can save the investor time and maximize where they spend their resources. By flagging potential success stories at an extremely early stage, our model could provide investors with a critical head start, allowing them to initiate discussions with founders before competitors become aware of the opportunity. It can also supplement investors' intuition and industry knowledge with empirical evidence, reducing the risk of overlooking promising opportunities or investing in less viable ones.

The firm can adjust the threshold based on their risk tolerance, time availability for meetings, and other factors. A lower decision threshold would flag more startups for further investigation but would also increase the chance of false positives. Conversely, a higher threshold would result in fewer flagged startups but with higher likelihood of success.

Besides the model's predictive power to guide investors in the right direction, the simplicity of the model could be of great benefit. With few inputs and a transparent algorithm, the model can provide insights into why a founder has been flagged as promising. This allows the investors to not only understand the predictions but also interrogate them critically. For example, they might probe deeper into certain characteristics during founder meetings or conduct additional due diligence on certain areas. Additionally, personality insights could lead to better communication with promising founders. These insights can be invaluable, merging machine intelligence with human judgement. The model could also be customized and iterated upon as the venture capital firm's needs to evolve. For example, the firm could decide to add more variables.

The application of our machine learning model has the potential to help venture capital practices to adapt. By efficiently parsing through vast amounts of data, our model serves as a strategic tool, identifying promising founders and startup teams much earlier than traditional methods. Despite its predictive power, the model doesn't eliminate the need for human expertise, but instead enhances it. It provides investors with invaluable insights, enabling them to delve deeper during founder meetings, enhancing communication, and aligning the investment focus with empirical evidence. The model's customizability allows it to remain an evolving tool, flexibly adapting to the changing needs of the firm. The fusion of machine intelligence with human judgment embodies the future of venture capital investing - more informed, efficient, and effective.

8.4 Limitations and Further Research

This thesis has several weaknesses and limitations that could influence the validity and precision of our machine learning model's performances.

There are multiple potential biases within the data that could influence the results of our machine learning models. As described in the data section, the structure of the dataset poses several challenges. Incorporation of company-level data to rows of individual founders disrupts the assumption of data independency within rows, thereby potentially affecting the results of our model. Moreover, this decision can exaggerate the role of individual founders in predicting startup success, as the total number of founders in each company is not considered. These concerns should be taken seriously, and implementation of hierarchical models and incorporating weights into the models could limit these problems. As the dataset becomes more complex with more company-level data, there arises a compelling need to explore alternative structures and more sophisticated models. While such models did not fit our purposes in this thesis, it would be interesting to see how the results differ.

Our data sources may carry inherent biases leading to imprecise results. Crunchbase tends to favour active and successful companies in their database, thereby not representing a broad spectrum of unsuccessful startups. Our effort to precisely categorize companies based on their level of success might not overcome the homogeneity of the sample. A too homogeneous pool of startups can obscure the nuanced differences in founder characteristics between the success categories. Furthermore, the personality traits obtained from Humantic AI could introduce large biases if the information is inaccurate. Although we were able to statistically ground some concern regarding their claimed accuracy, the conclusion remains uncertain, and a thorough study would be required to confirm the accuracy of these traits.

Our finding in this thesis would be more conclusive if we had more data. Our focus on scalable, technology-driven startups in Norway restricted the available data from public sources. This constraint made it impractical to concentrate on startups founded within the same year, an approach that would have offered a fair comparison between founders and companies. Expanding the scope to include other Scandinavian countries, or broadening the definition of startups, might be considered for future research.

In our pursuit of focusing on founder characteristics, our extensive data collection has resulted in a unique dataset with potential of which has not been fully explored in this thesis. The precise calculation of the performance scores for the companies can be used for different purposes and for both regression and multiple classification problems. Moreover, conducted tests showed that incorporating more company data, such as funding, greatly improved our models' performances, suggesting that the omitted variable could be strong indicators of startup success. Therefore, with the possibility of additional data collection to reduce the missingness of some variables, the dataset could prove beneficial for subsequent research efforts.

8.5 Ethical Considerations

Collecting and using data, especially personal data, for scientific research requires careful consideration of ethical principles. In this study, data on founders were obtained legally from public LinkedIn profiles, and all identities have been anonymized to ensure privacy. However, this does not diminish the importance of considering privacy concerns. Despite data being openly accessible, users may not anticipate the extent to which their information might be used, including their personality being analysed using AI technology.

In our rapid progressing society marked by the advance of artificial intelligence, the rate and scale at which personal data is collected, often without consent increases day by day (Pearce, 2021). Although users implicitly consent to some data collection when they create and use their LinkedIn profiles, there is a pertinent ethical question regarding the level of depth of that consent. Can consent to data collection for one purpose be extended to include more complex and intrusive forms of data analysis, like AI personality estimations? Users might reasonably expect their information to be viewed by potential employers or colleagues but may not anticipate an algorithmic analysis of their personality traits based on their public profile.

Furthermore, the use of AI technologies like Humantic AI to estimate personality traits raises additional ethical questions. While these technologies can provide valuable insights, they operate based on algorithms that may not be transparent or fully understood by those using them or those whose data is being analysed. Moreover, the accuracy and validity of AI predictions about deeply personal characteristics like personality traits are not absolute and could potentially lead to misinterpretation or misuse.

It is essential to reflect upon the potential biases and consequences that can arise from utilization of inaccurate data in both academic research and practical applications. The dataset of founder characteristics used in his study could be subject to both machine error and human error. It is important and necessary to again explicitly state that a full validity analysis of the personality traits obtained from Humantic AI, has not been conducted in this thesis. Guaranteeing data accuracy from such sources call for extensive studies over time.

A significant amount of unprecise or biased data can lead to inaccurate predictions for potential users and misguided interpretations of our findings. Such inaccuracies can lead to financial losses and misinformed decision-making. Thus, it is of utmost importance that our research will be interpreted with caution, especially when considering the unvalidated accuracy of the personality traits analysis performed by AI.

