The Journal of Finance and Data Science 9 (2023) 100099

Contents lists available at ScienceDirect

The Journal of Finance and Data Science

journal homepage: www.keaipublishing.com/en/journals/jfds

Research article

Making it into a successful series a funding: An analysis of Crunchbase and LinkedIn data

Yiea-Funk Te [a,*], Michèle Wieland [a], Martin Frey [a], Asya Pyatigorskaya [b], Penny Schiffer [b], Helmut Grabner [a]

[a] ZHAW School of Engineering, Zurich, Switzerland
[b] Raized.ai AG, Zurich, Switzerland

ARTICLE INFO

Keywords:
Success prediction
Machine learning
Startups
Venture capital
Crunchbase
Gradient boosting

ABSTRACT

Startups are a key force driving economic development, and the success of these high-risk ventures can bring huge profits to venture capital firms. The ability to predict the success of startups is a major advantage for investors to outperform their competitors. In this study, we explore the potential of using publicly available LinkedIn profiles as an alternative and complementary data source to Crunchbase for predicting startup success. We provide a comprehensive review of the existing literature on the factors that influence startup success to create a large set of features for predictive modeling. We train two models for predicting startup success employing light gradient boosting that use LinkedIn data as a standalone and as a complementary data source, and compare them to baseline models based on Crunchbase data. We show that using LinkedIn as a complementary data source yields the best result with a mean area under the curve (AUC) value of 84%. We also provide a thorough analysis of what types of information contribute most to modeling startup success using the Shapley value method. Our models and analysis can be used to develop a decision support system to facilitate startup screening and the due diligence process for venture capital firms.

1. Introduction

Venture capitals (VCs) are professionals who manage a pool of capital and provide funding to private companies. They function as financial intermediaries by matching investors with financial resources looking for investment opportunities to entrepreneurs with promising ideas (Peneder, 2010).

VCs make an important contribution to economic growth. As new companies need a large amount of money to accelerate their growth, the debt financing offered by banks is unsuitable from a cash management perspective. VCs have emerged to fill this gap in startup financing and often remain the only possible source of capital for many new ventures (Davila et al., 2003).

However, the venture capital industry faces challenges which are mainly driven by two factors: First, the amount of capital invested into startups has grown by 300–400% over the last decade while the number of new startups has remained approximately the same (Retterath, 2020). As a result, a rising number of VC firms have to invest an increasingly large amount of capital in a limited number of

* Corresponding author.
E-mail addresses: yfunk.te@gmail.com, teyi@zhaw.ch (Y.-F. Te), wiei@zhaw.ch (M. Wieland), fret@zhaw.ch (M. Frey), asya@raized.ai (A. Pyatigorskaya), penny@raized.ai (P. Schiffer), grbn@zhaw.ch (H. Grabner).

Peer review under responsibility of KeAi Communications Co., Ltd

https://doi.org/10.1016/j.jfds.2023.100099
Received 1 September 2022; Received in revised form 18 May 2023; Accepted 31 May 2023
Available online 20 June 2023

assets, leading to growing competition among VCs. They aim to invest before competitors do and at earlier stages of startups: Series A investors typically invest in the seed stage now, while seed stage investors often invest in the pre-seed stage. Second, globalization of the traditionally more local VC business is facilitated by use of digital technology (Agmon and Sjögre, 2016). "Geography and 'warm intros' via exclusive networks will eventually become less relevant" (Retterath, 2020). Moreover, deal-flow is shifting from mainly passive sourcing models of the past to an increasingly active sourcing model as most investors will compete for very few high potential deals. In addition, the Covid-19 pandemic has boosted these developments as more and more investors are willing to consider investment targets outside their geographic area (Retterath, 2020). Another critical factor affecting VC funding is the impact of the Federal Reserve quantitative easing (QE) on capital flows in 2020 as a response to the economic downturn caused by the Covid-19 pandemic, which led to a flood of excessive capital into the VC industry, driving up the valuations of startups and increasing competition among investors for the best deals (Aryoubi et al., 2020). A major challenge for VCs is to find startups with a higher probability of success. Not only have studies shown that VCs often underperform compared to the S&P 500 index (Harris et al., 2014), but also that they are highly susceptible to human error and subjective judgment (Matusik et al., 2008). Moreover, the process of sourcing and screening in conventional investment methods is still very tedious and time-consuming due to the complexity of the strategy to evaluate the risks and rewards behind each investment (Schmidt, 2019).

Therefore, the challenge of how to better evaluate and predict the likelihood of startups' success is undoubtedly of great importance. In recent years, machine learning has experienced a significant upswing and found numerous applications for data-driven investment approaches. However, previous studies were mostly limited to the analysis of structured data sources, such as databases of the startup ecosystem consisting of investors, incubators, and startups (Ferrati and Muffatto, 2021; Fragkiskos et al., 2021; Lencioni, 2020; Zhang et al., 2017). In particular, numerous studies have been conducted to build predictive models based solely on data provided by Crunchbase (Dalle et al., 2017; Ferrati and Muffatto, 2020; Żbikowski and Antosiuk, 2021). While Crunchbase provides a large dataset with extensive information about the company, investor funding, information about the team (i.e., founders and employees) are largely absent, which has proven to be critical to a startup's success. Thus, predictive models based solely on Crunchbase data cannot capture the full mechanism underlying a company's success. In addition, Crunchbase data provides information that is often not available at the time of decision making, rendering models trained on Crunchbase data very difficult to apply in a real-world scenario (Żbikowski and Antosiuk, 2021). This limitation is also found in the most closely related study by Sharchilev et al. (2018). They built a gradient-boosted decision tree model to predict the likelihood of receiving Series A funding for companies using a dataset collected by Crunchbase, enriched with data from publicly available LinkedIn profiles of people working in the companies in addition to other web sources. However, several limitations can be identified. First, the Crunchbase dataset was enriched with data from Linkedin and other web sources collected a few years after the last sample from the originally labeled Crunchbase dataset. The impact of this approach on the performance of the model is not investigated and is difficult to predict. Second, only a few features are created from LinkedIn profile data (e.g., number of founders, statistics on the success of their previous companies, previous startup experiences, etc.), which means that the full data potential of LinkedIn was not exploited. Third, LinkedIn data is used as feature enrichment for Crunchbase data, thus the potential of LinkedIn data as a stand-alone data source for building a success prediction model is not evaluated. Finally, the significance of the model's features is only briefly addressed, and it is our opinion that further analysis is needed with respect to the model's explainability.

In this study, we explore the potential of using publicly available LinkedIn profile data to augment the missing information about the team for predicting startup success. Specifically, we enrich the data provided by Crunchbase with information about the founders collected from publicly available LinkedIn profiles and use it to train a robust machine learning model with light gradient boosting (LGBM) (Fan et al., 2019). In addition, we evaluate the potential of using LinkedIn profile data as a stand alone data source for predicting startup success. We then compare our models to a benchmark model trained solely on Crunchbase. Finally, we analyze which types of information contribute the most in modeling the success of startups by applying the Shapley value method (Sundararajan and Najmi, 2020).

2. Related work

2.1. Survey on startup success factors

A large body of literature exists on understanding the determinants of business success, which is a multifaceted and rapidly evolving process shaped by a variety of internal and external business factors and macroeconomic environments, including the global business cycle, industry trends, and government policies (Worthington and Britton, 2009). The literature review focuses on identifying factors that are both well documented in the literature and available in our datasets to be investigated.

Recently, Tykvová (2018) and Corea (2019) conducted extensive literature reviews to identify specific variables that could explain, to varying degrees, the likelihood of a firm's success. We briefly review the key determinants identified in previous studies. This provides valuable insights into what types of data should be used and serves as a foundation and inspiration for feature engineering to model business success. In general, the variables can be categorized into three macro groups (Corea, 2019): Company-related, person-related and investment-related attributes.

