ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Study Program: Business Economics

Specialization: Data Science & Marketing Analytics

Successful Venture Capital Funding: The Role of Founder's Characteristics and Opportunity Costs

Author: J.J. Wallenburg
Email adress: 435432jw@student.eur.nl

Supervisor: dr. V. Avagyan
Second assessor: dr. P. Wan

Date: 06/08/2021

Abstract

Successfully raising venture capital (VC) is a critical element in the process of building a start-up. This paper uses three types of analytics to extract insights of VC funding and founder data of IT-related start-ups in the United states. Firstly, this paper uses descriptive analytics to determine founders' characteristics. Secondly, prescriptive analytics is used to discover the role of these founders' characteristics in the VC funding process. Thirdly, this paper uses predictive analytics to predict VC funding using these founders' characteristics. The results of the prescriptive analysis show positive effects from founder's education and social capital on the likelihood of raising VC. Furthermore, higher opportunity costs due to recent employment strengthen the positive effect of education. While industry-related and entrepreneurial experience do not influence the likelihood of VC directly, the extent of industry-related experience in founder's profile description positively affects the likelihood of receiving VC. The predictive analysis suggests that the best predictions of VC funding are obtained by a random forest model.

Keywords: Founders, Human Capital, Venture Capital Funding, Opportunity Costs, Start-ups

Table of Contents

1. Introduction

The founding of start-ups can lead to innovative technical solutions. According to economic theories, technical progress accounts for almost all of country’s economic growth (Solow, 1956; Hasan & Tucci, 2010). Entrepreneurs and investors are needed to successfully build new start-ups and eventually provide economic growth. On the one hand, entrepreneurs need to discover and execute opportunities. On the other hand, investors need to fund these opportunities. Since entrepreneurs are often wealth constrained, the fundraising of external financing is necessary to expand their business and reach their opportunities. Successfully acquiring funds is therefore a critical element in process of building a successful start-up.

However, acquiring funds through for example traditional bank loans is not always possible for start-ups with a limited operating history and untested business models. This is where venture capital (VC) comes in, which is a form of equity financing for these young firms with high uncertainty but also high growth potential. According to the National Venture Capital Association (NVCA) (2020) the number of US venture-backed firms and venture capital funds in existence raised respectively by 65 percent and 73 percent in the period 2010 to 2019. The VC assets under management of the US venture-backed firms increased with 87 percent in this same period. Also, globally, a positive trend of the importance of venture capital can be observed since the amount of global venture capital investments in 2019 is 889 percent higher compared to 2004 (NVCA, 2020).

1.1. Problem Statement and Research Question

Different examples of successful US ventures exist that used venture capital financing. WhatsApp, Facebook, and Groupon are examples in which venture capital investors successfully identified these ventures’ opportunities (CB Insights, 2021). The most important question for venture capital investors in this VC process is how to spot the next successful entrepreneur. Is there any way to spot the next Mark Zuckerberg (Facebook)? What characteristics of these entrepreneurs influence their success?

Unfortunately for efficiency, it is not easy to detect successful aspiring entrepreneurs. An important obstacle within this VC funding process is the information asymmetry problem between the funding seeker and the potential investors (Shane and Cable, 2002). VC firms have less information about the business opportunity and the funding seeker than the funding seeker himself. Entrepreneurs could act too opportunistically about the business opportunity and themselves to increase the likelihood of receiving VC funding. This information asymmetry problem can therefore lead to adverse selection of start-ups that are too opportunistic by venture capitalists (Akerlof, 1978). The consequence of adverse selection is a higher risk in general and therefore a higher overall risk premium charged by venture capitalists.

Adverse selection of start-ups in the VC process can be tackled by improving the screening process. Despite information asymmetry problems, information about founders is (at the moment of writing this paper) mostly publicly available. VC firms could use this information to decide whether a start-up should receive funding or not. This research investigates the role of founder's characteristics in United States in acquiring venture capital fundraising and tries to predict the VC fundraising. Therefore, the following research question is answered in this paper:

What is the role of founder's characteristics in acquiring venture capital fundraising in the United States and how can these characteristics be used to predict venture capital fundraising?

To answer this research question, I use data from Crunchbase to extract information regarding start-up's VC fundraising and founders. In this paper, I extract variables that capture founder's characteristics regarding education, industry experience, entrepreneurial experience, social capital, and opportunity costs.

1.2. Academic Relevance

This research contributes to two streams of literature. The first stream of literature is the literature regarding the role of entrepreneurs' human capital within a firm. Earlier research implies the importance of an entrepreneur in the success of a company (Matsuno et al., 2002). Also, recent research towards funding success suggests that entrepreneurs' characteristics, such as their prior education (Ko and McKelvie, 2018) and network size (Banerji & Reimer, 2019), are important determinants for start-up ventures. The contribution of my research to this stream is the engineering of multiple new features that capture different effects, such as proxies for the quality of last education and the size of the social capital. Additionally, this research contributes to this literature stream by exploring the relationships between the opportunity costs of recent employment and the VC funding likelihood. Quitting a job, that comes with a fixed salary, to start a new firm, that comes with an uncertain salary, can give a positive signal to venture capitalists regarding founders' drive.

The second stream of literature this paper expands is the literature regarding venture success. Recently, Żbikowski and Antosiuk (2021) and Arroyo, Corea, Jimenez-Diaz and Recio-Garcia (2019) predicted venture success using Crunchbase data. Żbikowski and Antosiuk (2021) adress the look-ahead bias that possibly exists in models derived by Arroyo et al. (2019) because different variables are used for prediction that are unknown at the investment decision moment. In contrast to Arroyo et al. and in line with Żbikowski and Antosiuk (2021), this study takes into account the look-ahead bias by carefully selecting variables available at the decision time. This paper contributes to the literature by using more detailed data about founder's prior experience. More variables are used that capture founder's prior education, industrial/entrepreneurial experience and social capital. In addition, this paper also explores the in text description of entrepreneurs using topic mining.

Additionally, this paper uses a large dataset which is collected using API calls. This structure enables future research to extract comparable Crunchbase datasets in different industries/regions. Using a large automatically extracted dataset is preferred in the exploration of the role of entrepreneurial human capital over small survey-based dataset, since the use of a larger dataset reduces the risk of a response and selection bias. Besides, using an automatically extracted dataset opens doors for future research since future data can be easily updated/extracted.

1.3. Managerial Relevance

This research is relevant for managers in the following ways. Firstly, this research is relevant for education policies, but also for the career path of people themselves who want to become a successful start-up founder. Business studies provide education about entrepreneurship and the process of becoming a successful entrepreneur. The VC fundraising process is an important part of the entrepreneurial success. The insights of this paper can be used as advice to adjust the curriculum of these business studies. Possible implications could be scheduling more time for industry related internships or more lectures that focus on for example social capital. People that want to be a successful start-up founder can also use insights of this paper to plan their career. This research can answer common questions among their career paths. Should these people start a start-up, or should they study first? Should these people gain industry experience after graduation by getting a job or should they start their start-up adventure straight after graduation?

Secondly, this research could be used to improve the accuracy and speed of the venture capital funding decision regarding the founder's characteristics. The resulting model could be used to automatically distinguish between founders that will be funded and founders that will not be funded. This decision is extracted from the VC history funding decisions. Automatically selecting potential successful founders or applying insights of this research could lead to a higher success rate of investments. It could bring VC funds closer to their goal of detecting and funding only successful future start-ups. Besides improving the success rate, selecting entrepreneurs using the in this research constructed models could speed up the VC funding process and save time.

1.4. Structure of Thesis

The rest of this paper has the following structure. Firstly, discussion of previous literature is present in section 2. The formulated hypotheses and conceptual model will be discussed in this section as well. Secondly, the data will be discussed in section 3 and the research methodology in section 4. Thirdly, section 5 presents data analyses and results. Finally, section 6 contains general conclusions of the results, academic contributions, and managerial implications of this research. Additionally, this section presents limitations of this paper and suggestions regarding future research.

2. Literature Review

This section focuses on earlier literature regarding the role of founders in the start-up process. This study builds on insights of two streams of literature. Firstly, this paper builds on the determinants of start-up venture capital (VC) funding (in section 2.2). This paper contributes to research regarding the role of entrepreneur's human capital in the success of a start-up (in section 2.3). Within this stream we focus on research regarding founder's education, work experience and social capital. Secondly, literature regarding the prediction of entrepreneurial success is discussed in section 2.4. This paper closes the gap between these two literature streams by considering the role of founders' human capital in successfully acquiring VC funding and using founder's human capital to predict VC funding.

2.1. Definition of Start-ups and VC Funding

To discuss the funding of start-ups by VC firms, it is important to first discuss most important concepts: start-ups and the VC funding. According to the definition of Ries (2011) a start-up is "a human institution designed to create a new product or service under conditions of extreme uncertainty". This extreme uncertainty shows up in a high start-up failure rate. Recent data suggests that 11 of 12 entrepreneurs fail with their start-up (Startup Genome, 2020). Venture capital funding refers to the form of private equity funding that comes from a venture capital firm. As the name already suggests, these firms do undertake high risk to accomplish above average returns. Therefore, VC funding is very suited for the external financing of the start-ups which operate under extreme uncertainty. Typically, different types of funding rounds are available for start-ups. The first equity funding stage is the seed funding stage which represents the first funding of the start-up. After successfully receiving seed funding, two out of the three (67%) (tech) start-ups will fail (CB Insights, 2018). The seed funding stage is followed by Series A Funding, which is typically funded by VC firms. Thereafter, different funding stages can take place (Series B Funding, Series C Funding, etc.) until an Initial Public Offering (IPO) becomes a solution to collect funding. The VC funding process can therefore be seen as a funnel with successful start-ups as a final product.

2.2. Determinants VC Funding

As stated in the introduction, economic theories of Solow (1956) and Hasan & Tucci (2010) argue that technical progress accounts for almost all economic growth in a country. Chandy and Tellis (2000) emphasize the importance of start-ups in this process by discovering that "over a 150-year period, small firms and nonincumbents introduce more radical product innovations than large firms and incumbents". Therefore, discovering drivers of VC funding, which helps building the perfect theoretical start-ups, can help to acquire more technical innovation (and thus economic growth).

A long history of research exists that tries to find important determinants for venture capitalists funding decisions. Using factor analysis on questionnaire data, Tyebjee and Bruno (1984) discover five dimensions in the investment decisions; market attractiveness, product differentiation, managerial capabilities, environmental threat resistance, and cash-out potential. MacMillian, Siegel and Narasimha (1985) also spread a questionnaire among venture capitalists and discover similar findings as Tyebjee and Bruno (1984); venture's product, market, financials, and entrepreneur determine the venture capitalists funding decisions. However, MacMillian et al. (1985) emphasize that the quality of the entrepreneur is the most important decision criteria.

In contrast with research using questionnaire data before funding took place (Tyebjee & Bruno, 1984; MacMillian et al., 1985), MacMillan, Zemann and Subbanarasimha (1987) considered the criteria weightings of VC firms using data after funding was granted to improve validity. In agreement with earlier research, the results still emphasize the importance of the quality of management. However, later research (Hall & Hofer, 1993) did not find a major role for the entrepreneur/entrepreneurial team in the funding decision, except at the extreme ends of the distribution entrepreneurial talent (very incompetent/competent). When testing these hypotheses, the underlying conceptual model in Figure 2.1 is tested.

Instead of determining funding decision criteria using factor analysis (Tyebjee & Bruno, 1984; MacMillan, Siegel & Narasismha, 1985; MacMillan, Zemann & Subbanarasimha, 1987), Riquelme and Rickards (1992) use conjoint analysis as a research method in the venture capital decision. The results of this research suggest that entrepreneur's experience and the existence of a prototype or unique features of the product are important criteria in the screening step.

2.3. The Role of Human Capital in Entrepreneurial Success

According to the discussed literature in section 2.2, the quality of the entrepreneur is an important driver of the VC capitalists' funding decision. In section 2.3 we focus on the literature regarding the role founder's human capital in the success of a start-up. The definition of 'success' in earlier research not only consists of receiving funding but can also depend on different proxies (See 'Response variable(s)' column in Table 2.1). We divide research towards the effect of human capital on entrepreneurial success in five different groups: education (section 2.3.1), industry experience (2.3.2), entrepreneurial experience (2.3.3), social capital (section 2.3.4) and opportunity costs (section 2.3.5). Table 2.1 contains a schematical overview of the to be discussed literature regarding the role of founder characteristics (section 2.3).

2.3.1. Education

A lot of research has been carried out to determine the effect of education on venture success. Van de Ven, Hudson, and Schroeder (1984) found that entrepreneurs with higher education levels are more successful in the development of (educational software) start-ups. A study of Sapienza and Grimm (1997) considered the effects of founder characteristics on the (multidimensional subjective assessment of) performance of small and regional short line railroads that were created since the 1980 in the United States. Sapienza and Grimm found out that more general education and more business education led to higher performance (of railroad ventures). Barringer, Jones and Neubaum (2005) found that founders of rapid-growth firms are better educated than founders of slow-growth firms. Later research of Ko and McKelvie (2018) confirms the positive signalling effect of founders' education level on the amount of acquired funding in all funding rounds. According to the literature, founders' education seems to play an important role in the venture's success.

Also, research to the role of top management team's education in business success can be used to better understand the effect of education. Homburg et al. (2014) considered the role of characteristics of chief marketing officers (CMO) on the likelihood of getting venture capital funding and the amount of the investment. Homburg et al. discovered that a company with a CMO who obtained an MBA degree has a higher likelihood of acquiring VC funding. In contrast, research on the role of top management teams by Nuscheler, Engelen & Zahra (2019) found no significant direct effect of top management team's education on achieving growth.