9 Conclusion

In this thesis we have investigated if machine learning models can prove themselves valuable in predicting founder success and identify advantageous founder characteristics at the earliest stages of startups in Norway.

The findings are encouraging, with the tree-based methods, XGBoost and Random Forest, being able to correctly predict successful companies well above random guess, despite relying on a limited feature set primarily consisting of founder-specific data. Logistic Regression also performed well and additionally proved to be highly valuable for interpretability purposes, and consequently is chosen as the best model overall. In alignment with existing literature, this thesis reaffirmed that the number of founders in a startup and the Big Five personality trait conscientiousness are significant predictors of success with a positive correlation. Yet, while other variables arose as promising indicators of success and improved accuracy, there is a lack of evidence to claim any significance beyond the above-mentioned predictors.

The thesis ever so slightly contributes towards bridging the literature gap existing in the Norwegian startup ecosystem. With thoughtful application and certain adaptations, the investigated machine learning models could become valuable tools to guide decision-making processes of venture capital firms.

However, while the results are intriguing, it is crucial to emphasize the limitations of this thesis. The structure and collected data sources introduce biases that have not been fully examined. The validity of the personality traits is of particular concern. Moreover, additional data could further solidify the findings.

We encourage further exploration into the realm of founders and startups within Scandinavian countries. As emerging artificial intelligence platforms offer novel data collection opportunities, there are undoubtedly more insights awaiting discovery in this intriguing field.

References

Abdullah, T. A. A., Zahid, M. S. M., & Ali, W. (2021). A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry*, *13*(12), 2439. https://www.mdpi.com/2073-8994/13/12/2439

Akmam, E. F., Siswantining, T., Soemartojo, S. M., & Sarwinda, D. (2019). Multiple imputation with predictive mean matching method for numerical missing data. 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS),

Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*(9), 735-739.

Albertsen, M. O., Johnsen, P. F. F., & Grimsby, G. (2021). *Startups and Scaleups in the Oslo Region* (No. 23). Menon Publication. https://oslobusinessregion.no/uploads/images/Scaleup-rerport-Menon/Final-report-Startups-and-scaleups-in-the-Oslo-region.pdf

Albertsen, M. O., Johnsen, P. F. F., & Grimsby, G. (2022). *Startups and Scaleups in the Oslo Region* (No. 140). Menon Publication. https://www.menon.no/wp-content/uploads/2022-140-Startups-and-scaleups-in-the-Oslo-region-2022.pdf

Azoulay, P., Jones, B. F., Kim, J. D., & Miranda, J. (2020). Age and high-growth entrepreneurship. *American Economic Review: Insights*, *2*(1), 65-82.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, *44*(1), 1-26.

Bengio, Y., Delalleau, O., & Simard, C. (2010). Decision trees do not generalize to new variations. *Computational Intelligence*, *26*(4), 449-467.

Blaine, B. E. (2018). Winsorizing. *The SAGE encyclopedia of educational research, measurement, and evaluation*, 1817.

Blumberg, D. L. (2021). The Secrets of Highly Successful Young Entrepreneurs. Retrieved 30. May 2023, from https://www.gsb.stanford.edu/insights/secrets-highly-successful-young-entrepreneurs

Bonelli, M. (2022). The Adoption of Artificial Intelligence by Venture Capitalists.

Box, G. (1979). All models are wrong, but some are useful. *Robustness in Statistics*, *202*(1979), 549.

Brønnøysundregistrene. (2022). *Åpne data* https://www.brreg.no/produkter-og-tjenester/apne-data/?nocache=1684315694090

Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, *111*(12), 4427-4431.

Buuren, S. v. (2023). *Predictive mean matching*. https://stefvanbuuren.name/fimd/sec-pmm.html

Carlsson, B., Braunerhjelm, P., McKelvey, M., Olofsson, C., Persson, L., & Ylinenpää, H. (2013). The evolving domain of entrepreneurship research. *Small business economics*, *41*, 913-930.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18,

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, *10*(8), 1283.

Chandra, Y. (2018). Mapping the evolution of entrepreneurship as a field of research (1990–2013): A scientometric analysis. *PloS one*, *13*(1), e0190228.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Clark, M. (2020, 10. April). *Factor Analysis with the psych package*. https://m-clark.github.io/posts/2020-04-10-psych-explained/

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, *74*(368), 829-836.

Crunchbase. (2022). *How does the machine learning model work for Similar Companies?* https://support.crunchbase.com/hc/en-us/articles/5006346272147

Dai, D., Xu, T., Wei, X., Ding, G., Xu, Y., Zhang, J., & Zhang, H. (2020). Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. *Computational Materials Science*, *175*, 109618.

Davila, A., Foster, G., & Gupta, M. (2003). Venture capital financing and the growth of startup firms. *Journal of business venturing*, *18*(6), 689-708.

Decker, R., Haltiwanger, J., Jarmin, R., & Miranda, J. (2014). The role of entrepreneurship in US job creation and economic dynamism. *Journal of economic perspectives*, *28*(3), 3-24.

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

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 1-37.

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, *61*(6), 2623-2640.

FasterCapital. (2023, 16. February). *The Rise of Technology Startups in the Business World*. https://fastercapital.com/content/The-Rise-of-Technology-Startups-in-the-Business-World.html

Fernandes, A. A. R. (2022). Computational Statistics with Dummy Variables. In *Computational Statistics and Applications*. IntechOpen.

Forbes. (2022). *What Is A Startup? The Ultimate Guide*. https://www.forbes.com/advisor/business/what-is-a-startup/

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gartner. (2021, 10. March). *Gartner Says Tech Investors Will Prioritize Data Science and Artificial Intelligence Above "Gut Feel" for Investment Decisions By 2025 [Press release]* https://www.gartner.com/en/newsroom/press-releases/2021-03-10-gartner-says-tech-investors-will-prioritize-data-science-and-artificial-intelligence-above-gut-feel-for-investment-decisions-by-20250

Gartner, W., Starr, J., & Bhat, S. (1999). Predicting new venture survival: an analysis of "anatomy of a start-up." cases from Inc. Magazine. *Journal of business venturing*, *14*(2), 215-232. https://doi.org/10.1016/S0883-9026(97)00063-3

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gjerde, K. Ø. (2023). *Oljebyen Stavanger*. Norsk Oljemuseum. https://www.norskolje.museum.no/forside/kunnskap/publikasjoner/artikler/oljebyen-stavanger/

Gompers, P., Kovner, A., Lerner, J., & Scharfstein, D. (2010). Performance persistence in entrepreneurship. *Journal of financial economics*, *96*(1), 18-32.