Company-related attributes include characteristics associated with business/operational aspects. The number of patents is positively related to the likelihood of exit (Cockburn and MacGarvie, 2009; Mann and Sager, 2007) and increases the likelihood of obtaining funding at a higher valuation (Greenberg, 2013). The same applies to government research grants (Islam et al., 2018), which increase the likelihood of funding in the six months following the grant, and participation in an accelerator program (Plummer et al., 2016). Strategic marketing (Morgan, 2012) and strong social media presence (Gloor et al., 2020) have been shown to correlate highly positively with business success. Strategic alliances (Hoenig and Henkel, 2015; Ozmel et al., 2013a) and board composition with experienced consultants (Giudici et al., 2020; Soriano and Castrogiovanni, 2012) have also been reported to correlate positively with startup success. In addition, Miloud et al. (2012) found that higher product differentiation, the industry in which the company operates and the growth rate of the industry, the number of founders, and the number of alliances have a positive impact on the likelihood of raising capital and a higher valuation. Finally, the completeness and gender diversity of the management team have been demonstrated to have a positive impact on business performance (Cassion et al., 2020; Hambrick, 1987). According to the study conducted by Gottschalk and Niefert (2011), the presence of at least one female co-founder appears to be associated with higher performance and exit probability.

Numerous studies have examined the impact of entrepreneurs' personal characteristics on the likelihood of startup success. Based on basic demographic characteristics, individuals in their 30s or early 40s are more likely to succeed (McKenzie and Sansone, 2017). A similar effect is observed for years of work experience (Barreira, 2005). Educational background, and in particular graduating from a top university, also increases the likelihood of receiving funding (Judge et al., 1995; Tanyel et al., 1999). In addition, serial entrepreneurs also get a better valuation (Hsu, 2007) and are on average better at timing the market (Ng and Stuart, 2016) i.e., choosing the best time to start a business and the industry they focus on. From a psychological perspective, previous studies show that grit (i.e., persistence in pursuing long-term goals) is an important trait of entrepreneurs that is related to the likelihood of success (Mueller et al., 2017). The same is true for resilience, resourcefulness, and optimism (Baum and Locke, 2004). In addition, social networks and relational capital also have a strong influence on the ability to raise funds and the likelihood of success. Shane and Stuart (2002) showed that relationships with reputable VC investors enhance the entrepreneur's ability to raise funds. Specifically, a strong and robust professional network actually has a positive effect on fundraising ability (Nann et al., 2010). In addition to entrepreneurial characteristics, the composition of the team has also been shown to have a strong influence on the likelihood of success of a start-up. Müller and Murmann (2016) found that a mix of business and technical skills is critical and has a positive impact on business performance. This is also confirmed by Jin et al. (2017), who found that team completeness and heterogeneity are strongly positively correlated with startup success.

Finally, the last group of studies focuses on the financial aspects of the company. Several studies have shown that companies backed by VCs with good reputations are more likely to exit through an IPO or acquisition (Chemmanur et al., 2011; Ozmel et al., 2013b). This effect is also stronger for VCs with prior experience in VC or startups (Zarutskie, 2010) or high specialization (Gompers et al., 2009). Angel investment and support from early investors also appear to increase the likelihood of growth and exit (Kerr et al., 2014). In addition, numerous studies have shown that deal structure has an impact on the probability of success. In particular, it depends on the equity share (Miettinen and Littunen, 2013), whether the company was financed by convertible notes (Cumming and Johan, 2008) or by debt (Cole and Sokolyk, 2018), and whether the deal was syndicated (Das et al., 2011). In addition, the size of the financing also plays a critical role in the likelihood of a startup's success (Groenewegen and Langen, 2012; Nanda et al., 2020).

2.2. Survey on startup success prediction studies

Machine learning has long been used to predict business success. Lussier and Corman (1995) used logistic regression to predict early-stage firm success using data collected through surveys of US firms. Tomy and Pardede (2018) used and compared different machine learning algorithms including k-nearest neighbors, naive Bayes and support vector machines to predict startup success. Bhat and Zaelit (2011) applied random forest algorithms to predict private company exits using data from different industries. In addition, they assessed the importance of the success determinants by ranking the features most relevant to late-stage investment decision-making. However, these studies relied heavily on either financial data provided by VCs, which is not accessible to the broader research community, or on qualitative data collected through questionnaires, which is very time-consuming and limited.

Recently, researchers have increasingly used data from Crunchbase, an open database of business information, to study company success due to the large amount of information that Crunchbase provides publicly (Dalle et al., 2017). For example, Xiang et al. (2012) used Crunchbase data and factual characteristics from TechCrunch articles to predict corporate acquisitions using Bayesian networks. Bento (2018) used Crunchbase data on startups in the U.S. to predict an acquisition or IPO using logistic regression, support vector machine, and random forest algorithms. Arroyo et al. (2019) also used data from Crunchbase to compare the performance of logistic regression, support vector machine, and other machine learning models in predicting startup success. These studies have proven that Crunchbase's data can effectively be used to identify promising startups without the need to conduct an elaborate qualitative assessment or rely on privileged financial records.

3. Data collection and preparation

The study was conducted using data from the Crunchbase database and publicly accessible LinkedIn profile data. Crunchbase is a platform with business information about private and public companies, founders or people in leadership positions, investors, and financing rounds (Ferrati and Muffatto, 2020). The Crunchbase data used for the studies and experiments were collected in September 2021. In addition to the Crunchbase data, we sampled LinkedIn profile data for a subset of companies with individuals who provided their LinkedIn profiles on Crunchbase. LinkedIn is the world's largest professional network with over 500 million members worldwide, containing detailed information on individuals' academic and professional backgrounds (Ramanath et al., 2018). LinkedIn profile data were collected in January 2022.

3.1. Dataset from Crunchbase

The Crunchbase dataset consists of several tables that can be broadly divided into three types of data: information regarding (1) the organization, (2) the people, and (3) the investments. The tables can be joined by unique identifiers, as shown in the simplified entity relationship diagram (ERD) in Fig. 1. The Crunchbase dataset consists of 1.5 million organizations. However, since our study focuses on the success or failure of an organization, a large part of the initial dataset can be excluded, which reduces the dataset to 54,675 organizations (see Section 3.3). Consequently, the descriptive analysis described in this section is limited to the labeled dataset.

3.1.1. *Organization data*

The information about the organizations is provided by the *organizations* and *organization descriptions* table. The *organization* table holds basic information such as name, HQ address, number of employees, website, social media links, email, and phone number. Crunchbase also keeps track of the status of the organization – active, closed, acquired, or IPO (public company). Each organization is also described by its primary role (company or investor) and the categories and subcategories that describe the industry it operates in. Furthermore, *organization descriptions* contain descriptions of the organizations' business. The dataset consists of 54,675 organizations from over 100 countries. Approximately 40% of all organizations originate from the U.S., followed by Chinese (14%) and British organizations (6%). The companies can be divided into 47 business categories, with 50% of all companies belonging to the software business (see Fig. 2a). Note, that the organizations can have multiple categories. In addition, the companies were founded in the last 10 years, with over 80% of all companies reporting fewer than 50 employees (see Fig. 2b). This forms a good data foundation to analyze the problem of startup success prediction.

3.1.2. *People data*

The people table describes individuals who are founders, investors, or employees of the organizations. The table includes the person's name, gender, address, social media account links, organization, and position within the organization. Information about an individual's education is held in the *degrees* table. Each entry might contain information about the subject of the degree, dates of matriculation and graduation, and the institution at which it was studied. Furthermore, information about past jobs connecting organizations and people is provided in the *jobs* table. It includes the position and duration of the job, and in which organization the work was performed. The dataset contains 137,180 individuals of which approximately 82% are male, 17% are female and less than 1% are of mixed gender (e.g., androgynous) with origins from over 100 countries. 73,403° are recorded which are distributed among 52,834 individuals. Approximately 40% of degrees consist of Bachelor of science, followed by Master of Science (20%), "other degree types" (19%), MBA (14%) and Doctoral degree (7%). Over 16,000 unique subject titles are recorded along with the 73,403 completed studies which can be broadly classified into 22 subject groups (see Fig. 3a). 20% of all completed studies are unknown (i.e., unspecified or marked unknown) and 10% cannot be clearly assigned to a group. 16% of all completed studies are related to Computer Science, followed by Business and Management Studies (9%), Economics (6%) and Accounting & Finance (6%). In addition, 206,244 jobs are recorded which are distributed among 168,533 individuals. Over 36,000 unique job titles are recorded which can be broadly categorized into 30 job groups. Approximately 15% of all job positions are CEO, followed by CTO (14%), other leadership positions such as chairman, president, etc. (13%) and founder (10%) (see Fig. 3b, left). Moreover, most jobs are held less than 8 years (see Fig. 3b, right).