2.3.2. Industry Experience

Van de Ven, Hudson, and Schroeder (1984) did not find a significant effect of industry experience on company development. Roure and Maidique (1986) continued to analyse the effect of the founder's industry expertise level on the high-tech venture's start-up success. In this research the degree of success is derived from the fact that a firm has been incorporated more than three years, has reached $20 million sales, and has achieved after-tax profits greater than 5 percent of the sales. In contrast with the results of Van de Ven et al. (1984), a positive relationship was found by Roure and Maidique (1986) between founders' prior experience and the start-up success. Founders of successful companies had two or more years of prior experience in the same position as their current position in the new start-up. Besides, the track records of these successful founders consist of companies that are larger than 500 employees and are characterized by more than 25% growth. In line with Roure and Maidique Barringer and in contrast with Van de Ven et al. (1984), Barringer, Jones and Neubaum (2005) discovered that founders rapid-growth firms have more relevant industry experience than slow-growth firms. In addition, Homburg et al. (2014) found similar results for CMO's industry experience; industry experience has a positive effect on the amount of venture capital funding. Baptista, Karaöz and Mendonça (2014) discover that industry experience mainly plays a role on the early survival of entrepreneurs who left previous employment to start a new firm. However, industry experience is less important for the survival of unemployment-driven entrepreneurs.

2.3.3. Entrepreneurial Experience

Besides researching the effect of education and industry experience, Van de Ven, Hudson, and Schroeder (1984) also examined the years of prior small business experience with the successful development of a (educational software) start-up. Surprisingly, prior experience in small businesses was found to be negatively related with start-up success. Van de Ven et al. explain this unexpected relationship by the strongly negative correlation between education and small business experience; the presence of education and small business experience seem to be mutual exclusive. Zhang (2007) focused on the role of prior entrepreneurial experience. Zhang expected that entrepreneurs with prior founding experience have more skills and social connections than novice entrepreneurs, which will lead to advantages in the venture capital raising process. Considering the first round of financing, the results suggests that experienced founders whose firms were not venture-backed before did not have an advantage of novice entrepreneurs. However, considering all rounds of financing, entrepreneurs with prior founding experience do appear to raise more venture capital than novice entrepreneurs. In line with Zhang (2007), Baptista, Karaöz and Mendonça (2014) found that entrepreneurial experience plays an important role in the survival of start-ups.

2.3.4. Social Capital

Literature also emphasizes the importance of founder's social capital. As one of the first researchers, Shane and Cable (2002) investigate the importance of social ties in the process of venture finance. Shane and Cable emphasize the existence of an information asymmetry problem (due to information disclosure and opportunistic behaviour from the entrepreneur's perspective) between the entrepreneurs and the potential investors. This information asymmetry (less information about ability of entrepreneur and the viability of the business plan than the entrepreneur) could lead to investors not willing to invest unless large, irreversible, credible commitments to the venture have been made by the entrepreneurs which lead to market failures according to the theory of Akerlof (1978). Shane and Cable (2002) argue that the existence of direct and/or indirect social ties could reduce the information asymmetry from the perspective of the venture capitalist by providing private information about the entrepreneurs and their 'opportunities' (Burt, 1992). Furthermore, social ties could lead to social obligations which cause generously behaviour between the entrepreneurs and founders (Gulati, 1995).

Shane and Cable (2002) showed that social ties are positively related to the probability of venture funding. However, since adding the reputation variable (information becomes public) mitigates the effect of social ties, the results suggests that investors use their social ties mainly to gather private information. However, Banerji and Reimer (2019) still found a positive association between social capital and fundraising amount since they found a positive correlation between the number of LinkedIn followers of a founder and the amount of VC fundraising of the start-up.

In contrast to Shane and Cable (2002) and Banerji and Reimer (2019), Hsu (2007) and Zhang (2007) investigated the importance of social ties because of founder experience and thus from the perspective of the entrepreneur directly. Hsu (2007) concluded that entrepreneurs with successful prior founding experience are more likely to receive venture capital funding through a direct tie compared through entrepreneurs with unsuccessful prior founding experience. Besides, the presence of a doctoral degree holder results in a higher likelihood of receiving venture capital funding via a direct tie in the (emerging) internet industry. Zhang (2007) discovered that founders with venture-backed founding experience tend to raise more venture capital, suggesting the importance of connections with venture capitalists. In accordance with the positive effects of social capitals found by Hsu (2007), Zhang (2007) and Banerji and Reimer (2019) in venture capital funding, Mollick (2014) discovers that a larger network size is associated with higher likelihood of successful funding though crowdfunding.

2.3.5. Opportunity costs

According to economists John Stuart Mill the amount of total foregone benefit of the best alternative of a certain choice should be recognized as costs (Stigler, 1995). These costs are also known as opportunity costs since they capture the costs of the best opportunity.

Cassar (2006) discover that entrepreneur's opportunity costs are a significant determinant of the intended scale of venture activity. He discovered that entrepreneurs with higher opportunity costs, measured through household income and managerial experience, intend on being involved in ventures with larger future sales revenue.

Until now, no research has been performed to determine the effect of opportunity costs on the likelihood of successful VC funding. Baptista, Karaöz and Mendonça (2014) found out that various forms of human capital factors of opportunity-based entrepreneur (who have higher opportunity costs) have more effect on the early survival changes than the pre-entry capabilities of necessity-based entrepreneurs.

2.4. Methodological Literature on Predicting Entrepreneurial Success

Earlier discussed research is focussed on discovering the size and the direction of the effect of a certain founder characteristic on entrepreneurial success. In contrast, some researchers and practitioners are solely focussed on predicting the success of business ventures. Due to the presence of aggregated data about start-ups through platforms, the ability to forecast entrepreneurial success is possible by using predictive analytics. One of largest of these platforms, which is used to collect data from thousands of companies, is Crunchbase. In Table 2.2 a schematic overview of the methodological literature is present.

Xiang et al. (2012), Arroyo et al. (2019) and Żbikowski & Antosiuk (2021) use the Crunchbase platform to collect aggregated company data. These researchers define entrepreneurial success differently. Xiang et al. (2012) define success as the occurrence of a company acquisition, while Arroyo et al. (2019) define venture success as the occurrence of an acquisition, any funding round or IPO's. Żbikowski & Antosiuk (2021) follow a similar definition as Arroyo et al. (2019), however a business venture is assumed only to be successful if series B venture capital fundraising is completed.

Although they use the same data provider, different variables are used to predict entrepreneurial success. While Arroyo et al. (2019) and Żbikowski & Antosiuk (2021) only use numerical features in building their predictive model, Xiang et al. (2012) use Latent Dirichlet Allocation (LDA) to discover topics in company related news. The most recent research from Żbikowski & Antosiuk (2021) points out the possible presence of a look-ahead bias in the research of Xiang et al. (2012) and Arroyo et al. (2019), since they use features that are not known at the time of assessment of a start-up (investing/funding or not). For example, Xiang et al. (2012) uses topics extracted from news articles as features which can contain information about funding events or venture capital backing.

Different predictive models are used. Xiang et al. (2012) use Bayesian Network to predict the occurrence of a merge or acquisition. Arroyo et al. (2019) use five different machine learning classifiers, such as Support Vector Machines (SVM) and tree-based ensemble algorithms. Besides these SVM and tree-based ensemble algorithms, Żbikowski & Antosiuk (2021) also use a logistics regression to build a predictive model, which is also a common method in earlier discussed research (Shane & Cable, 2002; Mollick, 2014; Baptista et al., 2014).

Table 2.1: Schematic overview literature regarding effects of founder characteristic.

Founder characteristic	Article	Explanatory variable(s)	Response variable(s)	Empirical Strategy	Method	Findings
Education	Van de Ven, Hudson & Schroeder (1984)	Level of education (high school, 1–3-year college, etc.)	Stage of the company (low vs. high-performing companies)	Data of 14 US based educational-software companies in 1983	t-test	Founders with higher education levels are more successful
Education	Sapienza & Grimm (1997)	Years of formal education (beyond high school) & Number of business courses	Goal Achievement and growth in employees	Survey data 70 US short line railroads	OLS regression	Positive effect general education and positive curvilinear effect number of business courses.
Education	Barringer, Jones & Neubaum (2005)	Presence of higher/college education, & relevant industry experience	3- year compound annual sales growth (>80%)	50 rapid-growth and 50 slow-growth firms	t-test	Founders of rapid growth have founders which are better educated.
Education	Ko and McKelvie (2018)	Number of years of highest education	Amount of VC funding	235 new ventures	Ordinary and Two-Stage Least	Founder’s education has a positive significant signalling

					Squares regressions	influence on the amount of VC funding
Education	Homburg et al. (2014)	CMO’s MBA education	Likelihood and amount of venture capital funding	2,945 high-technology new ventures	Hazard rate model	Positive effect of MBA education on the likelihood and amount of venture capital funding.
Education	Nuscheler, Engelen and Zahra (2019)	Dummy entrepreneurship education within top management team	Venture growth in employment	374 US based companies	GMM-SYS regressions	No significant effect from education on employment growth
Industry Experience	Van de Ven, Hudson & Schroeder (1984)	Level of education (high school, 1–3-year college, etc.)	Stage of the company (low vs. high-performing companies)	14 US based educational-software companies in 1983	t-test	Industry experience is unrelated to company development
Industry experience	Roure & Maidique (1986)	Dummy indicating founder has more than two years’ experience in similar position, and percentage of prior joint experience.	Incorporation years, sales, and profit	8 high-technology start-ups	Exploratory research	Position experience and prior joint experience is positively related with start-up success.

Industry experience	Barringer, Jones & Neubaum (2005)	Relevant industry experience	3- year compound annual sales growth (>80%)	50 rapid-growth and 50 slow-growth firms	t-test	Founders of rapid growth have higher industry experience.
Industry experience	Homburg et al. (2014)	Years of prior experience in same history	Likelihood and amount of venture capital funding	2,945 high-technology new ventures	Hazard rate model	Positive effect of industry experience on the likelihood of acquiring venture capital funding.
Industry experience	Baptista, Karaöz and Mendonça (2014)	Number of years since founder entered the labour market	Dummy if firm is operating three years after founding	Portuguese Survey covering 145,000 firms, 170,000 establishments and 2,000,000 workers	Logit regression	Industry experience mainly plays a role for opportunity-based entrepreneurs on the early survival of success.
Entrepreneurial experience	Van de Ven, Hudson & Schroeder (1984)	Years of small business experience	Stage of the company (low vs. high-performing companies)	14 US based educational-software companies in 1983	t-test	Small business experience is negatively related with the start-up success.
Entrepreneurial experience	Zhang (2007)	Dummy indicating founding experience	Time until VC funding and	11,029 venture-backed companies	OLS regression	Positive effect entrepreneurial experience on time until

			amount of VC funding	and 22,479 funding rounds		funding and the amount of funding (in all funding rounds).
Entrepreneurial experience	Baptista, Karaöz and Mendonça (2014)	Dummy: founded at least one firm	Dummy if firm is operating three years after founding	Portuguese Survey covering 145,000 firms, 170,000 establishments and 2,000,000 workers	Logit regression	Entrepreneurial experience mainly plays a role for all entrepreneurs on the early survival of success.
Social capital	Shane & Cable (2002)	Likert-scale survey questions regarding direct/ indirect social ties and reputation	VC funding and business angels funding	50 high-technology ventures	Logit regression	Positive effect social ties on funding decision, which mitigates if reputation variable is added.
Social capital	Banerji & Reimer (2019)	Network size: number of LinkedIn followers	VC fundraising per company year	150 US information and technology start-ups	Correlation	Positive correlation between network size and founding experience with the average amount of VC funding per year.
Social capital	Zhang (2007)	Dummy indicating venture-backed experience	Time until VC funding and	11,029 venture-backed companies	OLS regression	Positive effect of venture-backed experience on the

			amount of VC funding	and 22,479 funding rounds		time until funding and the amount of funding.
Social capital	Hsu (2007)	Number of founded start-ups, presence of MBA/PhD degree	VC funding via direct social ties	149 technology start-up firms	Probit and OLS regression	Positive effect of founding experience and positive effect PhD Degree in internet industry on likelihood VC funding via direct social ties.
Social capital	Mollick (2014)	Network size: Number of Facebook friends	Successful crowdfunding	48,526 crowdfunding projects	Logit regression	Personal networks are positively associated with the likelihood of reaching funding goal.
Opportunity costs	Cassar (2006)	Household income. Managerial experience in years	Intended future firm sales	5000 entrepreneurs	OLS regression	Individuals with higher current household income and greater managerial experience are associated with higher intended future firm sales

Table 2.2: Schematic overview methodological literature on predicting entrepreneurial success.

Article	Explanatory variable(s)	Response variable(s)	Empirical Strategy	Methods	Findings
Xiang et al. (2012)	Topics in news articles that contain funding news, basic statistics of a company, financial features, and managerial features/.	Company acquisition	59,361 companies and 38,617 news articles	Bayesian networks	Recall between 56.5% and 59.9% for companies with no matching articles.
Arroyo et al. (2019)	Founder characteristics	Company acquisition, IPO, or receiving any funding round	Data of 120,507 companies worldwide	Decision Trees, Random Forest, Extremely Randomized Trees, Gradient Tree Boosting, Support Vector Machines	Aggregated precision for positive classes between 45% and 68%
Żbikowski & Antosiuk (2021)	Founder characteristics	Company acquisition, IPO, or receiving at least a Series B funding round	Data of 213.171 entrepreneurs worldwide	Logistic regression, Support Vector Machines, XGBoost	Aggregated precision for positive classes between 45% and 68%

2.5. Hypothesis Development and Conceptual Model

Education

Earlier research suggests a positive effect of founder's education on the easiness of acquiring venture capital. Research of Van de Ven et al. (1984), Sapienza and Grimm (1997) and Barringer et al. (2005) conclude that education is an important driver of entrepreneurial success. Therefore, I expect that the presence of a degree acts as a positive signalling effect from founders towards venture capitalists, and the first hypothesis is formulated as follows:

Hypothesis 1. Having one or more degrees has a positive effect on the likelihood of acquiring venture capital funding.

Ko and McKelvie (2018) encoded founders' education as the number of years of the highest completed degree and found a positive signaling effect on the funding process. Sapienza and Grimm (1997) encoded founders' education in a similar way and found a positive effect on the growth of employees. Therefore, I expect that more years of education lead to a higher likelihood of acquiring venture capital funding:

Hypothesis 2. The effect of education on the likelihood of acquiring venture capital funding increases as the duration of the total education increases.

A novel element of this thesis is that the subject of the last degree is considered. Earlier discussed research (Sapienza & Grimm, 1997; Homburg et al., 2014) considered general and business education. Homburg et al. discovered a positive effect of business-related education on the likelihood of receiving VC. Therefore, in line with these results I expect that the presence of a related last degree subject will positively influence the VC process. The third hypothesis is:

Hypothesis 3. The effect of education on the likelihood of acquiring venture capital funding increases as the subject of the last degree is related to the start-up industry group.