Gompers, P., & Lerner, J. (2001). The venture capital revolution. *Journal of economic perspectives*, *15*(2), 145-168. https://doi.org/10.1257/jep.15.2.145

Graham, P. (2012). Startup= growth. *Paul Graham*.

Greenberg, J., & Mollick, E. R. (2018). Sole survivors: solo ventures versus founding teams. *Available at SSRN 3107898*.

Gross, B. (2015). *The Single Biggest Reason Why Start-Ups Succeed. [Video].* Youtube. https://www.youtube.com/watch?v=bNpx7gpSqbY

Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., & Tiwari, V. (2022). Comparing the performance of machine learning algorithms using estimated accuracy. *Measurement: Sensors*, *24*, 100432.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning* The University of Waikato].

Hashemi, M., & Karimi, H. (2018). Weighted machine learning. *Statistics, Optimization and Information Computing*, *6*(4), 497-525.

Healy, L. M. (2006). Logistic regression: An overview. *Eastern Michighan College of Technology*.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Humantic AI. (2022, 24. April). *How Does Humantic AI Work?* https://humantic.ai/blog/how-does-humantic-ai-work/

Humantic AI. (2023a). *One powerful API to understand every individual and personalize every interaction*. https://api.humantic.ai/

Humantic AI. (2023b). *Response Structure*. Retrieved 27. May from https://api.humantic.ai/#response_structure

Iguazio. (2023). *What is the Classification Threshold in Machine Learning?* https://www.iguazio.com/glossary/classification-threshold/

Innovation Norway. (2019). *Female Entrepreneurship in Norway*. Innovation Norway. https://www.innovasjonnorge.no/globalassets/0-innovasjonnorge.no/verktoy-og-temasider/verktoy-for-oppstart-av-bedrift/hvordan-bygge-gode-team/in_forskningsrapport_engelsk_190311_orginal_web_oppslag.pdf

Jain, V. (2022, 31. January). *Introduction to KNN Algorithms*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/01/introduction-to-knn-algorithms/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives.

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal*, *61*(2), 586-614.

Kollmann, T., Stöckmann, C., Hensellek, S., & Kensbock, J. (2016). *European startup monitor 2016*. Universität Duisburg-Essen Lehrstuhl für E-Business Graz.

Leutner, F., Ahmetoglu, G., Akhtar, R., & Chamorro-Premuzic, T. (2014). The relationship between the entrepreneurial personality and the Big Five personality traits. *Personality and individual differences*, *63*, 58-63.

Levesque, M., & Minniti, M. (2006). The effect of aging on entrepreneurial behavior. *Journal of business venturing*, *21*(2), 177-194.

Liang, Y. E., & Yuan, S.-T. D. (2016). Predicting investor funding behavior using crunchbase social network features. *Internet Research*.

Løset, G. K., & von Soest, T. (2023). Big five personality traits and physician-certified sickness absence. *European Journal of Personality*, *37*(2), 239-253.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

March-Chordà, I. (2004). Success factors and barriers facing the innovative start-ups and their influence upon performance over time. *International Journal of Entrepreneurship and Innovation Management*, *4*(2-3), 228-247.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), 175-215.

Metrick, A., & Yasuda, A. (2021). *Venture capital and the finance of innovation*. John Wiley & Sons.

Miloud, T., Aspelund, A., & Cabrol, M. (2012). Startup valuation by venture capitalists: an empirical study. *Venture Capital*, *14*(2-3), 151-174.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human performance*, *11*(2-3), 145-165.

Murphy, G. B., Trailer, J. W., & Hill, R. C. (1996). Measuring performance in entrepreneurship research. *Journal of business research*, *36*(1), 15-23. https://doi.org/10.1016/0148-2963(95)00159-X

Nassar, I. A., Almsafir, M. K., & Al-Mahrouq, M. H. (2014). The validity of Gibrat's law in developed and developing countries (2008–2013): Comparison based assessment. *Procedia-Social and Behavioral Sciences*, *129*, 266-273.

NHO. (2023). *Næringslivet i Nordland*. Næringslivets Hovedorganisasjon. https://www.nho.no/regionkontor/nho-nordland/naeringslivet-i-nordland/

Nielsen, D. (2016). Tree boosting with XGBoost. *Norwegian University of Science and Technology*.

NTB. (2023). *Leading online grocery store Oda raises NOK 1.5 billion* https://kommunikasjon.ntb.no/pressemelding/leading-online-grocery-store-oda-raises-nok-15-billion?publisherId=12583954&releaseId=17949388

Ostfeld, A., & Salomons, S. (2005). A hybrid genetic—instance based learning algorithm for CE-QUAL-W2 calibration. *Journal of Hydrology*, *310*(1-4), 122-142.

Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022). Impact of feature scaling on machine learning models for the diagnosis of diabetes. 2022 International Conference on Artificial Intelligence in Everything (AIE), Palanivinayagam, A., & Damaševičius, R. (2023). Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods. *Information*, *14*(2), 92.