3.1.3. *Investment data*

The information about the investments can be partly obtained in the *organization, the* funding *rounds* and *investments* table. The financial data include the number of funding rounds, the date of the last funding event, total funding, and the number of exits from investments. More detailed information about the investments include data about dates of funding events, amount of collected funds, and investment type (e.g., seed, angel funding, Series A, B, C, etc.). 81,774 funding rounds are recorded. Approximately 50% of all

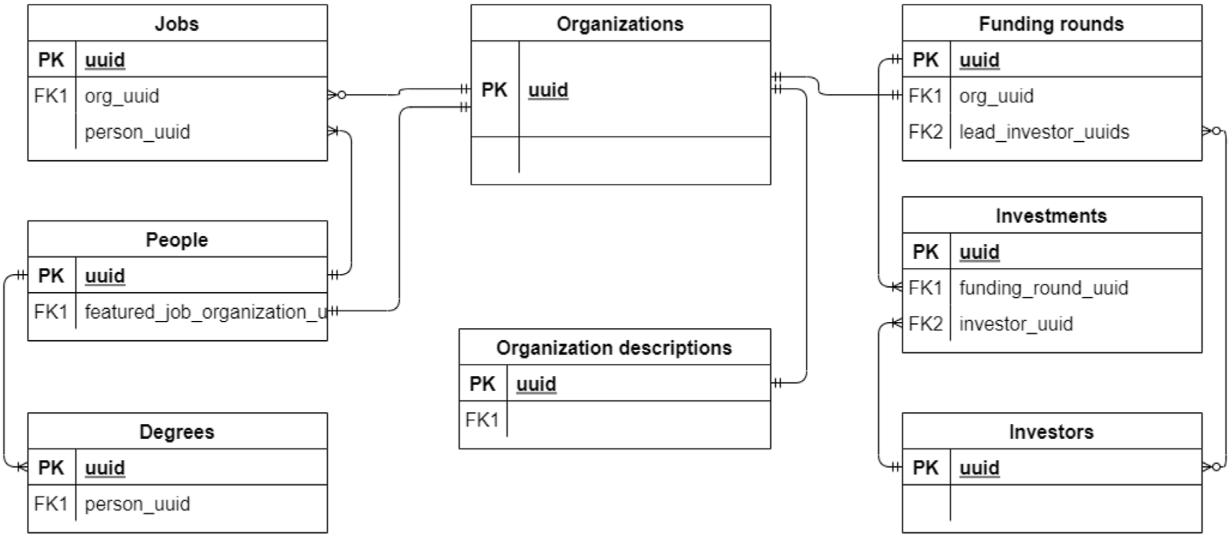
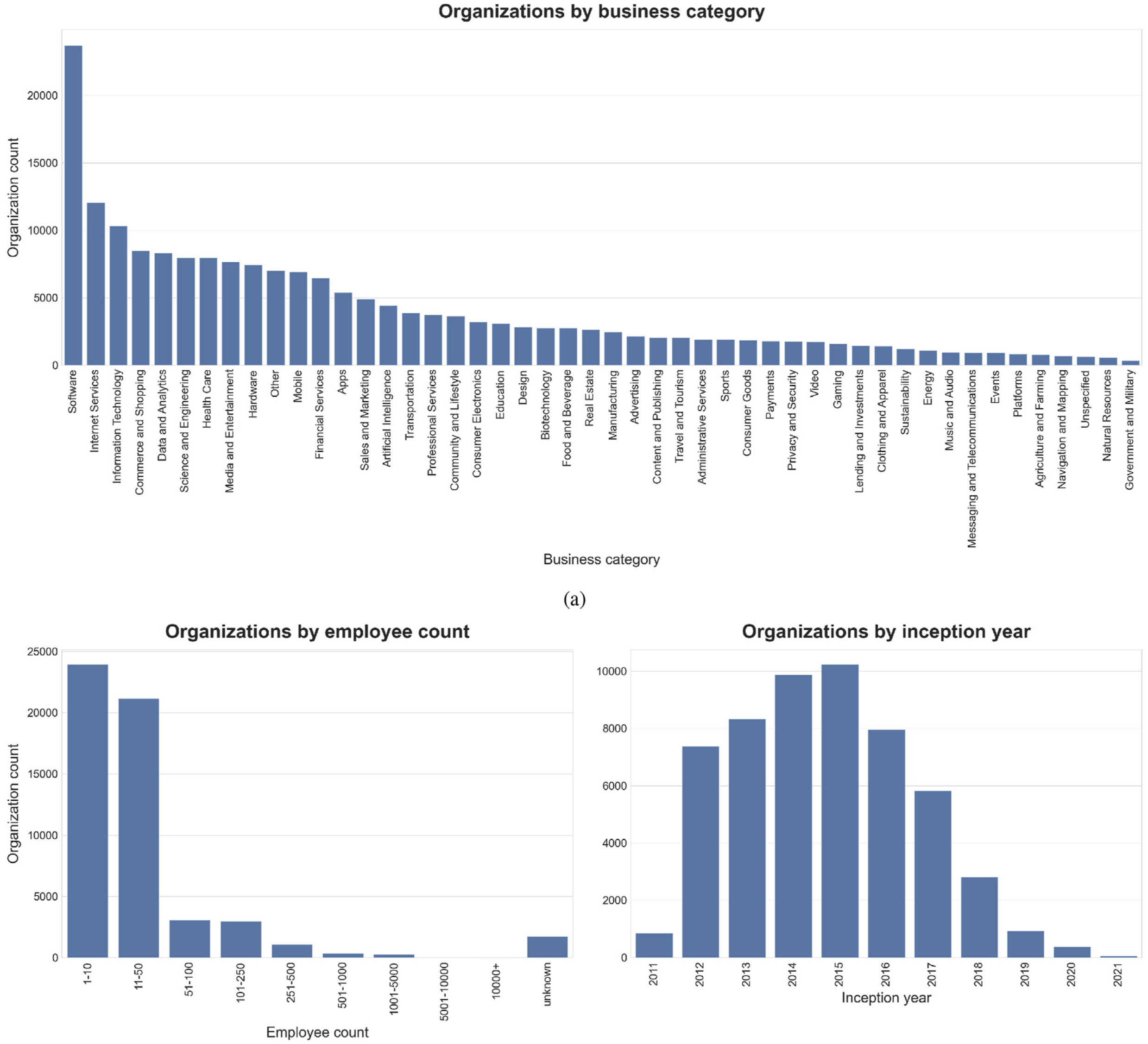


Fig. 1. Simplified ERD diagram of Crunchbase data.



(a)