Earlier research has discovered that higher education levels lead to more success in the development of start-ups (Van de Ven, Hudson & Schroeder, 1984). Later research of Barringer, Jones & Neubaum (2005) confirmed these conclusions by discovering that founders of rapid-growth firms are better educated than founders of slow-growth firms.

Following the rationale of these papers which states that better education led to more success, I expect that higher education levels increase the likelihood of acquiring VC:

Hypothesis 4. The effect of education on the likelihood of acquiring venture capital funding increases as the level of the last degree increases.

Industry Experience

The results of literature regarding the importance of industry experience show mixed effects. Van de Ven et al. (1984) concluded that industry experience is unrelated to company development and Baptista et al. (2014) concluded that this type of experience only plays a role for opportunity-based entrepreneurs. In contrasts, most of the discussed literature (Roure & Maidique, 1986; Barringer, Jones & Neubaum, 2005, Homburg et al., 2014) did find positive effects of industry experience on entrepreneurial success. The fifth hypothesis is formulated as follows:

Hypothesis 5. The presence of industry-related experience has a positive effect on the likelihood of acquiring venture capital funding.

Besides considering the industry of the last employer, I also consider the total working experience of the founder in a similar way as Żbikowski and Antosiuk (2021). Since it is expected that industry experience positively influences the VC funding process, I expect that this influence becomes bigger when the total working years increases.

Hypothesis 6. The effect of industry-related experience increases as the number of years between the graduation and start-up founding date increases.

As Żbikowski and Antosiuk (2021) suggest for future work, text data is explored as additional set of features to capture industry experience. This thesis captures features from founders' profile descriptions. I expect that the presence of industry related topics in founders' profile descriptions signals the extent of industry experience to venture capitalists. The hypothesis is formulated as follows:

Hypothesis 7. The extent of industry experience derived from founders' Crunchbase profile descriptions is positively associated with the likelihood of acquiring venture capital funding.

Entrepreneurial experience

While Zhang (2007) and Baptista et al. (2014) concluded that entrepreneurial experience has a positive relationship with entrepreneurial success, only Van de Ven et al. (1984) discovered a negative relationship between entrepreneurial experience and start-up success. Therefore, the eight hypothesis is:

Hypothesis 8. The number of founded organisations positively influences the likelihood of acquiring VC funding.

In a similar way to detect the extent of industry experiences, entrepreneurial related topics in founder's Crunchbase profile descriptions will be detected as signals for the extent of entrepreneurial experience. It is expected that the extent of entrepreneurial experience has a positive signalling effect towards venture capitalists and therefore the presence of entrepreneurial related topics positively influences the VC process:

Hypothesis 9. The extent of entrepreneurial experience derived from founders' Crunchbase profile descriptions is positively associated with the likelihood of acquiring venture capital funding and the amount of venture capital.

Social Capital

Shane and Cable (2002), Banerji & Reimer (2019), Zhang (2007), Hsu (2007), and Mollick (2014) have similar conclusions regarding the influence of social capital on the VC process. They discover positive effects of founder's social capital on the entrepreneurial success exists through acquiring VC funding. Therefore, I expect that higher social capital through the presence on social media platforms or start-up hubs positively influences the VC funding likelihood. Besides, it is expected that more press appearances can be a result of social capital (and vice versa) and as a result also positively influences the likelihood of receiving VC funding. The hypotheses are formulated as follows:

Hypothesis 10. The presence on social platforms positively influences likelihood of acquiring venture capital.

Hypothesis 11. More press appearances of founders positively influence the likelihood of acquiring venture capital.

Hypothesis 12. Having a headquarter located in a start-up hub positively influences the VC funding likelihood.

Opportunity costs

The potential benefits an entrepreneur misses when choosing to start his start-up are referred as opportunity costs. Quitting a job to build a new start-up go hand in hand with higher opportunity costs for the entrepreneur. In contrast, building a start-up without having any employment results in having relatively low opportunity costs. Cassar (2006) discovered that higher opportunity costs lead to a higher intended scale of venture activity. If an entrepreneur starts a new company while having relatively higher opportunity costs, this could signal to venture capitalists the business opportunity of the start-up and/or the internal motivation of the founder. Therefore, in the following hypothesis I expect a positive signalling effect of (quitting) past employment and thus having high opportunity costs.

Hypothesis 13. Presence of employment shortly before founding a new startup positively contributes to the likelihood of acquiring venture capital.

In addition, if somebody quits at a promising firm that received a lot of funding, the opportunity costs are higher since it is more likely that the employee had a better career perspective. Consequently, a higher positive signaling effect of presence of previous employment is expected:

Hypothesis 14. If previous employer received more funding, the positive signaling effect of previous employment on the likelihood of acquiring venture capital is bigger.

Baptista, Karaöz and Mendonça (2014) discover that specific human capital (education and work experience) plays a bigger role in business success for former employees who give up their employment to start a new business than for former unemployed founders. This is in line with their expectations since they argue that specific human capital may erode with unemployment spells (Neuman & Weiss, 1995; Albrecht et al., 1999). Therefore, in line with Baptista, Karaöz and Mendonça I expect a bigger presence of specific human capital at former employed founders. Specific human capital could therefore play a bigger role for these former employed founders in the VC funding process.

In addition, I expect a bigger role of specific human capital in the VC funding process for founders who gave up employment because they face higher opportunity costs. The founders who gave up their employment face more opportunity costs than founders who did not have any employment. I expect that these former employees therefore demand presence of more specific human capital before quitting their job and diving into the start-up adventure. All in all, because of this process and the process described by Baptista, Karaöz and Mendonça, former employee's (higher opportunity costs) who quit their employment to start a start-up have more human capital regarding education and work experience. Therefore, I expect that education and work experience play a bigger role in the VC funding process when opportunity costs are higher:

Hypothesis 15. Having higher opportunity costs through recent employment moderates the effect of education and work experience on the likelihood of venture capital such that the effect of education and work experience is higher when past employment is present than when it is not.

When testing these hypotheses, the underlying conceptual model in Figure 2.1 is tested.

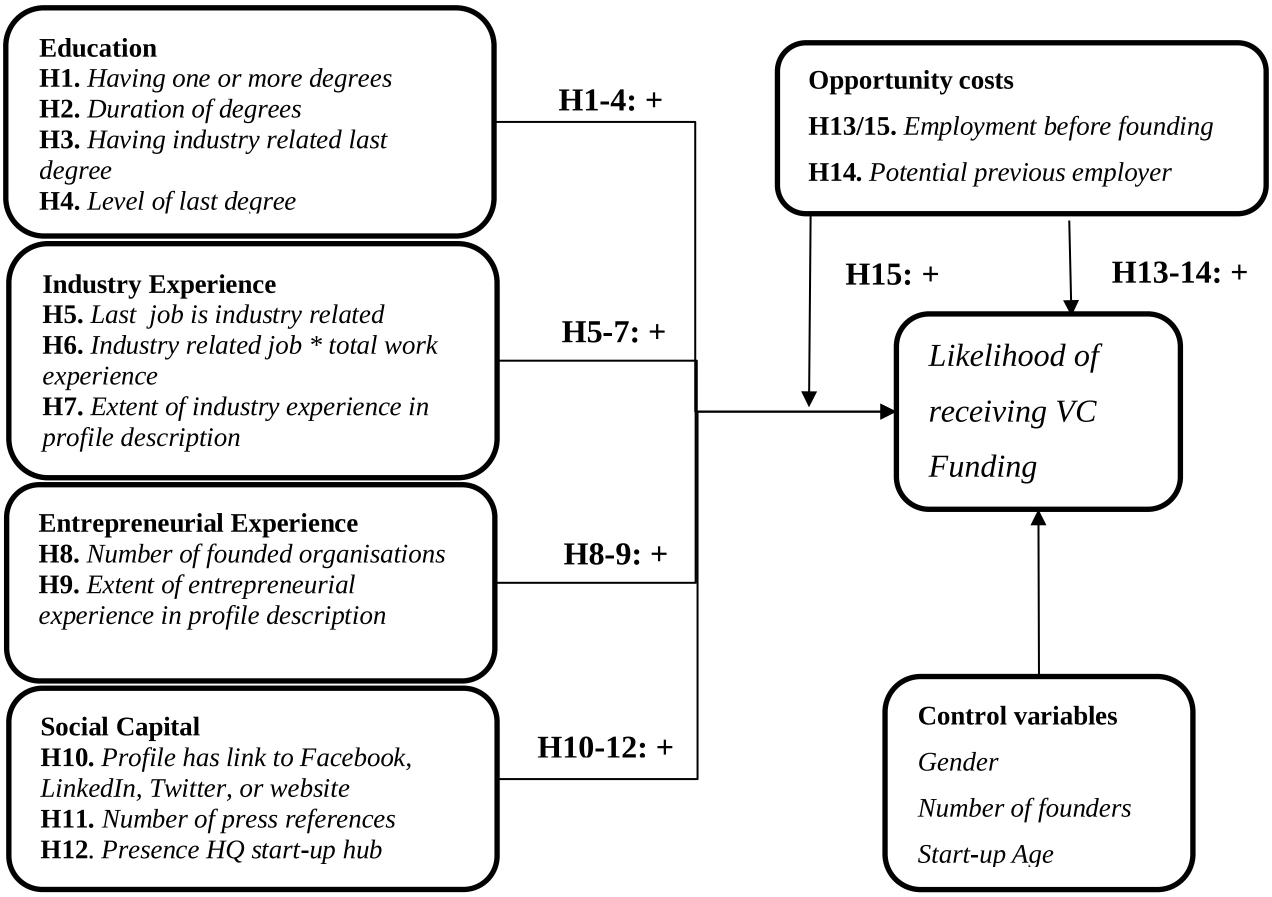

Figure 2.1: Conceptual model of founder characteristics and the likelihood of VC.

3. Data

3.1. Data Resources and Sample

This paper is focussed on finding founders' characteristics drivers for receiving venture capital funding. Multiple platforms, such as CB Insights, Crunchbase and Dealroom, were considered to collect aggregated business information of thousands of start-ups. Crunchbase is the platform with the most extensive information on founder level. Since this research focusses on founders' characteristics, Crunchbase's[1] database is used for this research. This database is also used in earlier discussed research (Xiang et al., 2012; Homburg et al., 2014; Ko & McKelvie, 2018, Banerji & Reimer, 2019; Arroyo et al., 2019, Żbikowski & Antosiuk, 2021) by using access to daily snapshots. In contrast, this thesis uses Crunchbase's API access which has one main advantage over daily snapshots: automatically updating of the dataset. This makes it possible to generate the results of this research automatically for different timeframes but also for different regions and industries.

The goal of this research is two-folded. Firstly, this research tries to understand the relationships between founder characteristics and the VC funding likelihood. Secondly, this paper tries to predict funding likelihood as accurate as possible using founder characteristics. When predicting VC funding it is important to stay away from the look-ahead bias, which is the result of using information as predictors that is not known at the decision moment (Żbikowski & Antosiuk, 2021). In practice, on a specific moment a venture capitalists can only use founder information about the past and not about the future for his funding decision. To prevent this look-ahead bias we use a 3-year warm-up window between t_c (January 2010) until t_s (December 2013) in which we gather founder information and a five-year simulation window t_s (January 2014) until t_r (December 2019) in which we gather VC funding information (see Figure 3.1). This more realistic time-aware approach is also used by Arroyo et al. (2019). Figuratively speaking, I simulate a situation in which a venture capitalist needs to predict the funding decision at the end of 2013.

[1] The author of this thesis applied for Crunchbase's academic research access program to receive access to the Crunchbase dataset. This program enables fully free or discounted access to Crunchbase datasets on case-by-case basis. After basic access API was granted, the author applied for full API access, which eventually was granted by Crunchbase for free.

Using only information available beforehand has also an advantage for understanding the effect of founder characteristics on the funding likelihood. Reversed causality, which assumes an effect of funding on founder characteristics (such as press references), is prevented by using only information available at the end of the warm-up window (December 2013). It is not possible that the funding process in the simulation frame influences the characteristics known at the end of the warmup period. However, vice versa it is possible that founder characteristics influence the venture capital likelihood. The use of a warmup and simulation window therefore successfully reduces the risks of reversed causality.

Since a start-up is assumed to create a new product or service (Ries, 2011), this paper needs to distinguish firms in new ventures and more adolescent ventures. By requiring a firm to be founded between t_c and t_s, we focus on companies that are founded between 2010 and 2013. This paper is focused on companies founded within one country, since comparing funding performance of start-ups within one country prevents the influence of differences in macro-economic factors across countries which are discovered in earlier research (Gompers & Lerner, 1999; Jeng & Wells, 2000). This paper focusses on the United States since this country has the most start-up observations in Crunchbase of all countries worldwide.

This research focusses on one single industry group (defined by Crunchbase) since using a single industry group reduces alternative explanations for human capital requirements, such as a medical degree, which is also confirmed by Ko & McKelvie (2018). Industry groups are broader subjects that encompass multiple industries in Crunchbase. For the exact definition/composition of industry groups we refer to Crunchbase industry group list (Crunchbase, n.d.-a). I focus on the “Information Technology” (IT) industry group since relatively many companies within this industry groups are represented in Crunchbase. The fact that Crunchbase is a spin-out of TechCrunch, which is focussed on technology related start-ups, can be the underlying reason for the representation of this industry group in the data of Crunchbase.

The initial data collection of the above API search query contained 11,982 IT start-ups that were based in the US and were founded between 2010 and 2013. The search query was fine-tuned taking account the following. First, non-profit start-ups are eliminated since the venture capital process may differ compared to the process of profit start-ups.