Pareto Securities. (2022, 19. December 2022). *Årets børsnoteringer: Disse aksjene gikk på børs i 2022*. Retrieved 8. May from https://www.paretosec.no/aktuelt/aarets-boersnoteringer-disse-aksjene-gikk-paa-boers-i-2022

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Pearce, G. (2021, 28. May). *Beware the Privacy Violations in Artificial Intelligence Applications*. ISACA. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2021/beware-the-privacy-violations-in-artificial-intelligence-applications

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Prabhakaran, S. (2023). *r-statistics.co*. http://r-statistics.co/Loess-Regression-With-R.html#google_vignette

Proff. (2023). *Proff – The Business Finder*. https://innsikt.proff.no/om-proff/

Proff Forvalt. (2023a). *Hva betyr karakterene i Forvalt ratingen?* Retrieved May 10. from https://forvalt.no/kredittsjekk/konkurs/konkursrating

Proff Forvalt. (2023b). *Hva er Proff Forvalt?* https://forvalt.no/

Qiu, X., Wu, H., & Hu, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC bioinformatics*, *14*, 1-10.

Rafiei, F. M., Manzari, S., & Bostanian, S. (2011). Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence. *Expert systems with applications*, *38*(8), 10210-10217.

Rengasamy, D., Mase, J. M., Kumar, A., Rothwell, B., Torres, M. T., Alexander, M. R., Winkler, D. A., & Figueredo, G. P. (2022). Feature importance in machine learning models: A fuzzy information fusion approach. *Neurocomputing*, *511*, 163-174.

Retterath, A. (2020, 20. Oct). *What's the best startup/VC database?* Medium. https://medium.com/@andreretterath/what-s-the-best-startup-vc-database-8237fc189830

Rhea, A. K., Markey, K., D'Arinzo, L., Schellmann, H., Sloane, M., Squires, P., & Stoyanovich, J. (2022). External Stability Auditing to Test the Validity of Personality Prediction in AI Hiring. *arXiv preprint arXiv:2201.09151*.

Rosvold, T. O., & Rosvold, E. J. (2022). *Entrepreneurial attributes and success: A study of motivation, personality, and human capital* Handelshøyskolen BI].

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rusdah, D. A., & Murfi, H. (2020). XGBoost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, *2*, 1-10.

Saha, S. (2018, 7. October). *Understanding the log loss function of XGBoost*. Medium. https://medium.datadriveninvestor.com/understanding-the-log-loss-function-of-xgboost-8842e99d975d

Sahani, N., & Ghosh, T. (2021). GIS-based spatial prediction of recreational trail susceptibility in protected area of Sikkim Himalaya using logistic regression, decision tree and random forest model. *Ecological Informatics*, *64*, 101352.

Sahlman, W. (2010). Risk and reward in venture capital. *Harvard Business School Entrepreneurial Management case*(811-036).

Santisteban, J., & Mauricio, D. (2017). Systematic literature review of critical success factors of information technology startups. *Academy of Entrepreneurship Journal*, *23*(2), 1-23. https://www.researchgate.net/publication/322094432_Systematic_literature_review_of_critical_success_factors_of_Information_Technology_startups

Schildt, H. A., Zahra, S. A., & Sillanpää, A. (2006). Scholarly communities in entrepreneurship research: A co–citation analysis. *Entrepreneurship theory and practice*, *30*(3), 399-415.

Sevilla-Bernardo, J., Sanchez-Robles, B., & Herrador-Alcaide, T. C. (2022). Success Factors of Startups in Research Literature within the Entrepreneurial Ecosystem. *Administrative Sciences*, *12*(3), 102. https://doi.org/10.3390/admsci12030102

Shane, S., Dolmans, S. A., Jankowski, J., Reymen, I. M., & Romme, A. G. L. (2015). Academic entrepreneurship: Which inventors do technology licensing officers prefer for spinoffs? *The Journal of Technology Transfer*, *40*, 273-292.

Sharma, S., Choudhary, M., & Shrotriya, V. (2022). Is Women's Personality Different From Men's Personality-Big Five Personality Traits And Gender Roles. *Journal of Pharmaceutical Negative Results*, 1755-1758.

Skjelsbæk, J. (2023, 31. March). Kickoff for håpefulle gründere. *Shifter*. https://www.shifter.no/nyheter/kickoff-for-hapefulle-grndere/275535

SSB. (2022, 13 September). *Newly established enterprises, survival and growth*. Statistisk sentralbyrå. Retrieved 8. May from https://www.ssb.no/en/virksomheter-foretak-og-regnskap/virksomheter-og-foretak/statistikk/nyetablerte-foretaks-overlevelse-og-vekst

SSB. (2023). *Standard for næringsgruppering (SN)*. Statistisk sentralbyrå. Retrieved 23. May from https://www.ssb.no/klass/klassifikasjoner/6

Startup Heatmap. (2023). *THE 10 BEST STARTUP DATABASES IN EUROPE*. https://startupheatmap.eu/best-startup-databases/

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.

Sven Blumberg, M. K., Elina Mäkelä, Henning Soller. (2023, 24. January). *Women in tech: The best bet to solve Europe's talent shortage*. Retrieved 26. May from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/women-in-tech-the-best-bet-to-solve-europes-talent-shortage#/

Szarek, J., & Piecuch, J. (2018). The importance of startups for construction of innovative economies. *International Entrepreneurship Review*, *4*(3), 389.

Taulli, T. (2012). *How to create the next Facebook: seeing your startup through, from Idea to IPO*. Apress.

The World Bank. (2023). *Stratified Random Sample*. Retrieved 25. May from https://dimewiki.worldbank.org/Stratified_Random_Sample

The World Financial Review. (2023, 10. May). *Top Industries To Learn From About Scalability*. https://worldfinancialreview.com/top-industries-to-learn-about-scalability/

Thébaud, S. (2015). Status beliefs and the spirit of capitalism: Accounting for gender biases in entrepreneurship and innovation. *Social Forces*, *94*(1), 61-86.

Tidymodels. (2023a). *Boosted trees*. Retrieved 20/5/2023 from https://parsnip.tidymodels.org/reference/boost_tree.html

Tidymodels. (2023b). *K-nearest neighbors*. https://parsnip.tidymodels.org/reference/nearest_neighbor.html

Tupes, E. C., & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. *USAF ASD Tech. Rep.*, *61-97*.