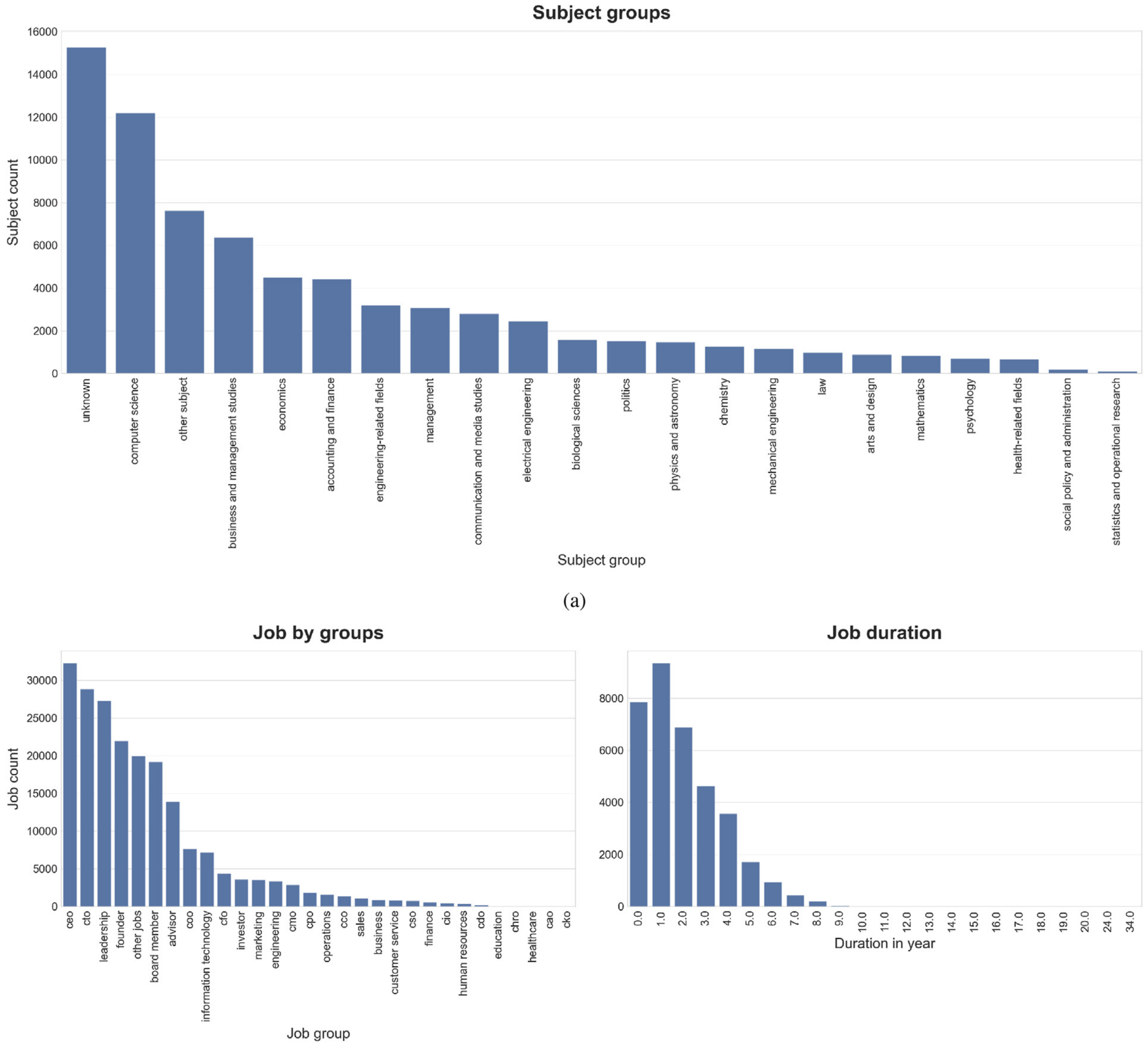
(b)

Fig. 2. (a) Distribution of companies by business category. Companies from the software, Internet and information technology sectors are most represented in the data set. (b) Distribution of companies by employee (left) and by inception year (right). Companies founded between 2012 and 2017 with less than 50 employees are most represented in the dataset.

funding rounds are seed funding, followed by Series A funding (18%), angel investments (6%) and Series B funding (6%) (see Fig. 4, left). The funding rounds are conducted between 2010 and 2020 (see Fig. 4, right).

3.2. *Dataset from publicly accessible LinkedIn profiles*

The LinkedIn dataset contains about 70 structured and unstructured fields with detailed information about individuals, such as work experience, skills, education, awards, certifications, groups and institutions they are part of, projects, and recommendations, amongst others. We collected LinkedIn data from 1247 labeled companies mainly originating from Germany, Switzerland, and Austria, which were provided by Crunchbase and where individuals provided their LinkedIn profile URLs. This resulted in 3204 profiles that can be further used for feature engineering and enrichment. The data is collected by web scraping, and the information is then compiled and stored in tables. The tables can be linked by unique identifiers, as shown in the ERD in Fig. 5. The education and job data provided by LinkedIn includes essentially the same type of information as the data from Crunchbase. The *education* table includes information about the subject of the degree, the date of enrollment and graduation, and the institution where the degree was earned. The *job* table contains information about past and current job positions such as start and end dates, job position and description, and company name and



(a)

(b)

Fig. 3. (a) Distribution of study subjects by groups. Besides the many unspecified study subjects (“unknown” and “other subjects”), computer science is the most commonly studied subject. (b) Distribution of jobs by groups (left) and by job duration (right). The most frequently mentioned job titles include ceo, cto, leadership and founder.

location. In addition, LinkedIn provides other valuable information not available in Crunchbase: detailed information about (1) awards and certifications held by individuals, including title, description, date issued, and name of issuing institution; (2) affiliations such as groups, institutions, and volunteer activities with which the individual is associated; (3) details about past and current projects, including title, description, start and end dates, and a link for more information; and finally, (4) activities and recommendations of individuals, which includes data about articles written, liked, or shared, and comments made.

3.3. Dataset and target creation

The dimension of success used by researchers and practitioners as an object of analysis is not uniform. Various definitions have been used in the studies that have attempted to explain the success of a company. Some researchers advocated the strict use of financial indicators such as sales and profit growth (Ahmad and Seet, 2006; Kotane and Kuzmina-Merlino, 2012), while others emphasized the importance of nonfinancial aspects of organizational success such as personal satisfaction and performance (Simpson et al., 2004; Walker and Brown, 2004). Meanwhile, others used merger and acquisition events or funding events as measures of business success. The fact that a company is acquired by a larger company or that a start-up company successfully obtains funding is a strong indicator of its

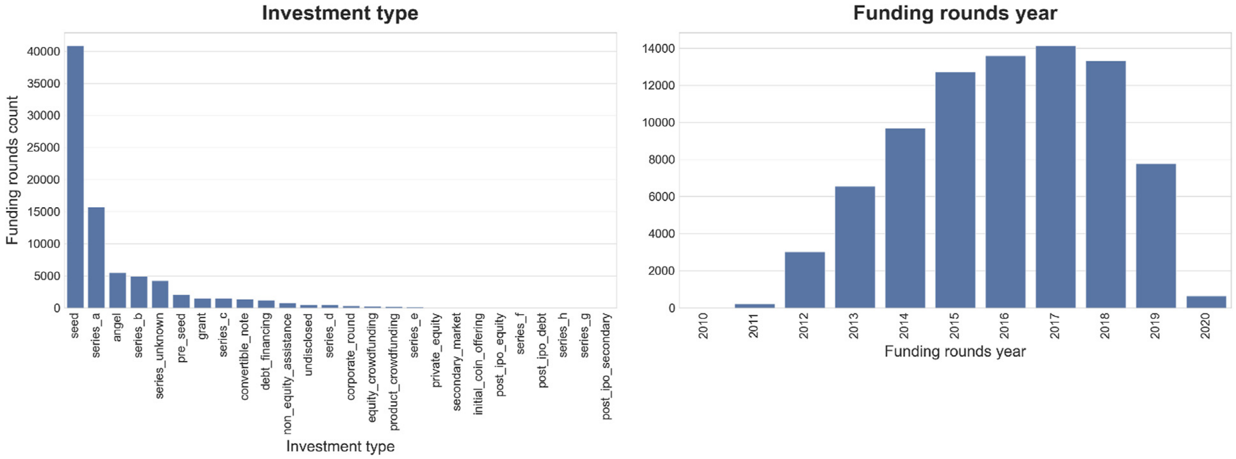


Fig. 4. Distribution of funding rounds by investment type (left) and investment year (right). About 50% of all funding rounds are seed funding. Most of the funding rounds were carried out between 2012 and 2019.