As a result, the number of start-ups captured in the data collection decreased to 11,522. Secondly, start-ups that underwent an IPO, were acquired, or closed before the start of the simulation timeframe were eliminated from the database since these firms are not able to receive venture funding in this timeframe. The total number of start-ups decreased to 11,340. Thirdly, start-ups with a missing link to the founder were deleted since in this case extracting founder characteristics is not possible. After this last elimination, the data contains information about 5,363 start-ups, which means that the initial Crunchbase database contains a lot of missing links with founders. These 5,363 start-ups are collectively founded by 9,504 different founders. Considering the results of an independent sample t-test (t=36.95, p=0.00), the mean funding ratio of the start-ups included in the final dataset (17.25%) is significantly higher than the mean funding ratio of start-ups that are not included in the final dataset due to a missing link (0.19%). Therefore, the link seems not missing at random which suggests a possible selection bias towards more successful start-ups and founders (that receive VC funding) in the final dataset.

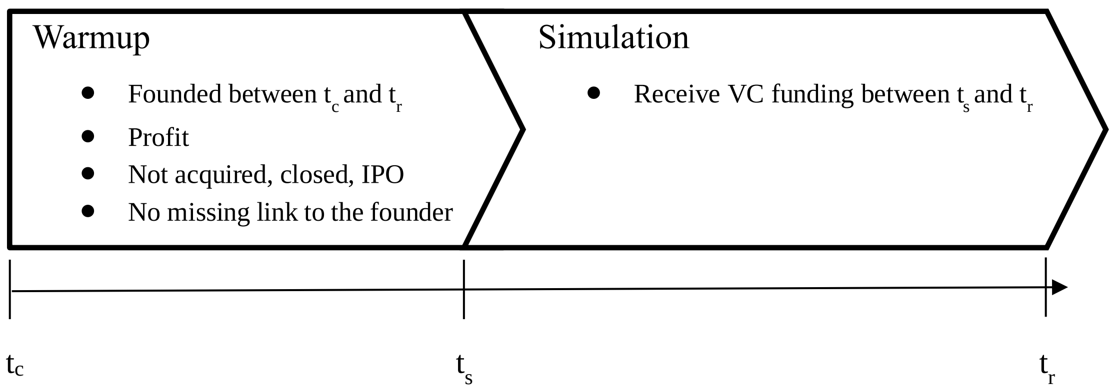

Figure 3.1 Warmup and simulation window.

3.2. Dependent Variable

Dummy VC funding: In this research a dummy variable is used which indicates whether a founder's start-up received VC funding or not. To distinguish between venture capital and non-venture capital investments, the variable *investment_stage* for each funding round is used which captures the funding stage of a funding round. This variable can be equal to 'early-stage venture' or 'late-stage venture'. Early-stage ventures consists of ventures that received funding in funding rounds Series A and/or Series B. Late-stage ventures received at least Series C funding (and maybe more VC funding rounds). If any of the funding rounds contain an early-stage or late-stage venture investment stage in the simulation timeframe, the *Dummy VC funding* is equal to 1 and 0 otherwise. For example, if no (known) funding or only a seed funding round (or earlier funding rounds) is reached, then the dummy variable is equal to 0.

3.3. Independent Variables

Education. To capture the education of a founder, different features are constructed. First, a dummy variable is created that captures if a founder has completed a degree (*has_degree*) before the start of the simulation date (01-01-2014). The subject text of the last degree (before the start of the simulation) is obtained to determine if the last degree's subject is industry-related (IT related). This is done by encoding a dummy variable equals to 1 if the subject description of the last degree contains the words "computer" or "information" (such as Computer Science, Information Science, Information Technology) and 0 otherwise. The type of the last degree is encoded by different dummy variables that capture if a founder's last degree was a PhD (*phd)*, MBA (*mba*), master's (*ms*), or bachelor's (*bs*) degree. The quality of all education is considered by calculating the length of all education in months (*duration_degree)*, which is also done by Żbikowski & Antosiuk (2021). If the graduation date is not available, the starting date of the first job is used as a graduation date to determine the duration of all degrees.

Industry Experience. To measure industry experience, dummy variables are constructed (*industry_related_job*, *industry_related_startup*) which capture if founder's last job or last founded start-up before founding the start-up is in the same industry as the start-up industry group (IT). The variable *total_work_experience* considers the number of years between the closing date of the last degree and the founding date of the new start-up. This variable captures total work/professional experience in years as done by Żbikowski and Antosiuk (2021). A missing graduation date of the last degree results in a missing value. Missing values are replaced with a 0 by Żbikowski and Antosiuk (2021). However, this results in people without having a degree (and thus a closing date of the last degree) in a zero *total_work_experience*. Therefore, if the graduation date is not available, *total_work_experience* is calculated using the difference between the start date of the first job and the founding date of the start-up. Missing values that still arise, are then replaced with a 0. In addition, the information inside founder's profile description is used to mine industry-related topics (*industry_related_topic_description*) in a similar way Xiang et al.

(2012) applied LDA on news articles. The topic mining is discussed extensively in section 3.6.

Entrepreneurial Experience. The number of organisations that person founded before the beginning of the simulation window measures the entrepreneurial experience of a founder (*num_founded_organization*). The presence of entrepreneurial topics (*entrepreneurial_related_topic_description*) is extracted in a similar way to capture industry experience using profile descriptions (Section 3.6).

Social capital. Banerji and Reimer (2019) use the number of LinkedIn followers as proxy for the size of the founder's social network. However, the number of followers is highly dependent on the time of assessing the social network. Instead of focussing on founders, Arroyo et al. (2019) adds dummy-variables that capture the presence on social media of start-ups. Since a high number of followers or start-up presence on social media platforms can be the result of business success through VC fundraising, a look-ahead bias can exist. Therefore, I create dummy variables which equal 1 if the founder personally is visible on a social media platform (*founder_has_facebook, founder_has_linkedin, founder_has_twitter, founder_has_website*). However, no date is available showing the creation of founder's social media profile. The risk of the look-ahead bias is thus not fully eliminated since a founder could create their social media profile later than the beginning of the simulation window. However, we assume that this bias is reduced compared to using the number of followers as feature since we assume that the presence on a social media platform is less dependent on business success than the number of followers. The variable *number_press_references* contains the number of press references which are linked by Crunchbase to a founder. The dummy variable *silicon_valley* takes the value 1 if a start-up's headquarter location group is equal to 'san-francisco-bay-area-california' and 0 otherwise to capture the social networking effects of the Silicon Valley area, which is known as the start-up hub of the United States.

Opportunity costs. As a proxy for the opportunity costs of a founder, a dummy called *opportunity_entrepreneur* is created that equals 1 if founder's last job ended within 12 months before or after the founding date of the start-up, and 0 otherwise. This variable *opportunity_entrepreneur* is an indication if a founder was employed, was used to receiving salary and thus had higher opportunity costs. The variable *last_job_funding* equals the total funding of the last employer measured in millions (USD) until the start of the simulation window and is a proxy for the potential growth of the last employer.

3.4. Control Variables

To control for omitted variable bias, different control variables are added. First, we control on founder level by adding variables that could be correlated with education or industry/managerial experience and influence VC funding. The founder's gender is added as dummy variable (*gender_male*) since research from Brush et al. (2018) suggests the presence of a gender gap in the venture capital fundraising process; all-men teams are four times more likely to receive VC funding than teams with even one woman on the team. Adding gender as control variable is in line with earlier research (Arroyo et al., 2019; Żbikowski & Antosiuk, 2021).

Second, we control on start-up level by adding the number founders (*num_founders*) as done by Ko and McKelvie (2018) since founding team size is linked to the growth of new firms (Eisenhardt & Schoonhoven, 1990). Ko and McKelvie (2018) aggregated founder's human capital of all founding team members. This is in contrast with Baptista, Karaöz and Mendonça (2014) who do not aggregate measures of entrepreneurial human capital. They argue that taking the sum or average of individual indicators would not produce an accurate measure of entrepreneurial human capital for the firm. In this research we focus on the effect of human capital of individual founders and not the effect of entrepreneurial teams. Therefore, we choose to only control for the number of founders and do not aggregate human capital for teams in a similar way as Baptista, Karaöz and Mendonça.

The firm's age is added as a control variable (*startup_age)*, in line with the research of Hsu (2007) and Ko and McKelvie (2018), because investor may prefer established ventures.

3.5. Descriptive Analysis: Descriptive Statistics

If the distribution of the dependent variable in Figure 3.2 is considered, it stands out that 23% percent of all start-ups received venture capital in the simulation timeframe. Given the fact that only 1 out of the 12 entrepreneurs succeed with their start-up (Startup Genome, 2020), this percentage seems to be relatively high. Crunchbase collects its data through contributors, publicly available sources, and various data partners (Crunchbase, n.d-b). The data collection therefore could be biased towards successful entrepreneurs and start-ups.

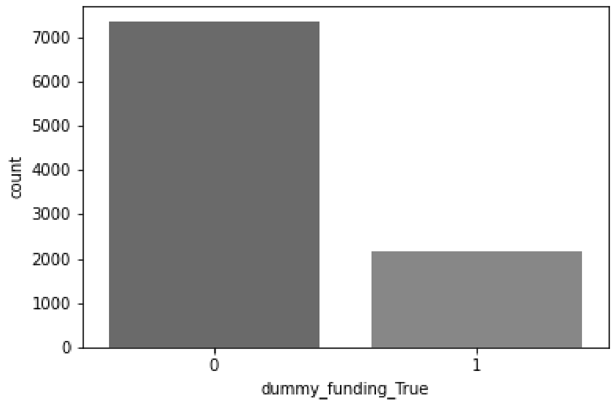

Figure 3.2 Distribution dummy Venture Capital funding.

If the education related variables are considered in Table 3.1, we see that 41% of all founders have at least one degree (before the start of the simulation window) on their Crunchbase page. From all founders only 10% of all founders followed IT related subject in his last degree, which is around 24% of all founders with a degree. Only 2%, 5%, 7% and 22% of all founders received a PhD, MBA, Master's, or bachelor's degree respectively. On average founders spend 8.36 months on education in this dataset. This average duration is relatively low since founders with no education have 0 months invested on education, which brings down the average.

Table 3.1 Descriptive statistics of used variable.

	mean	std	min	25%	50%	75%	max
dummy_funding_True	0.23	0.42	0.00	0.00	0.00	0.00	1.00
founder_has_degree_True	0.41	0.49	0.00	0.00	0.00	1.00	1.00
IT_True	0.10	0.30	0.00	0.00	0.00	0.00	1.00
phd_True	0.02	0.16	0.00	0.00	0.00	0.00	1.00
mba_True	0.05	0.23	0.00	0.00	0.00	0.00	1.00
ms_True	0.07	0.25	0.00	0.00	0.00	0.00	1.00
bs_True	0.22	0.41	0.00	0.00	0.00	0.00	1.00
duration_all_degrees	8.36	22.65	0.00	0.00	0.00	0.00	252.00
industry_related_job_True	0.15	0.36	0.00	0.00	0.00	0.00	1.00
industry_related_startup_True	0.04	0.20	0.00	0.00	0.00	0.00	1.00
total_work_experience	3.40	6.75	-4.00	0.00	0.00	4.00	45.00
last_job_type_executive	0.21	0.41	0.00	0.00	0.00	0.00	1.00
num_founded_firms	1.22	0.66	1.00	1.00	1.00	1.00	21.00
founder_has_linkedin_True	0.72	0.45	0.00	0.00	1.00	1.00	1.00
founder_has_facebook_True	0.19	0.39	0.00	0.00	0.00	0.00	1.00
founder_has_twitter_True	0.38	0.49	0.00	0.00	0.00	1.00	1.00
founder_has_website_True	0.18	0.38	0.00	0.00	0.00	0.00	1.00
silicon_valley_True	0.29	0.46	0.00	0.00	0.00	1.00	1.00
number_press_references	1.11	4.92	0.00	0.00	0.00	0.00	92.00
opportunity_entrepreneur	0.10	0.30	0.00	0.00	0.00	0.00	1.00
last_job_funding	9.24	133.05	0.00	0.00	0.00	0.00	6784000 .00
num_founders	2.32	1.21	1.00	1.00	2.00	3.00	8.00
gender_male	0.91	0.28	0.00	1.00	1.00	1.00	1.00
startup_age	24.98	13.61	0.00	12.00	24.00	36.0 0	48.00

15,00% of all founders had a job at an industry related employer (Table 3.1). The average number of founded firms on the start-up date is 1,121 (included the current start-up) and is skewed highly to the right (see Figure 3.3).

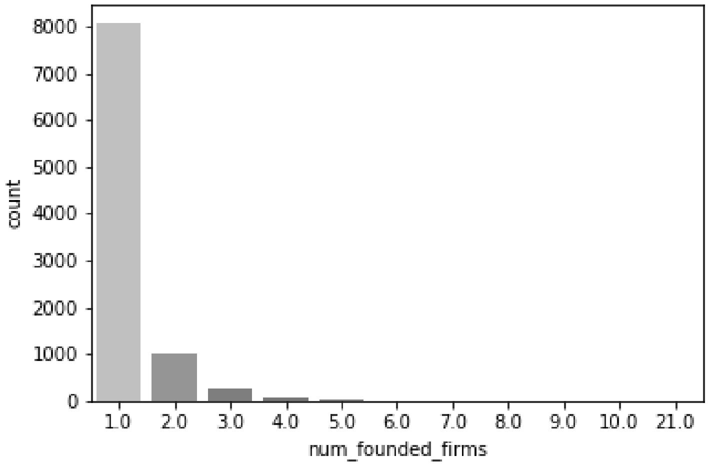

Figure 3.3 Distribution of number of founded firms before the start of simulation.

If consider the social related variables (Table 3.1) most of the founders have LinkedIn, followed by having Twitter, Facebook and an own website. On average founders are 1.11 times mentioned in a press reference before the simulation date started. Our dataset contains a lot of the start-ups are founded in Silicon Valley (29.48%).

If consider the mean of the variable *opportunity_entrepreneur* (Table 3.1), we observe that 10% of all founders ended their jobs within a timeframe of 12 months of the founding date. Considering the control variables, the average number of founders is 2.32 and the average start-up age is around 25 months. This suggests that the founding dates of the start-ups in the dataset are almost equally distributed over the warm-up period of 4 years.

3.6. Topic Analysis

Latent Dirichlet Allocation (LDA, also known as Topic Analysis) is applied on the Crunchbase's profile description of each founder. The main goal of LDA in this context is the detection of topic presence in each profile description. LDA is an unsupervised probability-based approach since it observes word frequency distribution among all profile descriptions to define the predefined number of topics (Kwartler, 2017).