Ünal, C. (2019). *Searching for a Unicorn: A Machine Learning Approach Towards Startup Success Prediction* Humboldt-Universität zu Berlin].

Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. *Machine Learning*, 1-12.

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599-606.

Wasserman, N. (2012). *The founder's dilemmas: Anticipating and avoiding the pitfalls that can sink a startup*. Princeton University Press.

Werth, D., & Greff, T. (2018). Scalability in consulting: insights into the scaling capabilities of business models by digital technologies in consulting industry. *Digital Transformation of the Consulting Industry: Extending the Traditional Delivery Model*, 117-135.

Westhead, P., & Wright, M. (1998). Novice, portfolio, and serial founders: are they different? *Journal of business venturing*, *13*(3), 173-204.

Widiger, T. A. (2017). *The Oxford handbook of the five factor model*. Oxford University Press.

Winther, F. (2023, 26. April). «AI i venture-investeringer: Slik er den nye skolen». *Shifter*. https://www.shifter.no/debatt/ai-i-venture-investeringer-slik-er-den-nye-skolen/277191

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.

Yang, Y., & Loog, M. (2018). A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, *83*, 401-415.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, *9*(2), 79-94.

Zhao, H., & Seibert, S. E. (2006). The big five personality dimensions and entrepreneurial status: a meta-analytical review. *Journal of applied psychology*, *91*(2), 259.

Zhao, H., Seibert, S. E., & Lumpkin, G. T. (2010). The relationship of personality to entrepreneurial intentions and performance: A meta-analytic review. *Journal of management*, *36*(2), 381-404.

Appendix

A.1: Calculation of Six Scores for Weighted Startup Success Score

The calculations and considerations taken for each of the six scores applied to create the weighted success score is described in detail in this appendix.

Growth

The growth variables are calculated by firstly calculating percentage change from year to year, getting one growth column for each year span for every row/company. We set a maximum growth value for each year span to be 500 percent. We recognize that real growth beyond 5x could happen organically on rare occasions, but as we find most of the inflated growth values to be from what we assume to be accounting errors we elect to limit these values to the 5x mark. Then we elected to weigh the years closer to the firm's infancy more. Our reasoning behind this choice was because some of the older companies' growth seemed to stagnate after they had reached a certain age. To make the growth of these companies comparable to younger companies we elected to weigh growth during the first years more. Lastly, we calculated the growth average by dividing the sum of weighed growth values by the sum of weights.

There were two reasons behind choosing turnover and employee growth specifically for our growth metrics. Firstly, because our background research supported these metrics to be representative and commonly used among startup success analysis. Secondly, because these metrics had few missing values in our accounting database Forvalt. The turnover growth was derived from the "Sum_Driftsinnt_year_x" columns, which translates to sum of all income related to operations. The employee growth variable was derived from the "Aarsverk_year_x" columns, which is the full-time equivalent (FTE), where the value 1.0 represents the workload of one full time employee.

Size

The calculation of the size variables was done simply by calculating a weighted average like described above in the growth section. Except, when we are analysing size, our intention is to evaluate the status que, and we are therefore weighing the accounting data of the most current years more. The reason behind not choosing solely to use the newest year's accounting data is because unfavourable fluctuations and noise might create a biased picture of the actual size of the company, for example the corona pandemic. The reasoning behind prompting for the use of turnover and employee data for size variables is as with the case of growth variables because it is founded in earlier research and the lack of missing data for these variables makes them well suited for analysis.

Profitability

The measure used to include profitability was the annual result, we divided each result by the turnover of said company each year, to make it a comparable ratio across companies. As with the size factors, we are interested in the status que, so we weigh the average accordingly so that results from recent years are more significant before calculating an average. The annual result was derived from the "Aarsresultat_year_x" columns and divided by the corresponding "Sum_Driftsinnt_year_x" columns to create a comparable ratio. The reasoning behind opting for the annual result to compute a profitability metric was due to this number having few missing values, and because the profitability ratios provided by Forvalt had a lot of unrepresentative inflated values due to missing values being used during calculations.

Financial health

For the financial health metric, there was no need for any calculations as Forvalt provided us with a representative bankruptcy score for each company. The score is a single value between 0 and 100 000 and is meant to provide an indication of bankruptcy probability, with companies with values closer to 0 being more likely to go bankrupt.

Calculating a representative score

Normalization of variables

After calculating the above-mentioned variables, we get the following summary:

Summary statistics for the six variables

	Income growth	**Employee growth**	**Employee size**	**Income size**	**Profitability (Turnover/Result)**	**Score Forvalt**
Min	-0.7771	-0.2769	0.000	-324333	-3664.769	0
1st Qu	0.1696	0.0000	1.200	1031977	-0.710	72000
Median	0.4809	0.1066	3.000	3017182	-0.015	88000
Mean	0.5320	0.1775	6.944	11833901	-5.608	79344
3rd Qu	0.8546	0.2876	6.500	8112490	0.071	95000
Max	2.2578	1.1795	457.833	1125977355	0.936	100000
NA	168	386	47	43	99	321

Table 25: Summary statistics for variables to create success score

As showcased by the summary table, our elected metrics are represented by vastly different scales and spread. To combine the variables into a single score some method of normalization would be necessary. We opted for the use of min-max normalization (Al Shalabi et al., 2006), a method which maps the value v' to the range [new_min$_a$, new_max$_a$] by computing:

$$\mathrm{v}' = \left((\mathrm{v} - \mathrm{min}_a) \,/\, (\mathrm{max}_a - \mathrm{min}_a) \right) * (\mathrm{new\_max}_a - \mathrm{new\_min}_a) + \mathrm{new\_min}_a$$

Equation 15: Min-max normalization

Put simply, it uses the maximum and minimum value to map the values on a scale between a new set of boundaries relative to the maximum and minimum values. Our reasoning behind opting for this type of normalization was that this method keeps relationship among original data and easy to interpret(Patro & Sahu, 2015). This simplicity was essential when performing trials of different score metrics and weights, furthermore Z-score normalization yielded highly similar results.