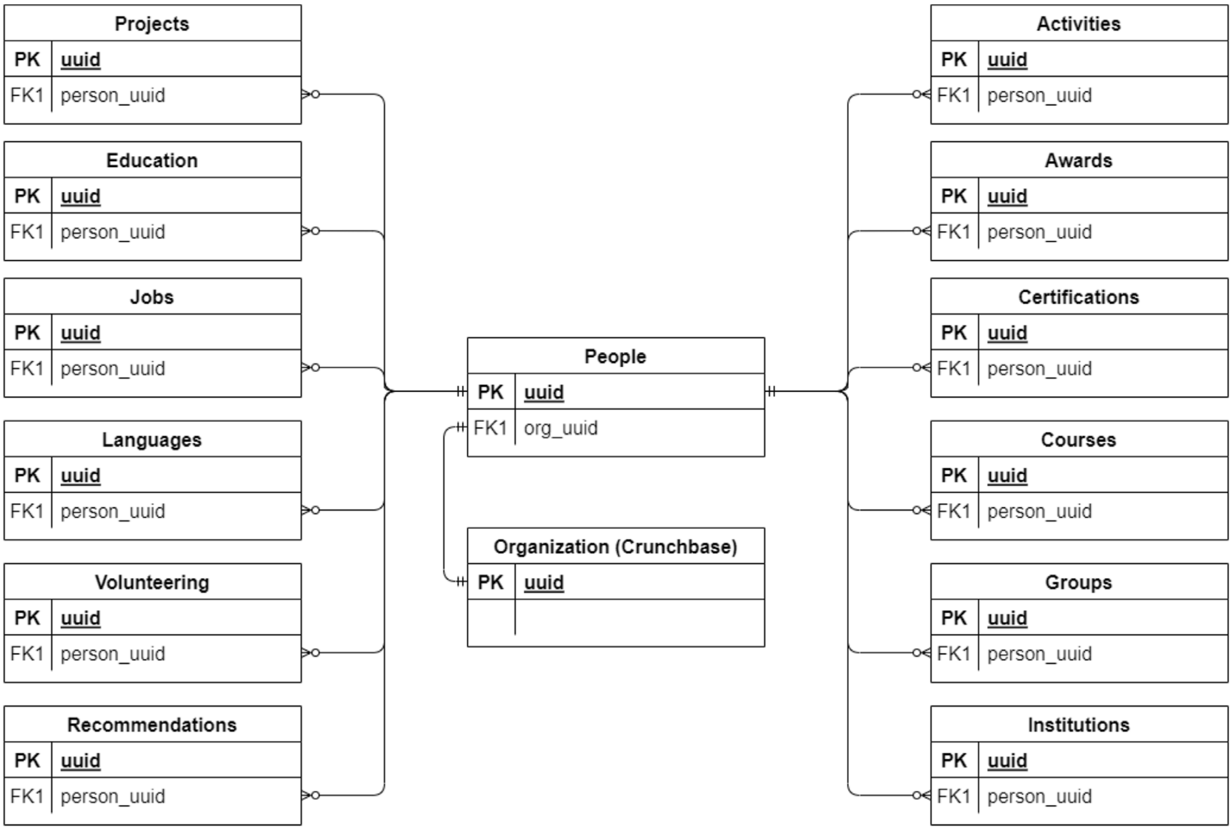


Fig. 5. Simplified ERD diagram of LinkedIn data.

current or potential business value (Davila et al., 2003; Wu and Gnanasambandam, 2017). In our study, we refer to the successful achievement of Series A funding as "success", as this is of great importance to VCs (Wu and Gnanasambandam, 2017). It represents a significant early investment into a startup and signals a successful demonstration of progress and a clear path to revenue growth. Moreover, the receipt of a Series A funding is an objective measure which reflects the potential future business value of the start-up. Therefore, the primary objective of this study is to classify organizations into successful and failed ones (i.e., whether they have received Series A financing or not).

Crunchbase dataset is used to construct a labeled dataset. In this process, funding history and development trajectory of organizations were used as input information for labeling organizations as success or failure. From an initial total of 1.5 million companies, the final subset consisted of 54,475 companies labeled success or failure. The main factors determining the creation of the dataset are presented in the decision flow diagram, which consists of company type, age and size, data availability, and funding information details (see Fig. 6). First, only organizations with a primary role “company” (i.e., not ” “investor”) are included in the dataset. Second, only organizations between 2011 and 2021 are considered in order to examine the determinants of success in the current, rapidly changing startup landscape. Consequently, companies without a valid startup date are discarded and the remaining organizations with a founding date between 2011 and 2021 are extracted. This filtering process resulted in a significant reduction of the original dataset, eliminating approximately one million organizations. Next, organizations without funding information are removed, further reducing the dataset by roughly 0.3 million organizations. From that moment on, the definition of success was straightforward - whether a company received Series A financing or not. Meanwhile, the definition of the failure label still required several steps in which the details of the previous financing as well as the company size were examined in more detail. Specifically, we define failed startups as those that received angel, pre-seed, or seed funding before 2018 but failed to obtain Series A funding thereafter. In addition, we excluded startups with data quality issues, such as startups that received Series B-J funding without successfully obtaining Series A funding. Finally, 22,455 companies were classified as successful, and 32,220 companies were classified as failures.

3.4. *Feature engineering*

Extensive feature engineering is conducted to cover a wide spectrum of success factors identified from the literature. In total, 66 types of features are generated which can be grouped into 3 macro groups according to Corea (2019): 1) organization-related features (company), individual-related features (demography, education, work experience, amongst others) and investment-related features (investments and investors). Feature engineering is applied separately to Crunchbase and LinkedIn data. While Crunchbase is characterized by a high information content in terms of organizations and investments, LinkedIn contains more detailed information on individuals. Table 1 provides an overview of the created features based on the availability of data in Crunchbase and LinkedIn.

3.4.1. *Organization-related features*

Features related to the company focus on its web presence, the nature of its business, the number of founders involved in the company, and whether the company has any advisors on board. Web presence involves two attributes, namely social media presence and website ending. Features related to the nature of the business are divided into category group (business categories) and category list (subcategories) both of which are only available in the Crunchbase dataset. In total, 47 category groups (e.g., software, internet services, etc.) and 711 subcategories (e.g., e-commerce, mobile, Fintech, etc.) exist. One-hot encoding is applied to transform this information into categorical features for machine learning.

Features related to the number of founders and the involvement of advisors are derived from jobs, since this information is not explicitly available in both datasets. To determine the number of founders, a keyword search is conducted on the job titles of individuals associated with a particular company. Keywords include founder, co-founder and owner. The number of founders is then simply the number of people to which the keywords apply. Similarly, the keyword advisor is applied to job descriptions to determine whether a particular company has hired advisors.

3.4.2. *Individual-related features*

Features related to the individuals focus on the founders' demography, education, work experience, awards, certifications, and other attributes specific to LinkedIn such as affiliations and activities. Demographic features include the people's gender and whether the

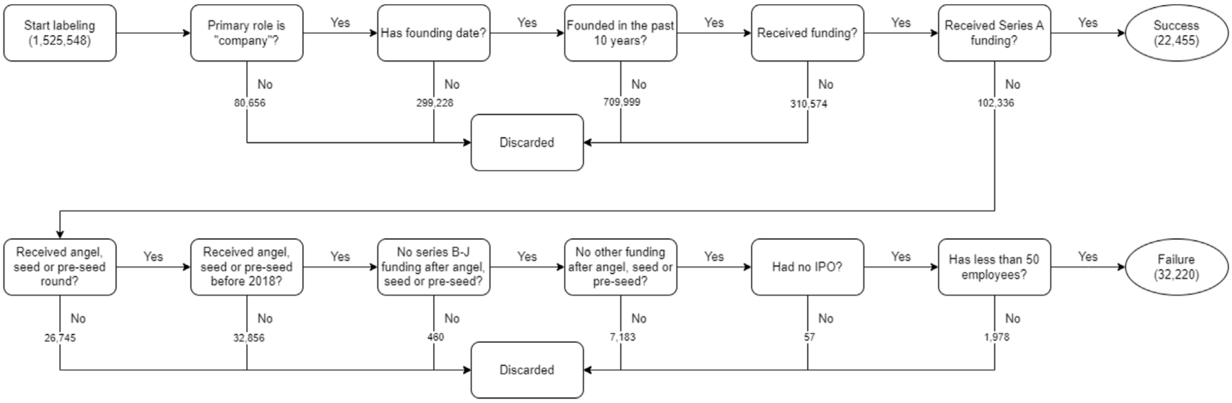

Start labeling (1,525,548)
Primary role is "company"?
Yes
Has founding date?
Yes
Founded in the past 10 years?
Yes
Received funding?
Yes
Received Series A funding?
Yes
Success (22,455)
No
80,656
No
299,228
No
709,999
No
310,574
No
102,336
Discarded
Received angel, seed or pre-seed round?
Yes
Received angel, seed or pre-seed before 2018?
Yes
No series B-J funding after angel, seed or pre-seed?
Yes
No other funding after angel, seed or pre-seed?
Yes
Had no IPO?
Yes
Has less than 50 employees?
Yes
Failure (32,220)
No
26,745
No
32,856
No
460
No
7,183
No
57
No
1,978
Discarded

Fig. 6. Decision flowchart for classifying companies into success and failure based on information related to company type, age and size, data availability, and funding information details. The numbers represent the number of startups retained at each filtering step, including the final dataset of successful and failed startups.

Table 1
Overview of created features based on the availability of data in Crunchbase and LinkedIn. Features can be broadly grouped into organizational-related, individual-related, and investment-related features.