Before applying LDA, each word is tokenized, and punctuations and unnecessary characters are removed as a pre-processing step. Stop words are removed and only unigrams are used since the detection of unigrams in topics seems sufficient to determine the extent of industry related and entrepreneurial related experience. After that, lemmatization of all words reduces words to their common base word form (lemma).

LDA assumes the following generative process. First, given the number of topics, get the topic weights for each profile description, which follows a Dirichlet distribution. Secondly, for each word in the profile description draw a specific topic randomly (using the topics distribution) and draw a specific term from the multinomial distribution (Hoffman, Blei & Bach, 2010). Since the profile descriptions are already known, the inverse of this generative process is used to determine the topic weights of each profile description. In this study I use an optimized LDA in Python which is based on the research of Hoffman, Blei & Bach (2010) because it can handle massive document collections and is available in Python through the *Gensim Package* (Rehurek & Sojka, 2011). A more extensive explanation of LDA (such as done by Tufts, n.d.) falls outside the scope of this paper.

It is necessary to determine the number of topics before running the LDA model. Therefore, hyperparameter tuning takes place to determine the number of predefined topics of the best LDA model considering topic coherence defined by Röder, Both and Hinneburg (2015). This Topic Coherence score is a measure that indicates the degree of semantic similarity between important words inside a topic. Less relative distance between words results in a high coherence score, and vice versa. Tuning the LDA model this way results in a LDA model with interpretable topics. If a range of topics between 2 and 40 topics is considered, the coherence score suggests a LDA model with 6 topics.

If the top 5 relevant words within each topic are considered in Table 3.1, I can see that the words ‘business’, ‘technology’, ‘company’, ‘experience’, and ’year’ are relevant within topic 6. Topics 3 and 4 have similar relevant words regarding the position of the entrepreneur. Therefore, these topics are the entrepreneurial related topics mentioned in hypothesis 9. Topic 0 seems to be more industry related since it contains the words: ‘security’, ‘university’, ‘system’, ‘engineering’, and ‘development’. Therefore, this topic can be seen as the industry related topic needed for hypothesis 7. No distinctive topics could be determined from topic 1 and 2. The dominant topic for each Crunchbase profile description is one-hot encoded, dropping topic 2 (general topic) to prevent perfect multicollinearity.

Table 3.2: Top 5 most relevant words from each topic learned by LDA. Topic 3 and 4 relates to entrepreneurial/managerial experience and topic 0 coincides to industrial experience.

Topic No	Top 5 words
0	security university system engineering development
1	company com work year help
2	company technology university year marketing
3	founder co ceo university technology
4	founder chief technology co officer
5	business technology company experience year

4. Research Methodology

The goal of this research is two-folded. On the one hand the goal is to determine the effect of founder's characteristics on the likelihood of raising venture capital (section 4.1). On the other hand, this research tries to predict VC fundraising as accurate as possible (section 4.2). This paper distinguishes between three categorical types of analytics: descriptive, prescriptive, and predictive analytics (Davenport & Harris, 2017). Descriptive analytics tells us what happened in the pasts, prescriptive analytics makes recommendations on how to improve the funding decision, and predictive analytics uses analytics to predict future VC funding. In section 3.5 the descriptive analysis is already discussed since it describes historical data of founders and VC funding data with no underlying model explaining this data. Section 4.1 presents the prescriptive analysis and section 4.2 presents the prescriptive analysis.

A logistic regression is preferred for the prescriptive analysis since the coefficients of a logistic (also called logit) can be interpreted to determine the effect of founder's characteristics and make recommendations. This logit version can also be used to predict VC fundraising. However, machine learning methods are preferred for prediction since more complex relationships can be modelled, which often results in better predictions in practice. Figure 4.1 displays the experimental setup of the model selection process for the prescriptive (left) and predictive analyses (right) schematically.

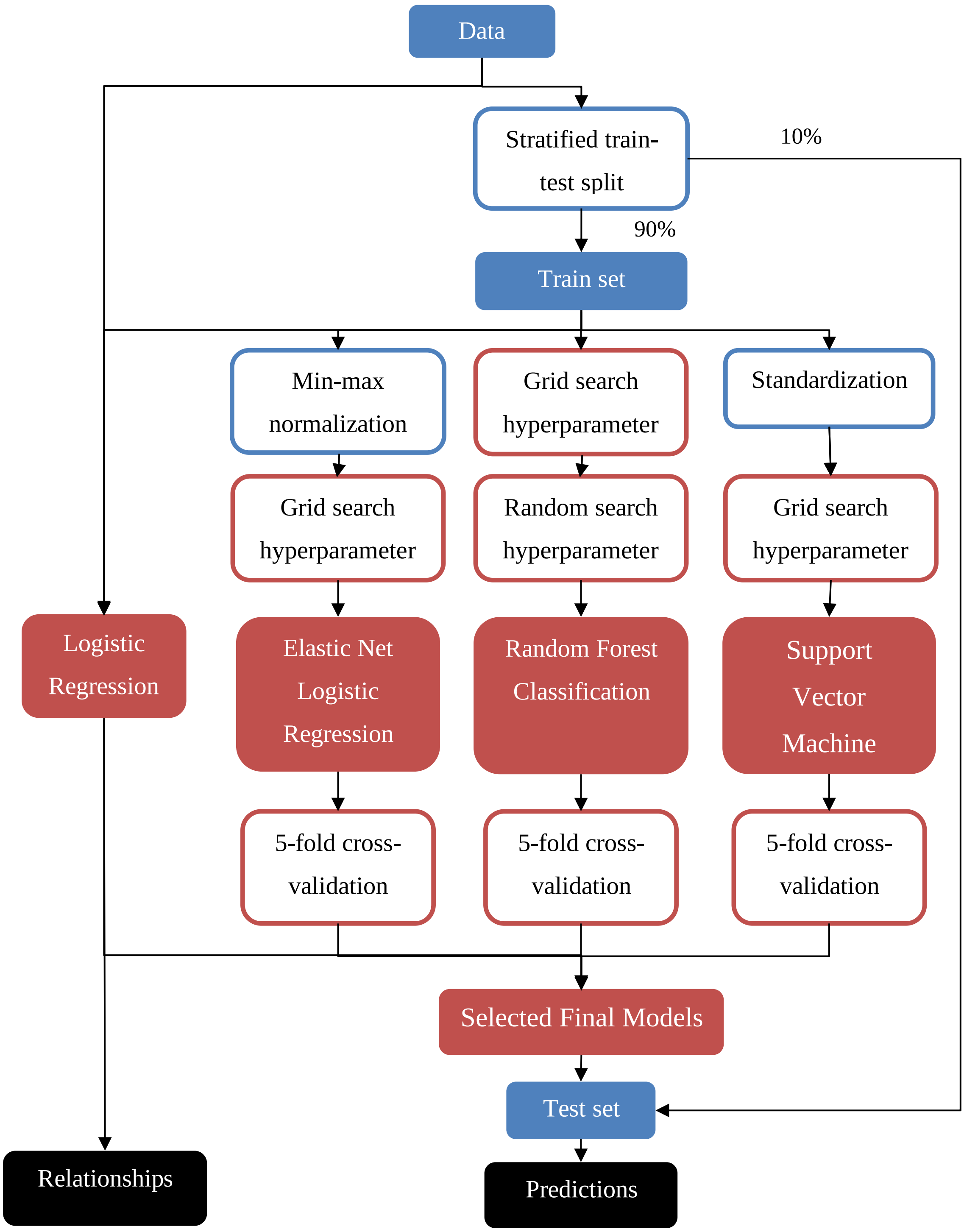

Figure 4.1: Experimental setup of prescriptive and predictive analyses

4.1. Prescriptive Analysis

Logistic regressions will be used to determine the effect of the independent variables on the likelihood of VC funding. Since receiving venture capital funding or not is a classification problem and the dependent variable therefore has bounded outputs (0/1), also known as binomial distribution. The underlying assumption of a linear model (LM) assumes a continuous normally distributed dependent variable. This assumption is violated due to the classification problem. Therefore, generalized linear models (GLM) offer a solution by providing a link (see equation 2) between the linear predictor (η_f, see equation 1) and the expected probability of the dependent variable equals 1 (π_f, see equation 3).

Using a logistic regression, which is a form of a GLM, to prescribe relationships is in line with most of the earlier discussed literature regarding the role of human capital in entrepreneurial success (Shane & Cable, 2002; Mollick, 2014; Baptista et al., 2014; Żbikowski & Antosiuk, 2021). The following logistic regression will be estimated:

$$\log\left(\frac{\pi_f}{1-\pi_f}\right) = \alpha_0 + \beta_1 E_f + \beta_2 I_f + \beta_3 M_f + \beta_4 S_f + \beta_5 O_f + \beta_6 L_f + \beta_7 O_f L_f + \beta_8 O_f E_f + \beta_9 O_f W_f + \boldsymbol{\beta_{10} C_f} + \varepsilon_f \quad (1)$$

Where π_f is the probability of interest for founder f, E_f represents the education related independent variables for founder f, I_f represents the industry experience related variables for founder f, M_f represents the entrepreneurial related variables for founder f, S_f represents the social capital related variables for founder f, O_f is the dummy variable that captures if founder f is an 'opportunity entrepreneur', L_f is the amount of funding of received by founders' last employer, and W_f is the working experience of founder f. $\boldsymbol{C_f}$ is a vector that contains the control variables discussed in section 3.4. The left side of equation 1 is called the logit link function ($g(\pi_f)$):

$$g(\pi_f) = \log\left(\frac{\pi_f}{1-\pi_f}\right) = \eta_f \quad (2)$$

Where η_f stands for the linear combination of all predictor variables (right side of equation 1). Equation 2 shows that using the logit link function assumes that the log odds of η for founder f is a linear combination all predictor variables (James, Witten, Hastie

& Tibshirani, 2013). The exponent of the β-coefficients of each predictor variable represent the effect on the odds ratios ($\frac{\pi_f}{1-\pi_f}$) of receiving venture capital funding.

If the inverse is derived from equation 2 (see equation 3), it can be concluded that the link function successfully maps the probability between the range 0 and 1 since $\frac{\exp(\eta_f)}{1+\exp(\eta_f)}$ can never be lower than 0 or higher than 1.

$$g^{-1}(\eta_f) = \frac{\exp(\eta_f)}{1 + \exp(\eta_f)} = \pi_f \tag{3}$$

In contrast to a linear model which minimizes the squares of residuals, a logistic regression maximizes the likilihood of the dataset. The R^2, which is used as indicator of fit, cannot be used to determine the fit of a logit model (Hauser, 1978; Hoetker, 2007). In this paper McFadden's likilihood ratio index (also known as McFadden's pseudo-R-squared, $R^2_{McFadden}$) is used to asses the fit of the logistic regression (McFadden, 1973):

$$R^2_{McFadden} = 1 - \frac{\log(L(X))}{\log(L_0)} \tag{4}$$

Where $L(X)$ stands for the maximized likihood of model X and L_0 is the likihood of the null model. For the execution of the logit model the *statsmodels* Python package is used (Seabold & Perktold, 2010) since the McFadden's pseudo-R-squared is available as property.

4.2. Predictive Analysis

I split the dataset into a training and test set using 90%-10% ratio. The distribution of the dependent variable is imbalanced, since only 22.74% of the start-ups in the dataset received venture capital (see also Table 3.1). Therefore, the split took place in a stratified fashion which resulted in an equal distribution of start-ups with and without funding in the simulation timeframe in the train and test set. This resulted in an availability of 8,551 instances for the training process and 951 instances for test process.

In a similar way as Arroyo et al. (2019) and Żbikowski and Antosiuk (2021), metrics such as precision (specificity), recall (sensitivity) and the F1 score will be used to determine the performance of prediction for each class. The Matthews correlation coefficient (MCC) measure is preferred as an overall performance measure since this measure only gives a high score if the prediction correctly classified a high percentage of the negative data instances *and* a high percentage of the positive data instances, with any class (im)balance (Chicco & Jurman, 2020). These metrics are derived from a confusion metrics:

Table 4.1: Confusion Matrix (no) funding classification

		Prediction	
Actual		No funding	Funding
	No funding	TN	FP
	Funding	FN	TP

The precision measure for the funding class in equation 5 is important since it is equal to the proportion of correctly predicted start-ups with a predicted ‘funding’ classification.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

The recall measure for the funding class (also known as specificity) in equation 6 is equal to the proportion of correctly predicted start-ups with an actual ‘funding’ classification.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Since we assume that VC investors have limited resources, the recall measure is not the most important measure. Due to the neccessity to invest in a few start-ups due to capital constraints, it is less interesting to know a lot of firms with a reasonably high success/funding probabilty than knowing a few firms with a high succes/funding probability. Therefore, as Arroyo et al. (2019) argue the recall should be high enough to find enough interesting firms, but the precision metric should be the main focus. The F1 measure is used as a balanced metric between precision and recall for each class:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

As earlier mentioned the MCC measure is used to determine the best model. The MCC measure takes into account two problems of the F1 measure; the F1 score is asymmetric for class swapping and does not take into account the number of true negatives (TN) (Chicco & Jurman, 2020). In contrast with the F1 measure that varies between 0 (perfect misclassification) and 1 (perfect classification), the MCC varies between -1 (perfect missclassifation) and 1 (perfect classification).

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (8)$$

4.2.1. Penalized Logistic Regression

Before applying the logistic regression minmax normalization took place to scale features. Main reasons to apply normalization is to ensure fair penalization across all features and to speed up the model's time to converge. To adjust for overfitting a penalized logistic regression is used for predicting using the same variables as the full model in equation 1 in section 4.1. Retaining a subset of the predictors and discarding other predictors, produces a more interpretable model and prevents overfitting (Hastie, Tibshirani & Friedman, 2009). Żbikowski and Antosiuk (2021) use a Lasso regression (L1 regularization [$\sum_j^p |\beta_j|$], $\alpha = 1$) and a Ridge regression (L2 regularization [$\sum_j^p \beta_j^2$], $\alpha = 0$) as shrinkage method (see equation 8). The penalty in the Lasso regression shrinks regression coefficients towards zero and the penalty of the Ridge regression shrinks coefficients close to zero. In this study I use an Elastic-Net shrinkage method to produce a logistic regression with L1 and L2 regularization which shrinks some coefficients to zero and other close to zero.

$$\sum_i^n (y_i - \hat{y}_i)^2 + \frac{1}{C}(\alpha \sum_j^p |\beta_j| + (1 - \alpha) \sum_j^p \beta_j^2) \quad (9)$$

The *LogisticRegressionCV* classifier implementation in the *scikit-learn* (Pedregosa et al., 2011) package is used to tune the regularization strength (C, lower C is higher penalization and vice versa) and the elastic-net mixing parameter (α). All combinations of the list hyperparameters displayed in Table 4.2 are tested (also called exhaustive grid search) using 5-fold cross validation and considering the highest F1 score (MCC is not available in the scikit-learn package as *scoring* metric).