We opted for a min-max range of 0-100, so that every company was given a score between zero and 100 for all six variables.

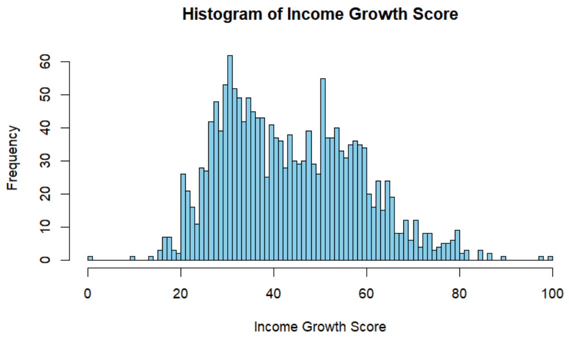

Figure 26: Histogram of growth score

However, for the size-variables, we opted to use winsorizing, a pre-processing method involving replacing outliers with less extreme values (Blaine, 2018). We selected the 99 percentiles as the cut-off value, so that the one percent highest values were set to the cut-off value. Our reasoning was that these ranges were highly skewed, by giving the largest companies the size score of 100 we provided them with a premium for being largest while keeping a high degree of diversity in size score for the remaining companies.

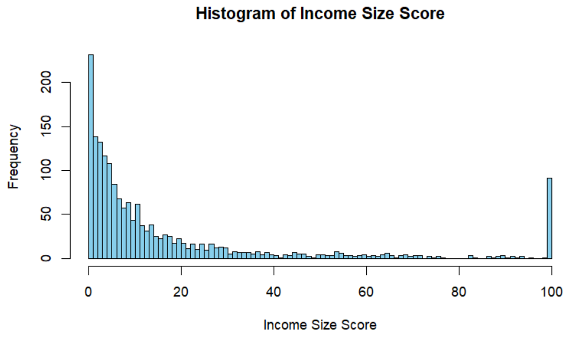

Figure 27: Histogram of income size score

Lastly, for the profitability score, we opted to use rank-based normalization instead of min-max, it is done by representing each observation by its fractional rank (Qiu et al., 2013). This was done to the nature of the variable consisting of a lot of negative values, and some extremely negative. Therefore, we reason that it provides a better insight of profitability to observe how the companies perform compared to each other.

A.2: Loess Smoothing Method

Locally Estimated Scatterplot Smoothing (LOESS) is a non-parametric method that can fit multiple least square regressions locally (local regression), on different subsets of the data (Cleveland, 1979). This method makes no assumptions about the underlying data structure and is particular proficient in fitting complex shapes but does not produce a clear mathematical formula of the regression function. This method is applied in figure 8 and 9 with a span of 0.7 and 0.1 respectively. The "span" argument in R, ranges between 0 and 1, and decides the degree of smoothing (Prabhakaran, 2023).

A.3: Factor Analysis

Factor analysis is a statistical method for identifying the underlying variables (factors) that can explain the pattern of correlation between the observed variables (Yong & Pearce, 2013). This method is often used in feature reduction, as it can point out variables that explain most of the variance in the data. In table 9 (Exploratory Data Analysis section), we used the "psych" package in R to perform the factor analysis. The fitting method used for the factors were chosen to be Minimum Residual (MR), which is the default (Clark, 2020). The "Cumulative Var" in the table 9 represents the cumulative sum of the "Proportion Var", where "Proportion Var" explains how much the overall variance the factor accounts for out of all the variables (Clark, 2020).

A.4: Imputation table

Variable	NA	Imputation method
“Age_f”	91	PMM (regression) + algorithm
“Similar_Companies_c”	196	PMM (regression)
“Gender_f”	166	Manual research
“Num_roles_proff”	161	PMM (regression)
“Num_Schools_f”	352	Manual research + imputing 1
“County_8_f”	440	PMM (polyreg) + algorithm
“Num_Employees_year0_c”	871	Imputing estimated number of founders
“Num_Founders_Estimated_c”	1928	Imputing number of founders we found
"Num_Founder_Titles_Work_f"	0	Imputing 1 for 0
"Country_Norway_f"	30	Manual research

Table 26: Imputation techniques applied

A.5: Formula for Logistic Regression Model

Note: Variables are standardized in this equation

Log(odds) = - 0.4470 + 0.1784*Num_Founders_c + 0.1215*Similar_Companies_c + 0.0333*Age_f - 0.1041*Num_roles_proff_f + 0.0506*Num_Schools_f - 0.107*Num_Jobs_f - 0.0342*Num_Founder_Titles_Work_f - 0.1008*Openness_score + 0.0464*Extraversion_score - 0.0485*Emotional_Stability_score + 0.0743*Agreeableness_score + 0.1764*Conscientiousness_score + 0.4809*County_c_INNLANDET - 1.1212*County_c_M.U.00D8.RE.OG.ROMSDAL - 0.8001*County_c_NORDLAND - 0.5967*County_c_OSLO - 0.8434*County_c_ROGALAND - 0.8222*County_c_TR.U.00D8.NDELAG - 0.9374*County_c_TROMS.OG.FINNMARK - 1.0608*County_c_VESTFOLD.OG.TELEMARK - 0.3963*County_c_VESTLAND - 0.8289*County_c_VIKEN - 0.6469*County_8_f_Oslo + 0.5025*County_8_f_Other + 1.3079*County_8_f_Rogaland + 1.0031*County_8_f_Tr.U.00F8.ndelag + 0.666*County_8_f_Vestfold.og.Telemark + 0.1716*County_8_f_Vestland + 0.7631*County_8_f_Viken + 0.4787*Country_Norway_f_yes + 0.1591*DISC_Summary_group_dominant + 0.0086*DISC_Summary_group_influential + 0.2044*DISC_Summary_group_steady + 0.2209*Gender_f_male + 0.5408*NACE_group_7_c_C...Industri - 0.2548*NACE_group_7_c_G...Varehandel..reperasjon.av.motorvogner - 0.2456*NACE_group_7_c_J...Informasjon.og.kommunikasjon - 0.6464*NACE_group_7_c_K...Finansierings..og.forsikringsvirksomhet - 0.0455*NACE_group_7_c_M...Faglig..vitenskaplig.og.teknisk.tjenesteyting - 0.0331*NACE_group_7_c_N...Forretningsmessing.tjenesteyting