Feature group	Feature type	Feature	Crunchbase	LinkedIn
Organization	Company	Social media presence	x	
		Category group	x	
		Category list	x	
		founder count	x	x
		has experienced advisor	x	x
Individual	Demography	is founder	x	x
		gender of founders	x	x
		location		x
		number of connections		x
	Education	number of degrees	x	x
		degree group	x	x
		number of subjects	x	x
		subject group	x	x
		number of universities	x	x
		latest degree completed on	x	x
		attended top university	x	x
		number of courses		x
		number of languages		x
		English speaking		x
		German speaking		x
	Work experience	number of jobs	x	x
		job type	x	x
		job title group	x	x
		years of work experience	x	x
		serial entrepreneur	x	x
		number of current position		x
		longest stay in company		x
	Awards and certifications	number of awards		x
		award_type (prize, etc)		x
		number of certifications		x
		certification type (online courses, etc)		x
	Affiliations	number of groups		x
		number of organizations		x
		number of volunteerings		x
		volunteering type		x
		years of volunteering		x
	Activities and recommendations	number of activities		x
		activities, number of authored articles		x
		activities, number of liked/shared articles		x
		activities type (hire, investment, partnerships)		x
		number of recommendations		x
	Projects	number of projects		x
		number of current project		x
		years project experience		x
		longest project		x
Investment	Investments	number of funding rounds	x	
		time incorp. until first investment before Series A	x	
		time incorp. until last investment before Series A	x	
		investment size	x	
		money currency type	x	
		investment type (seed, angel, pre_seed)	x	
		investor count	x	
		country	x	
	Investors	type (person, org)	x	
		investor type: angel/seed/pre-seed investors	x	
		web presence	x	
		country, region	x	
		total funding usd	x	
		investment count	x	

individuals involved in the company is a (co-) founder. Extracting gender is a straightforward process as this information is directly available in the people table. To determine whether an individual involved in the company is a founder, a keyword search using “ounder, co-founder, and owner is applied on the job titles as described earlier.

Features related to education include the number and types of degrees and study subjects, the number of distinct universities attended, the elapsed time in days since the last degree was completed, and whether the individual attended a top 100 university according to QS World University Rankings (2021). The number of degrees is determined by counting the number degrees an individual holds. Similarly, the number of subjects is calculated by counting distinct subject titles an individual studied. Creating features related to degree groups and subject groups required additional work, because the degree and subject fields are free texts which need to be unified before they can be used to create features. Therefore, to determine the degree groups and subject groups, a list of degree types and study subjects is constructed based on our own domain expertise and QS World University Rankings by Subject (2021). The list of degree types includes MBA, PhD, MSc, BSc, and "other degree types". The list of study subjects includes computer science, economics, electrical engineering, accounting and finance, business and management studies, mechanical engineering, amongst others. A keyword search approach similarly to above is applied to create the degree groups and subject groups categorization. In addition, one-hot encoding is used to create binary input features from the degree and subject categories for the machine learning models. The number of universities is determined by counting the individual universities where the degrees were completed. In addition, a string fuzzy-matching algorithm (Cohen, 2011) is applied to determine whether an individual attended a top 100 university according to QS World University Rankings 2021.

Features related to work experience include the number of prior jobs, years of work experience, prior job type and job title groups, and whether the individual is a serial entrepreneur. The number of prior jobs is determined by counting the number of jobs a person has performed in the past. The years of work experience are calculated by summing the duration of all jobs, which can be calculated by subtracting the start date from the end date of the jobs. Job type is provided by Crunchbase and describes the function of the person in a company in general, comprising the following 5 types: employee, executive, advisor, board member, and board observer. Job positions are derived from the job titles which are provided as free texts in Crunchbase data. Therefore, similar to creating the subject groups feature, a list of job positions is constructed based on most commonly used job titles provided by the job portal Indeed (2021). The list of job positions consists of 35 job title groups, namely: board member, advisor, investor, chief executive officer (CEO), chief operating officer (COO), chief financial officer (CFO), chief information officer (CIO), chief technology officer (CTO), chief compliance officer (CCO), to name a few. A keyword search approach is then applied to create the job title groups categorization. Finally, to determine whether a person is a serial entrepreneur, keywords such as founder, co-founder and owner are used to count previous activities related to startup creation.

Furthermore, the creation of features must be conducted on an organizational level since we are aiming at predicting the success of a startup. Therefore, features related to individuals are aggregated on an organizational level by applying mathematical operations such as *min*(), *max*(), *mean*(), *std*(), *sum*(), etc. (see Fig. 7). As a result, over 600 numerical and categorical features are created from the

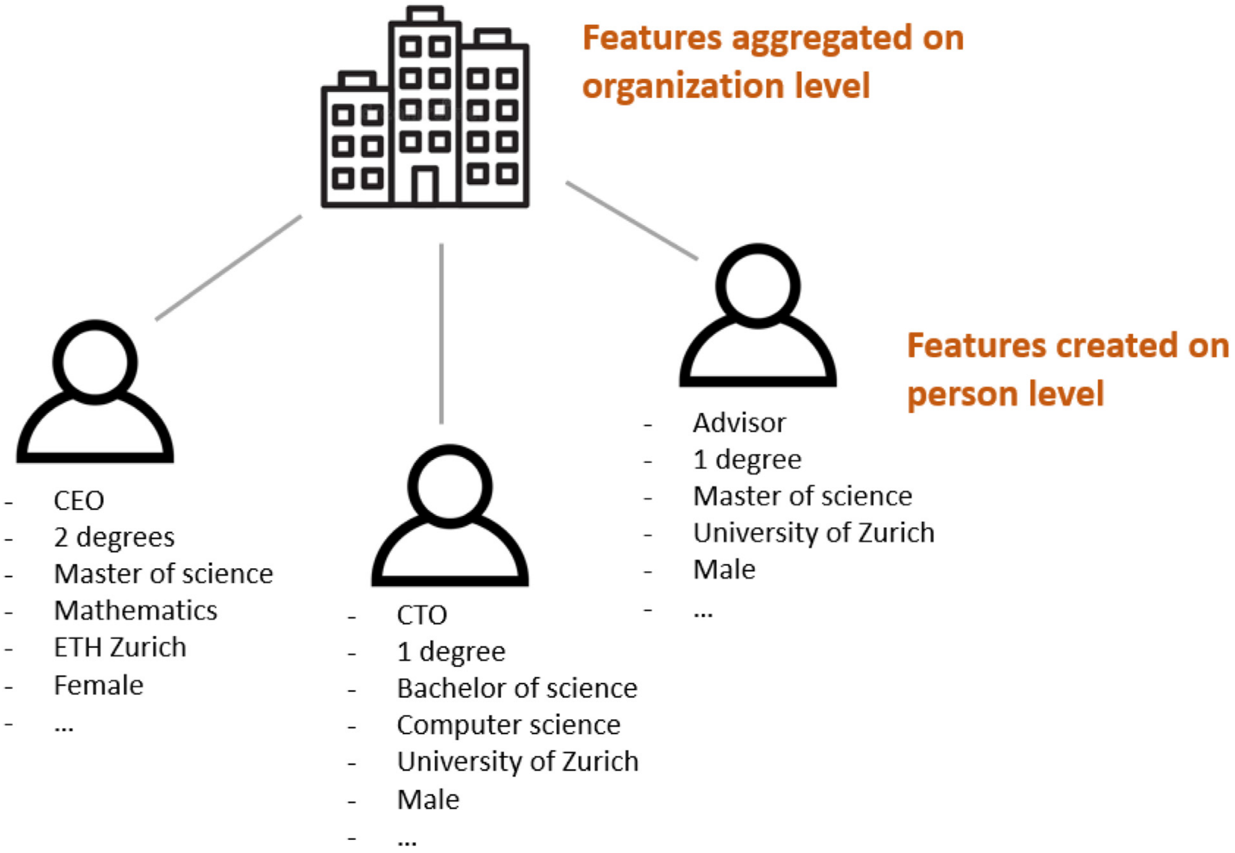


Fig. 7. Illustration of the feature creation process for individual-related features. Features are first created for each person. The features are then aggregated at the organization level using mathematical operations such as *min*(), *max*(), *mean*(), *sum*(), etc.

individual-related features which serve as an input to supervised machine learning algorithms.