Table 4.2: Hyperparameters tested in random forest using randomized search cross-validation.

Hyperparameter	List
C	0.001, 0.01, 0.1, 1, 10, 100, 1000
α	0, 0.2, 0.4, 0.6, 0.8,1

4.2.2. Random Forest

A random forest is used as a tree-based method for classification. This method is an improvement of bagging, which uses bootstrapped datasets from the original dataset to fit separate decision trees and then aggregates the individual predictions (by voting) to form a final prediction. The random forest method uses a random number of variables as predictors of each individual decision tree, which results in a decorrelation between all trees on each bootstrapped dataset in the random forest and as a result in a variance reduction of the predicted values (compared to taking the average of correlated decision trees) (Kuhn & Johnson, 2013).

In a random forest different parameters can be tuned. In this paper I decided to tune the number of trees in the forests (*number of estimators*), the maximum depth of the decision tree (*maximum depth*), the maximum number of features to be considered (*maximum features*), the minimum number of samples required to split an internal node (*minimum samples leaf*), and the minimum number of samples required to split an external node (*minimum samples split*).

Due to computational restraints a randomized search parameter optimization is used to find optimal hyperparameters for the random forest. In contrast to the exhaustive grid search method. which tests every combination of parameters, the randomized search cross-validates different number (*n_iter*) of random parameters combinations from a predefined distribution of parameters (see Table 4.3).

Table 4.3: Hyperparameters tested in random forest using randomized search (*n_iter=50)* with cross-validation.

Hyperparameter	List
Number of estimators	100, 250,500,750,1000,1250
Maximum depth	None, 5, 10, 15, 20,25,30
Maximum features	6
Minimum samples leaf	1,2,4
Minimum samples split	2,5,10

After knowing the optimal random parameters combinations, exhaustive grid search is performed with 5-fold cross-validation using the F1 *scoring* metric near the optimal random parameter's combination according to the randomized search (see Table 4.4) to check for further improvements of the random forest. The *RandomizedSearchCV* and

GridSearchCV implementations are used to tune the hyperparameters of the *RandomForestClassifier* model, which are all part of Python's *scikit-learn* (Pedregosa et al., 2011) package.

Table 4.4: Hyperparameters tested in random forest using grid search cross-validation.

Hyperparameter	List
Number of estimators	650, 650, 850
Maximum depth	25
Maximum features	4,5,6
Minimum samples leaf	1,2,3
Minimum samples split	4,5,6

4.2.3. Support Vector Classifier

In addition, the support vector machine (SVM) approach is used since this approach is intended for the binary classification of two classes by using a boundary. The support vector classifier is an extension of the maximal marginal classifier, which assumes perfect separability by a linear boundary while the support vector classifier does not. While the support vector classifier only assumes a linear boundary (independent of the degree of perfect separability of instances), SVM admits for a non-linear boundary by enlarging the feature space using kernels (Hastie, Tibshirani & Friedman, 2009).

Before applying the SVM, standardization takes place to scale all features (center to mean and scale to unit variance) because SVM is sensitive to the scale of the features (Juszczak, Tax & Duin, 2002) and using scaled features will speed up the process of fitting the SVM. Different parameters can be tuned to improve the prediction power of the support vector machine. The SVM approach can be used as a non-linear classifier. In this study two hyperparameters are tuned in a similar fashion as done by Żbikowski & Antosiuk (2021); the nonnegative regularization parameter (C) and the kernel coefficient (γ). The tuning of the regularization parameter determines the tolerated training errors that occur by a specific boundary/margin. A higher C will lead to more toleration of errors to the margin and thus less regulation. A lower C will lead to less toleration of errors to the margin and thus higher regulation. The C parameter controls the bias-variance trade-off; higher C will result more bias and less variance, lower C will result in less bias and more variance (Hastie, Tibshirani & Friedman, 2009). As suggested by Pedregosa et al. (2011), a logarithmic grid of C between 10^{-3} and 10^{3} are tested within a 5-fold cross-validation process. In line with the study of Żbikowski & Antosiuk (2021), a radial basis function kernel is used as kernel. The γ parameter is tuned since this parameter defines the influence of a single training example; a low γ parameter suggests far influence of support vectors while a high γ parameter suggests a narrow radius of influence (Pedregosa et al., 2011). The two γ values suggested by the *SVC* implementation in the *scikit-learn* package (Pedregosa et al., 2011) are used as tuning numbers (see Table 4.5).

In line with the elastic-net logit, hyperparameters in the *SVC* function are 5-fold cross-validated using the *GridSearchCV* function. The F1-score metric is used as *scoring* function in the cross-validation process.

Table 4.5: Hyperparameters tested in SVM using grid search cross-validation.

Hyperparameter	List
C	0.001, 0.01, 0.1, 1, 10, 100, 1000
Kernel type	Radial basis function kernel
γ	$\frac{1}{n_{features} * X.var()}$ (also known as 'scale'), $\frac{1}{n_{features}}$ (also known as 'auto')

5. Data Analysis and Results

5.1. Prescriptive Analysis

The main results of the prescriptive analysis are displayed in Table 5.1. Each column in Table 5.1 represents a logistic regression using *dummy_funding_True* as response variable. Columns 1 until 7 contain a subset of the main predictor variables used in the full model, which is displayed in column 8. The control variables *num_founders* and *gender_male* influences the VC likelihood, as expected, positively. The variable *startup_age* is not significant. All non-categorical predictor variables are centred for the ease of interpretation.

In Table 5.2 an overview of all hypotheses is displayed. Since the hypotheses in this paper are focused on the existence of a certain effect, I will report the significance and the sign of the logit coefficients to answer the hypotheses. I will only report the size of variables on the odds ratio, since interpreting marginal effects of predictors on the probabilities of receiving VC funds depends on the change of the predictor, the starting value of the predictor, and the value of the other predictors (Hoetker, 2007; Zelner, 2009).

The coefficients of the educational related effects suggest a general positive effect of education on the likelihood of receiving venture capital. Considering model 1, a significant positive effect of having a degree is observed on the log odds. Having a degree, results in 47.0% higher odds ratio of receiving VC funding according to the full model in column 8 ($= \exp(0.385) - 1$). This effect is without considering the duration of all degrees which is captured in the variable *duration_all_degrees*. The variable *duration_all_degrees* does have a significant and positive effect on the likelihood of receiving VC funding. If the size of the coefficient of the full model in Table 5.1 in column 8 is considered, one month longer education increases the odds of getting VC funding by 0.8% ($= \exp(0.008) - 1$). This is in line with the research of Van de Ven et al. (1984), Sapienza and Grimm (1997) and Barringer et al. (2005) who conclude that education is an important driver of entrepreneurial success. Therefore, there is support for hypothesis 1 since a positive object is observed of having one or more degree(s) on the likelihood of acquiring venture capital funding. The total duration of education has a positive effect on the likelihood of receiving VC funding and consequently there is support for hypothesis 2.

Model 1, model 7 (full model without interaction), and model 8 (full model) show a robust significant positive effect of the variable *IT*, which indicates that having an IT-related degree increases the effect of having a degree. This is in line with the conclusion of Homburg et al. (2014) who discovered a positive effect of related education of CMO's on the VC process. Having an IT-related degree increases the effect of having degree with 19.0% ($= \exp(0.174) - 1$) on the odds ratio. Therefore, there is support for hypothesis 3.

Table 5.1: Results logit: effect of founder characteristics on the likelihood of receiving venture capital funding (Note: *p<0.1, **p<0.05, ***P<0.01).

Dependent variable: dummy_funding_True	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	-2.174***	-1.712***	-1.762***	-2.813***	-1.720***	-1.186***	-2.503***	-2.510***
	(0.105)	(0.097)	(0.104)	(0.118)	(0.097)	(0.130)	(0.165)	(0.165)
founder_has_degree_True	0.574***						0.368***	0.385***
	(0.083)						(0.092)	(0.092)
IT_True	0.281***						0.174**	0.174**
	(0.084)						(0.088)	(0.088)
phd_True	0.263						0.042	0.051
	(0.161)						(0.169)	(0.169)
mba_True	0.217*						0.135	0.125
	(0.116)						(0.121)	(0.121)
ms_True	0.237**						0.100	0.104
	(0.111)						(0.116)	(0.116)
bs_True	0.044						0.010	0.012
	(0.091)						(0.094)	(0.094)
duration_all_degrees	0.009***						0.009***	0.008***
	(0.001)						(0.001)	(0.001)
industry_related_job_True		0.142**					-0.023	-0.055
		(0.070)					(0.074)	(0.079)
industry_related_startup_True		0.209*					0.311**	0.294*
		(0.121)					(0.138)	(0.153)
total_work_experience		0.041***					0.002	-0.001
		(0.003)					(0.004)	(0.005)
total_work_experience:industry_related_job								0.010
								(0.009)
total_work_experience:industry_related_startup								0.004
								(0.015)
num_founded_firms			0.077**				-0.125**	-0.128**
			(0.036)				(0.050)	(0.050)
founder_has_linkedin_True				1.193***			0.956***	0.966***
				(0.075)			(0.079)	(0.079)
founder_has_facebook_True				-0.360***			-0.340***	-0.341***
				(0.075)			(0.077)	(0.077)
founder_has_twitter_True				0.246***			0.137**	0.141**
				(0.062)			(0.066)	(0.066)
founder_has_website_True				-0.433***			-0.455***	-0.449***
				(0.076)			(0.079)	(0.079)
silicon_valley_True				0.696***			0.653***	0.651***
				(0.055)			(0.057)	(0.057)
number_press_references				0.024***			0.023***	0.023***
				(0.005)			(0.005)	(0.005)
opportunity_entrepreneur					0.610***		0.155*	0.045
					(0.076)		(0.086)	(0.106)
last_job_funding					-0.000		-0.000	-0.000
					(0.000)		(0.000)	(0.000)
opportunity_entrepreneur:last_job_funding								-0.000
								(0.000)
opportunity_entrepreneur:duration_all_degrees								0.007**
								(0.003)
opportunity_entrepreneur:total_work_experience								0.009
								(0.011)
profile_topic_0						0.356***	0.303**	0.308**
						(0.130)	(0.139)	(0.140)
profile_topic_1						-0.517***	-0.535***	-0.538***
						(0.135)	(0.145)	(0.145)
profile_topic_3						-0.708***	-0.641***	-0.650***
						(0.111)	(0.118)	(0.118)
profile_topic_4						-0.107	-0.117	-0.122
						(0.145)	(0.154)	(0.154)
profile_topic_5						-0.594***	-0.238**	-0.241**
						(0.096)	(0.103)	(0.103)
num_founders	0.465***	0.471***	0.463***	0.463***	0.461***	0.461***	0.465***	0.467***
	(0.021)	(0.020)	(0.020)	(0.021)	(0.020)	(0.020)	(0.022)	(0.022)
gender_male	0.535***	0.386***	0.397***	0.407***	0.378***	0.395***	0.501***	0.500***
	(0.103)	(0.100)	(0.100)	(0.103)	(0.100)	(0.100)	(0.106)	(0.107)
startup_age	0.002	0.001	0.001	0.002	0.001	0.001	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	9,502	9,502	9,502	9,502	9,502	9,502	9,502	9,502
McFadden's pseudo R-squared	0.101	0.057	0.061	0.120	0.063	0.070	0.151	0.152
Residual Std. Error	1.000 (df=9491)	1.000 (df=9495)	1.000 (df=9497)	1.000 (df=9492)	1.000 (df=9496)	1.000 (df=9493)	1.000 (df=9474)	1.000 (df=9469)
F Statistic	(df=10; 9491)	(df=6; 9495)	(df=4; 9497)	(df=9; 9492)	(df=5; 9496)	(df=8; 9493)	(df=27; 9474)	(df=32; 9469)

At first glance (model 1) the quality of the last education seems to have a positive effect on the likelihood of receiving VC funding; an MBA as last degree increases the odds ratio with 24.2% ($= \exp(0.217) - 1$), and a master as last degree increases the odds ratio with 26.7% ($= \exp(0.237) - 1$). However, in the final model (column 8, Table 5.1) these effects are not significant, and therefore there is no support for hypothesis 4.

While a positive effect of industry related experience from a last job or start-up exists in the model in column 2 of Table 5.1, the industry experience observed in a last job (*industry_related_job_True*) has no significant effect on the odds ratio of the dependent variable in the models displayed in column 7 and 8 of Table 5.1. However, industry experience obtained through a previous start-up increases the odds ratio with 34.2% (=$\exp(0.294) - 1$) considering a p-value of 0.1. Since the effect of industry experience obtained through a last job is not considered to be significant in the final model, there is no support for hypothesis 5. The interaction effect between the length of the working experience and the presence of an industry related last job/start-up is not significant. There is no support for hypothesis 6. The length of the working experience has no significant effect on the likelihood of receiving VC funding.

Considering the final model (column 8 of Table 5.1), I do find that industry-related topic 0 has significant positive effect on odds ratio of receiving VC funding compared to the base topic 2. Therefore, hypothesis 7 is supported. Except the positive effect of a dominant industry-related topic, no effect seems to exist for industry related experience acquired through employment. This is in contrast with earlier research that did find positive effects of industry experience on entrepreneurial success (Roure & Maidique, 1986; Barringer, Jones & Neubaum, 2005, Homburg et al., 2014).