A.6: Crunchbase Dataset

Number of rows: 2768
Number of columns: 101

Variable	Type	# NA-values
[1] “Organization.Name”	character	0
[2] "Organization.Name.URL"	character	0
[3] "Number.of.Private.Notes"	character	2768
[4] "Last.Funding.Type"	character	2160
[5] "Operating.Status"	character	0
[6] "IPO.Status"	character	0
[7] "Acquisition.Status"	character	2608
[8] "Company.Type"	character	132
[9] "Full.Description"	character	1470
[10] "Founded.Date"	character	0
[11] "Founded.Date.Precision"	character	0
[12] "CB.Rank..Company."	character	0
[13] "Last.Funding.Date"	character	2160
[14] "Founders"	character	1809
[15] "Headquarters.Location"	character	0
[16] "Description"	character	0
[17] "Number.of.Employees"	character	124
[18] "Estimated.Revenue.Range"	character	998
[19] "Exit.Date"	character	2604
[20] "Exit.Date.Precision"	character	2604
[21] "Closed.Date"	character	2758
[22] "Closed.Date.Precision"	character	2662
[23] "Website"	character	15
[24] "LinkedIn"	character	637
[25] "Aberdeen...IT.Spend"	integer	2612
[26] "Aberdeen...IT.Spend.Currency"	character	2606
[27] "Aberdeen...IT.Spend.Currency..in.USD."	integer	2612
[28] "Apptopia...Number.of.Apps"	integer	2588
[29] "BuiltWith...Active.Tech.Count"	integer	424
[30] "SEMrush...Global.Traffic.Rank"	character	1796
[31] "SEMrush...Average.Visits..6.months."	character	2240
[32] "Top.5.Investors"	character	2275
[33] "Number.of.Funding.Rounds"	integer	2160
[34] "Funding.Status"	character	2292
[35] "Total.Funding.Amount"	numeric	2313
[36] "Total.Funding.Amount.Currency"	character	2160
[37] "Total.Funding.Amount.Currency..in.USD."	numeric	2313
[38] "Number.of.Founders"	integer	1809
[39] "Number.of.Founders..Alumni."	character	2768
[40] "Investment.Stage"	character	2760

[41] "Number.of.Articles"	integer	2256
[42] "Hub.Tags"	character	2763
[43] "Actively.Hiring"	character	2759
[44] "Facebook"	character	985
[45] "Twitter"	character	1849
[46] "Investor.Type"	character	2754
[47] "Number.of.Exits"	integer	2761
[48] "Number.of.Investments"	integer	2749
[49] "Number.of.Alumni"	integer	2767
[50] "Industry.Groups"	character	62
[51] "Industries"	character	62
[52] "Last.Funding.Amount"	numeric	2335
[53] "Last.Funding.Amount.Currency"	character	2335
[54] "Last.Funding.Amount.Currency..in.USD."	numeric	2335
[55] "Total.Equity.Funding.Amount"	integer	2349
[56] "Total.Equity.Funding.Amount.Currency"	character	2198
[57] "Total.Equity.Funding.Amount.Currency..in.USD."	numeric	2349
[58] "Last.Equity.Funding.Type"	character	2198
[59] "Last.Equity.Funding.Amount"	Integer	2371
[60] "Last.Equity.Funding.Amount.Currency"	character	2371
[61] "Last.Equity.Funding.Amount.Currency..in.USD."	numeric	2371
[62] "Number.of.Investors"	integers	2274
[63] "Number.of.Acquisitions"	integer	2713
[64] "Transaction.Name"	character	2654
[65] "Transaction.Name.URL"	character	2654
[66] "Acquired.by"	character	2654
[67] "Acquired.by.URL"	character	2654
[68] "Announced.Date"	character	2654
[69] "Announced.Date.Precision"	character	2654
[70] "Price"	integer	2748
[71] "Price.Currency"	character	2748
[72] "Price.Currency..in.USD."	integer	2748
[73] "Acquisition.Type"	character	2658
[74] "Acquisition.Terms"	character	2757
[75] "Valuation.at.IPO"	numeric	2759
[76] "Valuation.at.IPO.Currency"	character	2759
[77] "Valuation.at.IPO.Currency..in.USD."	numeric	2759
[78] "IPO.Date"	character	2715
[79] "Last.Leadership.Hiring.Date"	character	2748
[80] "Last.Layoff.Mention.Date"	character	2766
[81] "Number.of.Events"	integer	2648
[82] "CB.Rank..Organization."	character	0
[83] "Trend.Score..7.Days."	numeric	0
[84] "Trend.Score..30.Days."	numeric	0
[85] "Trend.Score..90.Days."	numeric	0
[86] "Similar.Companies"	integer	237
[87] "SEMrush...Monthly.Visits"	character	1796
[88] "SEMrush...Monthly.Visits.Growth"	character	2023

[89] "SEMrush...Visit.Duration"	character	1796
[90] "SEMrush...Visit.Duration.Growth"	character	2167
[91] "SEMrush...Page.Views...Visit"	numeric	1796
[92] "SEMrush...Page.Views...Visit.Growth"	character	2023
[93] "SEMrush...Bounce.Rate"	character	2796
[94] "SEMrush...Bounce.Rate.Growth"	character	2085
[95] "SEMrush...Monthly.Rank.Change...."	character	2023
[96] "SEMrush...Monthly.Rank.Growth"	character	2023
[97] "Apptopia...Downloads.Last.30.Days"	character	2693
[98] "G2.Stack...Total.Products.Active"	integer	2322
[99] "Tags"	character	2768
[100] "Number.of.Lead.Investors"	integer	2486
[101] "IPqwery...Patents.Granted"	integer	2395