3.4.3. *Investment-related features*

Characteristics related to investments can be divided into investments and investors, all of which are available exclusively in the Crunchbase dataset. Investments features include the number of funding rounds, investment size in USD, type of currency in which an investment is transacted, investment type (e.g., pre-seed, seed, angel, etc.), the number of investors participating in a funding round, and the elapsed time in days from a company's inception until the first and last investment is completed. The number of funding rounds is determined by counting the historical funding rounds. Since the target variable is the occurrence of a Series A funding, only funding rounds up to Series A-H (if exists) are considered for all features related to investment (see Fig. 8). Likewise, the elapsed time in days until the first and last investment only considers investments up to Series A. The investment size is calculated by summing up the money collected in each funding round. The number of investors is determined by counting unique investors participating in the funding rounds. Features related to investors include the type of investor (i.e., person or organization), investor's investment focus, web presence, country, total investment in USD, and number of investments transacted.

All investment-related features must be created on an organizational level since we are aiming at predicting the success of a startup. Therefore, features related to investments are aggregated on an organizational level by applying mathematical operations such as *min*(), *max*(), *mean*(), *std*(), *sum*(), etc. (see Fig. 9). As a result, over 500 numerical and categorical features are created from the investment-related features which serve as an input to supervised machine learning algorithms.

After creating the initial set of features, additional steps are undertaken to optimize the performance of machine learning algorithms. Crunchbase's data originates from a structured database, where the information often depends on the user's input. Therefore, much of the information provided is incomplete, resulting in incomplete features. Despite the frequent occurrence and relevance of the missing data problem, many machine learning algorithms handle missing data quite naively. The processing of missing data should be handled carefully, otherwise biases can be introduced into the induced knowledge (Batista and Monard, 2003).

To address this issue, the following measures were taken based on the business and structural characteristics of the features: 1) deletion of samples if too many information is missing (i.e. missing data more than 50%), 2) deletion of features when imputation is not appropriate, 3) imputation of missing numerical values with missForest (Stekhoven and Bühlmann, 2012), and 4) imputation of missing categorical values with −1, which indicates that the information is missing. Furthermore, features with zero variance are discarded and high correlated features are removed iteratively based on variance inflation factor to avoid multicollinearity which may lead to reduced model performance (Folli et al., 2020). Finally, the initial complete feature set is reduced from 1200 features to approximately 400 features for the purposes of supervised machine learning.

4. Methodology

The success of a startup is a highly complex matter which is influenced by a variety of factors. Thus, predicting the success of startups requires machine learning algorithms which are capable of handling a high level of complexity. Therefore, we use Light Gradient Boosting (LGBM), which can model complex interactions between the input variables and thus, share a predominant role in a range of research domains (Fan et al., 2019; Li et al., 2018; Sharma and Naaz Mir, 2019; Taha and Malebary, 2020).

The label creation and dataset selection (i.e., startups only) are extensively covered in Section 3.3. The dataset used for model training consists of approximately 54,000 startups labeled either "success" or "failure". As in any supervised machine learning algorithm, the general purpose is to model the relationship between inputs and outputs in the training set such that it allows generalization, or to generate meaningful results for new inputs not included in the training data, also called generalization performance (Wang et al.,

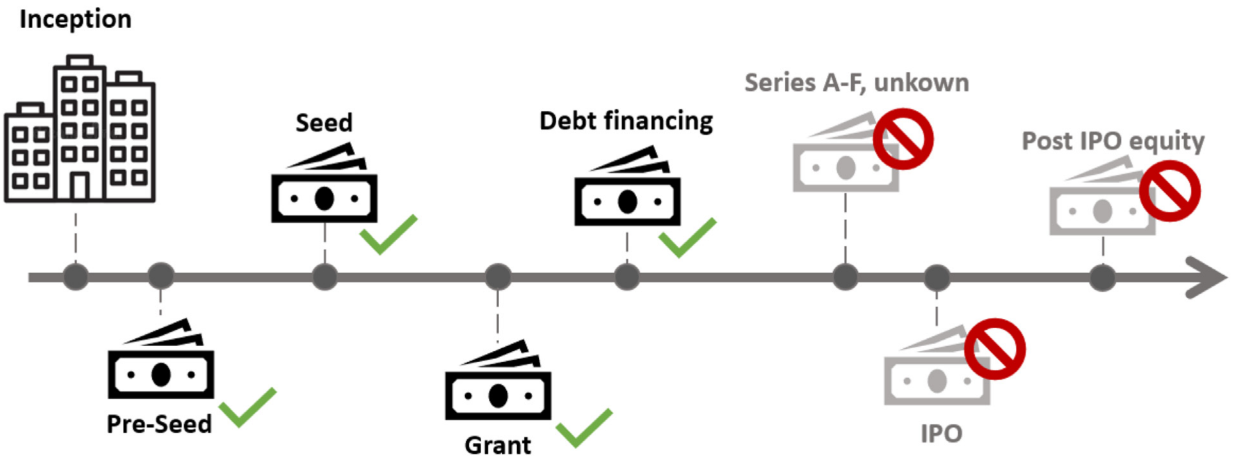


Fig. 8. Selection process for funding rounds to create investment-related features. Only information on funding rounds that predate the occurrence of a Series A-H is considered.

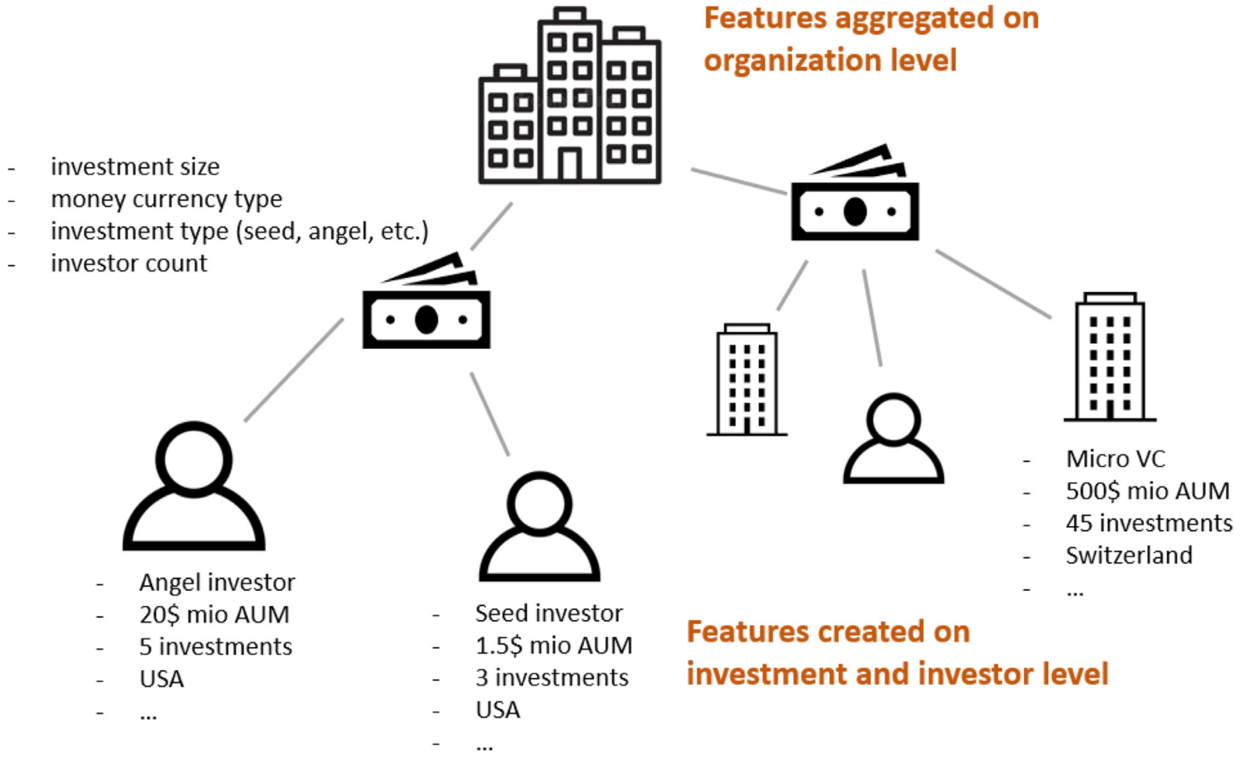


Fig. 9. Illustration of the feature creation process for investment-related features. The features are first created at the investment and investor level. The features are then aggregated at the organization level using mathematical operations such as *min*(), *max*(), *mean*(), *sum*(), etc.