The number of founded firms, which tends to capture entrepreneurial experience, has a negative significant effect on the likelihood of receiving VC funding. For one additional founded firm, I expect a decrease in the odds ratio of 13.66% ($= \exp(-0.128) - 1$). This suggests no support for hypothesis 8. This is in line with the results of Van de Ven et al. (1984) who explain the negative effect by the strongly negative correlation between education and small business experience. If we consider the presence of dominant managerial/entrepreneurial topics in the founder's Crunchbase description (topics 3 and 4), a significant negative effect of the dominant presence of topic 3 and no significant effect of topic 4 is observed. Therefore, no support for hypothesis 9 is found. The extent of entrepreneurial-related experience does not influence the VC funding decision positively. This is in line with earlier research of Van de Ven, et al. (1984), but in contrast with Zhang (2007) and Baptista et al. (2014), who discovered a positive role for entrepreneurial experience on business success.

The presence of links to social media links seems to have a huge impact on the VC decision. Founders' profiles containing a link to LinkedIn, Facebook, Twitter or a website result in a +162.7% ($= \exp(0.966) - 1$), -28.9% ($= \exp(-0.341) - 1$), +15.1% ($= \exp(0.141) - 1$), - 36.1% ($= \exp(-0.448) - 1$) change in the odds-ratio. Since not all links to social platforms positively influence the likelihood of acquiring venture capital, no support is found for hypothesis 10. The number of press references has a significant positive effect on the odds-ratio; keeping all other variables constant, a new number of press reference increases the odds-ratio with 2.3% ($= \exp(0.023) - 1$). Consequently, there is support for hypothesis 11. A headquarter located in the San Francisco Bay Area in California significantly positively influences the odds ratio of receiving VC funding strongly by 91.7% ($= \exp(0.651) - 1$), which suggests support for hypothesis 12. Confirming earlier literature (Shane & Cable, 2002; Banerji & Reimer, 2019; Mollick, 2014), overall social capital seems to have a significant positive effect on the likelihood of receiving VC.

No stand-alone positive signaling effect is discovered of the presence of employment shortly before founding a start-up (also known as opportunity entrepreneur; *opportunity_entrepreneur*). Cassar (2006) discovered that being an opportunity entrepreneur, characterized by recent employment, leads to higher intentions. Considering the model in column 8 (Table 5.1), being an opportunity entrepreneur (and thus having higher intentions/motivation) does not significantly influence the odds-ratio of receiving VC funding. There is no support for hypothesis 13.

A bigger effect of being an opportunity entrepreneur is expected if somebody quits a promising firm, which I defined by a firm that received a lot of funding. It is expected that the internal motivations of a founders are bigger when founders quit employment with good career perspective when diving into the start-up adventure than employment with bad career perspective (higher opportunity costs). If we consider the final model (column 8, Table 5.1) the interaction effect between being an opportunity entrepreneur and the funding amount of the last employer is insignificant. The results do not imply that the effect of being an opportunity entrepreneur is higher when the total funding amount is larger than when it is not. This suggests no support for hypothesis 14.

Beforehand, I expected that the effect of education and industry experience on the likelihood of VC receiving VC funding is higher when past employment is present than when it is not. Main reason is the depreciation of specific human capital due to unemployment (Baptista, Karaöz & Mendonça, 2014) and the founders' attached importance to specific human capital to compensate the higher risk of diving in the start-up adventure due to opportunity costs. Considering the interaction effect between being an opportunity entrepreneur and the duration of all degrees, a positive significant effect is observed in the final model (column 8, Table 5.1). It seems the length of education for opportunity entrepreneurs plays a bigger role in the VC process than the length of education for non-opportunity entrepreneurs. The effect of the length of education is 0.7% $(= \exp(0.007) - 1)$ bigger on the odds-ratio of VC funding for opportunity-entrepreneurs than for non-opportunity-entrepreneurs (Buis, 2010). The interaction effect between being an opportunity entrepreneur and the duration of the work experience is not significant in the final model. Hence, there is partly support for hypothesis 15 since only education plays a bigger role in the VC funding process for opportunity entrepreneurs.

Table 5.2: Overview of hypotheses (dependent variable: likelihood of receiving VC funding)

Hyp.	Founder characteristic	Independent Variable	Proposed Relationship	Result
1	Education	Having one or more degrees	+	Supported
2	Education	Duration of degrees	+	Supported
3	Education	Having industry related last degree	+	Supported
4	Education	Level of last degree	+	Not supported
5	Industry experience	Having industry related last job	+	Not supported
6	Industry experience	Industry related last job * total work experience	+	Not supported
7	Industry experience	Extent of industry experience in profile description	+	Supported
8	Entrepreneurial experience	Number of founded organizations	+	Not supported
9	Entrepreneurial experience	Extent of entrepreneurial experience in profile description	+	Not supported
10	Social capital	Profile has links to social media	+	Not supported
11	Social capital	Number of press references	+	Supported
12	Social capital	Presence HQ start-up hub	+	Supported

13	Opportunity costs	Employment before founding	+	Not Supported
14	Opportunity costs	Employment before founding * funding amount previous employer	+	Not Supported
15	Opportunity costs	Employment before founding* specific human capital	+	Partly Supported

5.2. Predictive Analysis

The performance metrics of the final models' predictions on the test set are shown in Table 5.3. While the non-penalized logistic regression (basic model) is not tuned, the other models are tuned using 5-fold cross-validation. The final tuned elastic net regression uses a value of 0.4 for the α parameter and a value of 100 for the C parameter (relatively low penalization). The final Random Forest Classification model used relatively deep and many trees (25, 500) and has relatively low requirements for splitting an internal node and external nodes (2). The support vector machine algorithm scored the highest score, using a C of 1 and a gamma equal to $\frac{1}{n_{features} * X.var()}$ (also known as 'scale').

Table 5.3 Results prediction tests set.

Model	Class	Precision	Recall	F1	MCC
Logit	No VC	0.82	0.95	0.88	
	VC	0.63	0.30	0.40	0.334
Penalized logit	No VC	0.82	0.95	0.88	
	VC	0.62	0.29	0.40	0.326
Random forest	No VC	0.82	0.97	0.88	
	VC	0.69	0.25	0.37	0.353
SVM	No VC	0.80	0.98	0.88	
	VC	0.74	0.19	0.30	0.301

Looking at the precision, recall and F1-score of all models in general, the first thing that stands out is that all models perform well in the prediction of start-ups that do not receive VC capital. The proportion of correctly predicted start-ups with a true no funding classification is for all models higher than 0.95 and for the SVM it is even 0.98 which is close to predicting all start-ups with no funding in practice correctly. The precision score for all models is 0.80 or higher for the 'No VC' class, indicating that the proportion of correctly predicted start-ups with a 'non-funding' prediction is 80%. Therefore, 20% or less of all predicted non-funding start-ups in fact is classified wrongly.

The success rate of VC investment decisions is the most important metric (Arroyo et al., 2019). Therefore, the precision metric of the 'funding' class in Table 5.3 is the most interesting for evaluating the model. The penalized logit has a slightly lower precision score than the non-penalized logit. The random forest shows an improvement in the precision score (0.69) compared to the penalized logit (0.62). The tuned SVM model has the highest precision score of 74%. This precision score is higher than the precision scores from the models performed by Arroyo et al. (2019), which vary between 45% and 68% for the 'positive' class, and Żbikowski and Antosiuk, which vary between 49% and 67%. The use of the extra created variables seems to have improved the prediction of the 'VC' class.

The recall measure of the SVM model in Table 5.3 reveals that in contrast to the high precision score, the proportion of correctly classified start-ups which received VC funding is very low (0.19). Therefore, using the SVM model will be the best model for smaller firms that only find important the success rate on relatively few investments (such as medium VC firms). However, in general a random forest can be used since this model is the 'best' model because the MCC measures of this model outperforms the other models.

Venture capital firms can use this model to support investment decisions by assessing the chances of founder's getting VC funding. Since the trained model provides a classification whether a founder will receive VC funding, the model can be used as a decision-support system for all venture capitalists. In practice, venture capitalists should integrate the trained model into dashboards that indicate if a founder is predicted to receive funding. Venture capitalists should check their funding decisions by consulting the dashboard, which could improve the performance of the VC firm. Besides integrating dashboards, this model can be integrated to make the funding decision automatically to speed up the VC funding process. For example, if large VC firms receives a lot of VC funding requests, these VC firms can save a lot of time by letting the model automatically deny or accept VC funding requests. As a result, the VC firm can proceed the VC process with founders that are predicted to be funded. Implementing this model in dashboards or directly in the VC funding process brings venture capitalists closer to the goal of automatically spotting for example the next Mark Zuckerberg.

The following scenario shows how VC firms can use this model in practice (see Figure 5.1). At the beginning of 2020 100 random start-ups, all founded between January 2017 and December 2020, apply for VC funding at the beginning of 2021. All start-ups are linked to 180 different founders in Crunchbase, which makes it possible to automatically extracts founder's characteristics of all 180 founders. A venture capitalist uses the random forest model to automatically spot entrepreneurs who are likely to receive VC funding and are therefore the most interesting to fund. The suggested model selects 15 different founders from all founders. This will speed-up process of the VC funding process since 165 founders' applications are automatically denied. According to the precision metric, around 69% of these selected 15 founders are likely to receive venture capital funding and are therefore interesting to fund.

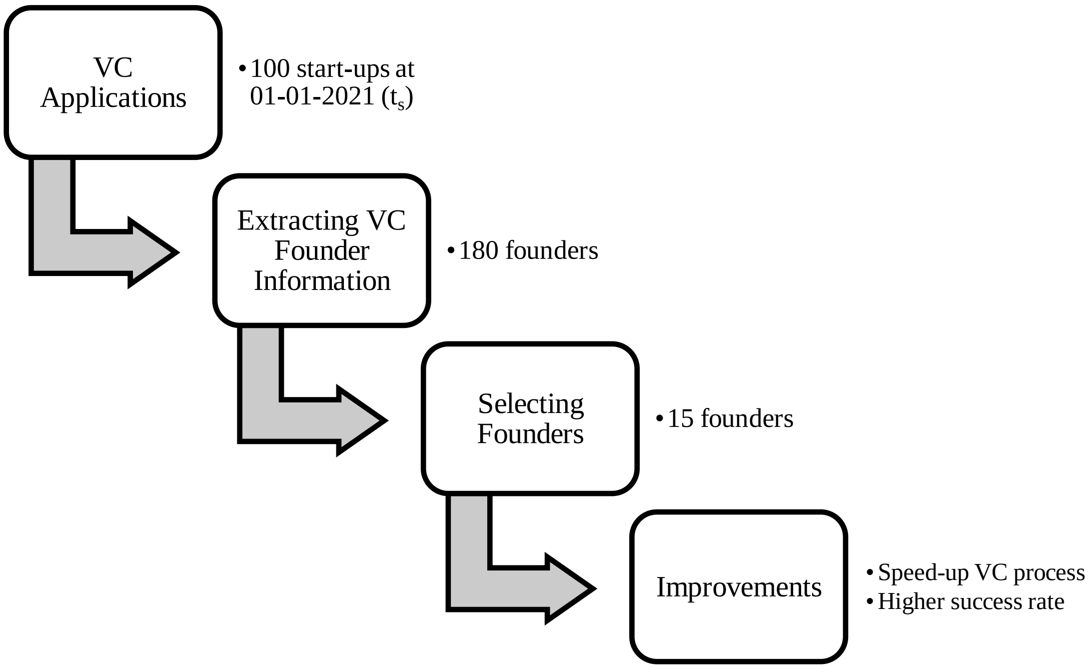

Figure 5.1: Schematically overview of VC funding process in the scenario of 100 random IT US-based start-ups (founded between 2017 and 2020) applying at the beginning of 2021.

6. Conclusion

6.1. General Discussion

Despite earlier research regarding the role of founder characteristics in business success, little is known regarding the role of founder's characteristics in the venture capital process. To shed a light on this role a prescriptive analysis is performed. Additionally, I use founder characteristics to predict the occurrence of VC funding. This prescriptive and predictive analysis answer the following research question:

What is the role of founder's characteristics in acquiring venture capital fundraising in the United States and how can these characteristics be used to predict venture capital fundraising?

The effects of founder's characteristics on VC fundraising are determined using logistic regressions. Summarising the results, this study concludes that founder's education positively influences the VC funding decision. A less clear effect of industry-related experience follow from the results; while no significant positive effects from previous industry-related employment and start-up experience are discovered, a dominant industry related topic in the founder's Crunchbase profile increases the likelihood of raising venture capital. The effects of variables that capture social capital show mixed effects. Being presents on some social platforms showed positive effects and having an account on other platforms showed negative effects on the VC funding likelihood. Having a headquarter located in a start-up hub and the number of press references boost the VC process. Being an opportunity entrepreneur does not affect the VC funding decision in general. However, the role of education in the VC funding process becomes bigger when a founder is classified as an opportunity entrepreneur.

In this study, receiving VC funding is predicted using different supervised machine learning methods; elastic-net logistic regression, random forest, and support vector machine. The tuning of hyperparameters takes place using cross-validation. Evaluating the predictions of the test set shows us that the random forest method brings the best performing model considering the MCC measure. However, the support vector machine is the best model if only the precision metric of the positive class is considered.

6.2. Academic Contribution

Until now, no study is published that uses and collects thousands of instances of founders and start-ups in a fully automated way to investigate the role of founder's characteristics on the likelihood of receiving VC funding (prescriptive analysis). In addition to the variables that are used in similar studies regarding venture success, feature engineering takes place to extract new features from the Crunchbase database using text data (of founder's profiles and degree descriptions) that capture the extent of founder's human capital. Furthermore. this research contributes to the literature regarding the role of entrepreneur's human capital by considering a proxy for opportunity costs. By using Crunchbase's API, a large dataset is constructed which diminishes the presence of a response and selection bias present in survey data. The underlying code structure of this paper can be used to open doors for future research by easily creating datasets of thousands of founders in different regions or industries.

The predictive analysis in this paper uses more features than earlier research (Żbikowski and Antosiuk, 2021; Arroyo et al., 2019) that capture multiple types of founder's characteristics. At the same time I keep into account the look-ahead bias as pointed out by Żbikowski and Antosiuk (2021) that can exists in the studies of Arroyo et al. (2019) and Xiang et al. (2012). As suggested by Żbikowski and Antosiuk (2021), founder's profiles are explored using topic mining and dominant topics are used for prediction. The predictions on the test set show that the models created in this research perform better than the models created by Żbikowski and Antosiuk (2021) and Arroyo et al. (2019).