Table 27: Crunchbase dataset

A.7: Brønnøysund Dataset

Variable
[1] "Organisasjonsnummer"
[2] "Navn"
[3] "Organisasjonsform"
[4] "Organisasjonsform.beskrivelse"
[5] "Næringskode 1"
[6] "Næringskode 1.beskrivelse"
[7] "Næringskode 2"
[8] "Næringskode 2.beskrivelse"
[9] "Næringskode 3"
[10] "Næringskode 3.beskrivelse"
[11] "Hjelpeenhetskode"
[12] "Hjelpeenhetskode.beskrivelse"
[13] "Antall ansatte"
[14] "Hjemmeside"
[15] "Postadresse.adresse"
[16] "Postadresse.poststed"
[17] "Postadresse.postnummer"
[18] "Postadresse.kommune"
[19] "Postadresse.kommunenummer"
[20] "Postadresse.land"
[21] "Postadresse.landkode"
[22] "Forretningsadresse.adresse"
[23] "Forretningsadresse.poststed"
[24] "Forretningsadresse.postnummer"
[25] "Forretningsadresse.kommune"
[26] "Forretningsadresse.kommunenummer"
[27] "Forretningsadresse.land"
[28] "Forretningsadresse.landkode"
[29] "Institusjonell sektorkode"
[30] "Institusjonell sektorkode.beskrivelse"

[31] "Siste innsendte årsregnskap"
[32] "Registreringsdato i Enhetsregisteret"
[33] "Stiftelsesdato"
[34] "FrivilligRegistrertIMvaregisteret"
[35] "Registrert i MVA-registeret"
[36] "Registrert i Frivillighetsregisteret"
[37] "Registrert i Foretaksregisteret"
[38] "Registrert i Stiftelsesregisteret"
[39] "Konkurs"
[40] "Under avvikling"
[41] "Under tvangsavvikling eller tvangsoppløsning"
[42] "Overordnet enhet i offentlig sektor"
[43] "Målform"

Table 28: Brønnøysund dataset

A.8: Forvalt Dataset

Variable
[1] "Status"
[2] "Orgnr"
[3] "Juridisk selskapsnavn"
[4] "Kommune"
[5] "Fylke"
[6] "Gateadresse (forretningsadresse)"
[7] "Postnr (forretningsadresse)"
[8] "Poststed (forretningsadresse)"
[9] "Antall ansatte (Aa-registret - månedlig oppdatert)"
[10] "NACE-bransjekode"
[11] "NACE-beskrivelse"
[12] "Org.form"
[13] "Daglig leder"
[14] "Styrets leder"
[15] "Reg.dato"
[16] "Stift.dato"
[17] "Aksjekap."
[18-29] "Sum salgsinntekter" 2012-2023
[30-41] "Sum driftsinnt." 2012-2023
[42-54] "Andre driftsinnt.," 2012-2023
[55-66] "Driftsres." 2012-2023
[67-78] "Ord. res. f. skatt" 2012-2023
[79-90] "Årsresultat" 2012-2023
[91-102] "Lønnskostnader" 2012-2023
[103-114] "Andre driftskostnader" 2012-2023
[115-126] "Sum driftskostn." 2012-2023
[127-138] "Sum eiend." 2012-2023
[139-150] "Sum egenkap." 2012-2023

[151-152] "Sum gjeld" 2012-2023
[153-164] "Sum innskutt egenkapital" 2012-2023
[165-176] "Sum opptjent kapital" 2012-2023
[177-188] "Lønnsomhet" 2012-2023
[189-200] "Likviditet" 2012-2023
[201-212] "Soliditet" 2012-2023
[213-224] "Valutakode" 2012-2023
[225-236] "Ant. ansatte/regnskapsår" 2012-2023
[237-248] "Årsverk" 2012-2023
[249] "Score"
[250] "Rating"

Table 29: Forvalt dataset

A.9: Humantic Dataset

Number of rows: 2925
Number of columns: 51

Variable	Type	# NA-values
[1] "Display Name"	character	0
[2] "User Name"	character	0
[3] "Location"	character	8
[4] "User ID"	character	0
[5] "User Description"	character	23
[6] "Demographics"	character	1791
[7] "Languages"	character	2901
[8] "Social Profiles"	character	0
[9] "Education"	character	355
[10] "Work History"	character	13
[11] "Tech Usage"	character	2915
[12] "Interests"	character	2925
[13] "Social Interactions"	character	2925
[14] "Social Activity"	character	785
[15] "Content Affinity"	character	2699
[16] "Hiring Interests"	character	2925
[17] "Conversation Starters"	character	2925
[18] "Stability Potential"	character	424
[19] "Learning Ability"	character	424
[20] "Teamwork Skills"	character	424
[21] "Need for Autonomy"	character	424
[22] "Attitude and Outlook"	character	424
[23] "General Behavior"	character	424
[24] "Action Orientedness"	character	424
[25] "Sales Interests"	character	2925
[26] "Risk Appetite"	character	424

[27] "Ability To Say No"	character	424
[28] "Speed"	character	424
[29] "Decision Drivers"	character	424
[30] "Calculativeness"	character	424
[31] "Influence"	character	424
[32] "Dominance"	character	424
[33] "Steadiness"	character	424
[34] "DISC Summary"	character	424
[35] "OCEAN Summary"	character	424
[36] "Openness"	character	424
[37] "Extraversion"	character	424
[38] "Emotional Stability"	character	424
[39] "Agreeableness"	character	424
[40] "Conscientiousness"	character	424
[41] "Confidence Level"	character	0
[42] "Skills"	character	266
[43] "Followers"	numeric	0
[44] "Experience in Years"	numeric	132
[45] "Social Activity Status"	character	0
[46] "Designation"	character	129
[47] "Education Level"	character	1047
[48] "Job Level"	character	752
[49] "Status"	character	0
[50] "Analysis Status"	character	0
[51] "Confidence Score"	character	0

Table 30: Humantic dataset