6.3. Managerial Implications

In the introduction I considered if people should start an IT-related start-up or should study first to maximize the likelihood of getting venture capital funding. In this paper positive effects of having one or more degrees and the duration of all degrees are discovered. Therefore, I recommend people who want to become a successful start-up founder to start studying for a relatively long time since this increases the likelihood of getting funded by venture capitalists. It is recommended to study industry-related subjects since this paper discovered a positive effect on VC funding likelihood if founder's last degree is industry-related.

Getting industry-related experience through a job or start-up does not improve the chance of getting VC funding. However, the extent of industry-related experience in the profile description, does positively influence the likelihood of receiving VC funding. Therefore, aspiring founders are recommended to have a profile description that displays the extent of industry experience. In addition, the curriculum of business studies should not be too focussed on getting industry experience through employment or the founding of a start-up (for example through internship programs), since this does not directly improve the odds of getting VC funding.

In this paper positive effects of the number of press references and the presence in Silicon Valley on the likelihood of receiving VC funding are discovered. Therefore, my recommendation for IT-related start-ups is to get a lot of press references and move their headquarters to Silicon Valley.

No positive signalling effect of having opportunity costs is discovered. Nevertheless, the results suggests that education related variables play a bigger role in the VC funding process for opportunity entrepreneurs than for non-opportunity entrepreneurs. Aspiring founders should be aware that quitting a promising job on its own does not increase the VC funding likelihood but increases the role of education in the VC funding process.

The models constructed in the predictive analyses can be used to support the investment decision of venture capitalists. Since the best model predicts 69% of the as funded classified founders correctly, venture capitalists are recommended to integrate this model into their investment decision (for example through a dashboard environment). Using this model as a decision support system will increase the success rate of VC funding because the VC funding process can be seen as a funnel with successful start-ups as a final product.

Besides improving the success rate, venture capitalists could speed up the VC funding process by using the models to make funding decisions automatically. The scenario elaborated in this paper showed that venture capital funding applications from 180 founders can be filtered down to 15 founders who are likely to be funded. Besides speeding up the process, using the automatic decision system will decrease the overhead costs of the VC funding in general because less applications must be revised.

6.4. Limitations and Directions for Future Research

This research has some limitations. Firstly, 23% of all founders used in the dataset received VC funding. Since 11 of 12 entrepreneurs fail with their start-up, founders with VC funding seem to be overrepresented, which can result in a selection bias. The significant difference in the mean funding ratio of the final dataset and the mean funding ratio of the start-ups that are not included in the final dataset (due to a missing link to the founders) also suggests this selection bias. Future research could undersample founders that receive VC funding or fill in the missing link to founders by consulting LinkedIn to prevent a selection bias.

Secondly, in the feature creation process, I assume that missing data equals 0. Contributors are one of the resources that create Crunchbase founder's profiles (Crunchbase, n.d.-b). Contributors could prefer completing profiles of more successful entrepreneurs who are more likely to receive VC funding. This could result, for example, in more successful founder instances with registered education than not successful founders. Eventually, this leads to an upward biased effect of the role of founder characteristics.

Only considering features that apply on the period before the funding decision, does not fully eliminate the look-ahead bias. Using profile descriptions of founders and links to social media platforms, that could be updated later, increases the risk of a look-a-head bias. Future research could eliminate the look-ahead bias completely by only considering data that is updated before the funding decision date.

Fourthly, the data consists of start-ups operating in the IT industry and have a headquarter in the United States in a specific time frame. This selection criteria prevents the influence of variables that are correlated with other features and influence the VC process (for example country's fiscal policy regarding VC). The main disadvantage of using this method is that a selection bias arises and that the results are not representative for different regions and industries. Future research should extend the range of their research to multiple countries and industries.

Future studies could improve the quality of the data by adding different sources to the Crunchbase data. For example, LinkedIn data can be used to fill in missing Crunchbase data. This study predicts the role of founder's characteristics on receiving VC funding in general, but the role of specific characteristics may change in different funding stages. Future work could investigate this by using a dummy indicating if a specific funding stage (such as Serie A, B, etc.) is received as response variable. This can lead to a more differentiated models and thus better predictions of VC funding in each funding rounds.

7. Appendix

Table 7.1

Dependent variable name	Description
Dummy_funding_True	Dummy variable that equals 1 if funding round is received which is determined as an investment stage in the early-stage (VC up to Series B) or late-stage (VC from Series C and onwards).
Independent variables	**Description**
Education	
Founder_has_degree_True	Dummy variable that equals 1 if founder has completed a degree before the simulation date.
IT_True	Dummy variable that equals 1 if the subject of the last degree contains the word “computer” or “information”
Phd_True	Dummy variable that equals 1 if the type of the last degree was equal to ‘phd’
Mba_True	Dummy variable that equals 1 if the type of the last degree was equal to ‘mba’
Ms_True	Dummy variable that equals 1 if the type of the last degree was equal to the following regular expression: 'msc?|ma|masters?
Bs_True	Dummy variable that equals 1 if the type of the last degree was equal to the following regular expression: 'ba|bsc?|bachelors?'
Duration_all_degrees	The length of all education in months by taking (1) the sum of all differences between the start and completion date, (2) or if not available the difference between the start date of the first degree and completion date of the last degree, or if not available (3) the

	differences between the founding date of the start-up and the start date of the first degree.
Industry Experience	
Industry_related_job_True	Dummy variable that equals 1 if the last employer 'information technology' registered as a category group in Crunchbase.
Industry_related_startup_True	Dummy variable that equals 1 if previously founded start-up had 'information technology' recorded as a category group in Crunchbase.
Total_work_experience	The difference in years between (1) the completion date of the last degree and the founding date of the startup, or if not available the difference in years between the founding date and the start date of the first job.
Entrepreneurial/Managerial Experience	
Num_founded_organizations	Number of founded organisations before the beginning of the simulation date
Social Capital	
Founder_has_facefook_True	Dummy variable equals 1 if a link to founder's Facebook account is present
Founder_has_linkedin_True	Dummy variable equals 1 if a link to founder's LinkedIn account is present
Founder_has_twitter_True	Dummy variable equals 1 if a link to founder's twitter account is present
Founder_has_website_True	Dummy variable equals 1 if a link to founder's website is present
Silicon_valley_True	Dummy variable equals 1 if the start-up's headquarter location group is equal to 'san-francisco-bay-area-california'

Number_press_references	Number of press references before the start of the simulation
Opportunity	
Opportunity_entrepreneur	Dummy variable equals 1 if founder's last job ended within 12 months before or after the founding date of the start-up
Last_job_funding	Total amount of funding before the start of the simulation window (x 1000,000 USD)

Control variables	**Description**
Gender_male	Dummy variable equals 1 if founder's gender is equal to 'male'
Num_founders	Number of founders of a start-up
Startup_age	Age in years of the start-up at the beginning of the simulation window.

8. Reference List

Akerlof, G. A. (1978). The market for “lemons”: Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235-251). Academic Press.

Albrecht, J. W., Edin, P. A., Sundström, M., & Vroman, S. B. (1999). Career interruptions and subsequent earnings: A reexamination using Swedish data. *Journal of human Resources*, 294-311.

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

Banerji, D., & Reimer, T. (2019). Startup founders and their LinkedIn connections: Are well- connected entrepreneurs more successful? *Computers in Human Behavior*, *90*, 46-52.

Barringer, B. R., Jones, F. F., & Neubaum, D. O. (2005). A quantitative content analysis of the characteristics of rapid-growth firms and their founders. *Journal of business venturing*, *20*(5), 663-687.

Baptista, R., Karaöz, M., & Mendonça, J. (2014). The impact of human capital on the early success of necessity versus opportunity-based entrepreneurs. *Small Business Economics*, *42*(4), 831-847.

Buis, M. L. (2010). Stata tip 87: Interpretation of interactions in nonlinear models. *The stata journal*, *10*(2), 305-308.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology, 92*(6), 1287-1335.

Brush, C., Greene, P., Balachandra, L., & Davis, A. (2018). The gender gap in venture capital-progress, problems, and perspectives. *Venture Capital*, *20*(2), 115-136.

Cassar, G. (2006). Entrepreneur opportunity costs and intended venture growth. *Journal of business venturing*, *21*(5), 610-632.

CB Insights (2018). Venture Capital Funnel Shows Odds of Becoming a Unicorn Are About 1%. *Research Briefs.*

CB Insights (2021). From Alibaba to Zynga: 45 of the Best VC Bets of All Time and What We Can Learn from Them. Consulted via: https://www.cbinsights.com/research/best-venture-capital-investments/

Chandy, R. K., & Tellis, G. J. (2000). The incumbent's curse? Incumbency, size, and radical product innovation. *Journal of marketing*, *64*(3), 1-17.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1-13.

Crunchbase (n.d.-a). What Industries are included in Crunchbase? Consulted via: https://support.crunchbase.com/hc/en-us/articles/360043146954-What-Industries-are-included-in-Crunchbase-

Crunchbase (n.d.-b). How does Crunchbase get my information? Consulted via: https://support.crunchbase.com/hc/en-us/articles/360001360088-How-does-Crunchbase-get-my-information-

Davenport, T., & Harris, J. (2017). *Competing on analytics: Updated, with a new introduction: The new science of winning*. Harvard Business Press.

Eisenhardt, K. M., & Schoonhoven, C. B. (1990). Organizational growth: Linking founding team, strategy, environment, and growth among US semiconductor ventures, 1978-1988. *Administrative science quarterly*, 504-529.

Gompers, P. A., & Lerner, J. (1999). *What drives venture capital fundraising?* (No. w6906). National bureau of economic research.

Gulati, R. (1995). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of management journal*, *38*(1), 85-112.

Hauser, J. R. (1978). Testing the accuracy, usefulness, and significance of probabilistic choice models: An information-theoretic approach. *Operations Research*, *26*(3), 406-421.

Hall, J., & Hofer, C. W. (1993). Venture capitalists' decision criteria in new venture evaluation. *Journal of business venturing*, *8*(1), 25-42.

Hasan, I., & Tucci, C. L. (2010). The innovation–economic growth nexus: Global evidence. *Research policy*, *39*(10), 1264-1276.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*.

Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal*, *28*(4), 331-343.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856-864).

Homburg, C., Hahn, A., Bornemann, T., & Sandner, P. (2014). The role of chief marketing officers for venture capital funding: Endowing new ventures with marketing legitimacy. *Journal of Marketing Research*, *51*(5), 625-644.

Hsu, D. H. (2007). Experienced entrepreneurial founders, organizational capital, and venture capital funding. *Research policy*, *36*(5), 722-741.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jeng, L. A., & Wells, P. C. (2000). The determinants of venture capital funding: evidence across countries. *Journal of corporate Finance*, *6*(3), 241-289.

Juszczak, P., Tax, D., & Duin, R. P. (2002). Feature scaling in support vector data description. In *Proc. asci* (pp. 95-102). Citeseer.

Ko, E. J., & McKelvie, A. (2018). Signaling for more money: The roles of founders' human capital and investor prominence in resource acquisition across different stages of firm development. *Journal of Business Venturing*, *33*(4), 438-454.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

MacMillan, I. C., Siegel, R., & Narasimha, P. S. (1985). Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business venturing*, *1*(1), 119-128.

MacMillan, I. C., Zemann, L., & Subbanarasimha, P. N. (1987). Criteria distinguishing successful from unsuccessful ventures in the venture screening process. *Journal of business venturing*, *2*(2), 123-137.

Matsuno, K., Mentzer, J. T., & Özsomer, A. (2002). The effects of entrepreneurial proclivity and market orientation on business performance. *Journal of marketing*, *66*(3), 18-32.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, *29*(1), 1-16.

National Venture Capital Association (2020). NVCA Yearbook 2020. Consulted via: https://nvca.org/wp-content/uploads/2020/04/NVCA-2020-Yearbook.pdf

Neuman, S., & Weiss, A. (1995). On the effects of schooling vintage on experience-earnings profiles: theory and evidence. *European economic review*, *39*(5), 943-955.

Nuscheler, D., Engelen, A., & Zahra, S. A. (2019). The role of top management teams in transforming technology-based new ventures' product introductions into growth. *Journal of Business Venturing*, *34*(1), 122-140.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Riquelme, H., & Rickards, T. (1992). Hybrid conjoint analysis: An estimation probe in new venture decisions. *Journal of Business Venturing*, *7*(6), 505-518.

Ries, E. (2011). Lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses. New York: Crown Business.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Roure, J. B., & Maidique, M. A. (1986). Linking prefunding factors and high-technology venture success: An exploratory study. *Journal of business venturing*, *1*(3), 295-306.

Sapienza, H. J., & Grimm, C. M. (1997). Founder characteristics, start-up process, and strategy/structure variables as predictors of shortline railroad performance. *Entrepreneurship Theory and Practice*, *22*(1), 5-24.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*

Shane, S., & Cable, D. (2002). Network ties, reputation, and the financing of new ventures. *Management science*, *48*(3), 364-381.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, *70*(1), 65-94.

Startup Genome (2020). Global Startup Ecosystem Report 2019. Consulted via: https://startupgenome.com/reports/global-startup-ecosystem-report-2019

Stigler, G. J. (1955). The nature and role of originality in scientific progress. *Economica*, *22*(88), 293-302.

Tufts, C. (n.d.) The Little Book of LDA. Consulted via: https://miningthedetails.com/LDA_Inference_Book/

Tyebjee, T. T., & Bruno, A. V. (1984). A model of venture capitalist investment activity. *Management science*, *30*(9), 1051-1066.

Van de Ven, A. H., Hudson, R., & Schroeder, D. M. (1984). Designing new business startups: Entrepreneurial, organizational, and ecological considerations. *Journal of management*, *10*(1), 87-108.

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012, May). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 6, No. 1).

Zhang, J. (2007). The advantage of experienced start-up founders in venture capital acquisition: evidence from serial entrepreneurs. *Small Business Economics*, *36*(2), 187-208.

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

Zelner, B. A. (2009). Using simulation to interpret results from logit, probit, and other nonlinear models. *Strategic Management Journal*, *30*(12), 1335-1348.