2016). Therefore, to optimize the generalization performance, the dataset is split into train and test dataset at 85:15 ratio. The training set is used for hyperparameter tuning and model training, while the test data set is used to report models' performance. To determine the optimal set of model parameters, we use the random grid search method with 100 iterations combined with ten-fold cross-validation (Bergstra and Bengio, 2012). Furthermore, to reduce the variance due to the training-test split, and to obtain reliable performance estimation for model comparison, we repeated the procedure multiple times. Therefore, the dataset is successively split into training and test set, and the proposed procedure is executed five times. In this approach, the dataset is reshuffled before each round, and the average performance of the models is reported. Furthermore, we find that the performance results on the training set are on average 4% (±0.8%) higher than on the test set.

To compare and evaluate the classification performance of our models, we use four performance measures, namely area under the Receiver Operating Characteristic Curve (AUC) - a commonly used measure for model comparison and effective evaluation of the accuracy measure, accuracy - the overall percentage of samples correctly classified, precision - the fraction of successful startups correctly classified as successful, sensitivity - the fraction of samples correctly classified as successful startups, and specificity - the fraction of samples correctly classified as failed startups. Here, the performance measures are determined for each repeat, and finally averaged and reported as the mean performance of the classification method along with the standard deviation.

While the AUC is a global measure of the model, the choice of a meaningful cut-off point in the ROC curve is critical for specifying optimized accuracy, sensitivity, and specificity. Several methods have been proposed for selecting optimal cut-off points (Hajian-Tilaki, 2018). In the present study, we focus on the Youden index method which is widely used in many research fields (Bantis et al., 2019; Fluss et al., 2005). From a graphical point of view, the Youden index is the maximum vertical distance between the ROC curve and the imaginary diagonal random line from (0, 0) to (1, 1). In summary, accuracy, sensitivity, and specificity are reported for the optimized cut-off point. The overall procedure is illustrated in Fig. 10.

To validate our models, two baseline models are trained based on Crunchbase data: first, a baseline model using the full Crunchbase dataset (Baseline 1), which contains 54,475 startups and second, baseline model using a smaller dataset containing 1247 startups (Baseline 2), which is limited by the amount of data scraped from LinkedIn (see Section 3.2). Furthermore, to assess the added value of LinkedIn data, we trained two models: a model based on LinkedIn (Li model) data to evaluate the possibility to build a success prediction model solely based on publicly available LinkedIn data, and a model which uses both Crunchbase and LinkedIn data (Cb-Li model). Finally, to study how different feature groups contribute to the models, we conduct several experiments based on the feature groups according to Corea (2019). Baseline 1 exploits the full data potential of Crunchbase, which allows us to compare results from related studies and to validate our experiments. Baseline 2 enables a direct comparison of the trained models.

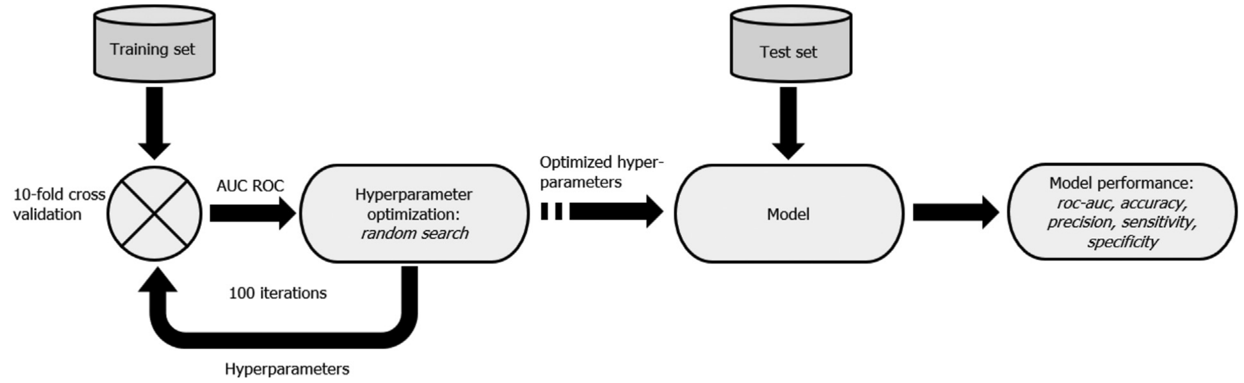

Training set
Test set
10-fold cross validation
AUC ROC
Hyperparameter optimization: random search
Optimized hyper-parameters
Model
Model performance: roc-auc, accuracy, precision, sensitivity, specificity
100 iterations
Hyperparameters

Fig. 10. Overview of model training and performance reporting. The dataset is split in a ratio of 85:15 into training and test data. The training dataset is used to optimize the hyperparameters and train the model, while the test data is used to evaluate the model performance.

5. Model results

Fig. 11 provides an overview of the mean ROC curves for 5 repetitions along with the mean AUC and standard deviations. Table 2 further summarizes the model performances including accuracy, precision, sensitivity and specificity. Our experiments demonstrate that Baseline 1 performs better than Baseline 2. Baseline 1 with the complete feature set (see Table 2: column "complete") achieves the best overall model performance (AUC = 0.87 ± 0.02). Moreover, the performance of Baseline 1 is comparable to the performance of related studies (Li, 2020; Sharchilev et al., 2018; Żbikowski and Antosiuk, 2021). Both baseline models indicate that features related to investment have high impact on model performance, while features related to people have low predictive value. Features related to organization show mixed results: for Baseline 1, the features have a moderate impact on model performance, and their predictive value is rather small for Baseline 2. Arguably, this can be attributed to the fact that the features related to organization contain mainly categorical features with high cardinality (e.g., business type) and thus inevitably lead to lower performance with little training data (Gupta and Asha, 2020). The Li-model, as a stand-alone model on the Linkedin database, performs worse than the two baseline models (AUC = 0.73 ± 0.02). While features related to organization have a moderate impact on model performance, features related to individuals show a higher predictive value compared to their effects on baseline models. The effect of features related to investment cannot be assessed because no information on investment is included in the LinkedIn data.

The Cb-li model combining features created from Crunchbase and LinkedIn data (AUC = 0.84 ± 0.04) outperforms both Li-model and Baseline 2. Features related to investments contribute most to model performance, followed by features related to individuals and organizations. However, the Cb-li model performs slightly worse compared to Baseline 1. Presumably, this is due to the fact that Baseline 1 was trained on approximately 40 times more data than the Cb-li model.

6. Model interpretation

Understanding model decisions is essential for evaluating the consistency of predictions and identifying potential sources of model bias. In addition, the interpretability of models is also critical for acquiring knowledge from modeling practice (Balfer and Bajorath, 2015; Ribeiro et al., 2016). This is particularly important when applying artificial intelligence in the VC industry for decision making, where wrong investment decisions can cause fatal consequences (Jain, 2018).

In recent years, the Shapley approach has proven to be a powerful resource for explaining complex models (Datta et al., 2016; Merrick and Taly, 2020; Štrumbelj and Kononenko, 2014). The Shapley value is a concept from game theory used to determine the contribution of each player in a coalition or cooperative game (Roth, 1988). It can be applied in machine learning to explain the contributions of features, where the features are the players and the model prediction is the payoff of the game. To calculate the importance of feature j, the process can be intuitively represented as drawing feature values in random order for all features except feature j for each iteration, before computing the difference of the prediction with and without feature j. Essentially, the Shapley value is the average marginal contribution of a feature given all possible combinations (Winter, 2002).

In this study, we apply SHAP (Shapley Additive exPlanations), an interpretation method based on Shapley values introduced by Lundberg and Lee (2017). SHAP provides a more comprehensive view of feature importance when compared to conventional feature importance scores derived from tree-based machine learning models, and can additionally be used to explain individual predictions of any machine learning model (Zafar and Khan, 2021). We focus on explaining the Li model and the Cb-li model which are the main subject of our study. Fig. 12 shows the SHAP summary plots for Li model (left) and Cb-li model (right), combining the importance of the top 20 features (ranked from top to bottom) with the Shapley values for each prediction. In the Cb-li model, the data sources of the features are indicated by the prefixes "cb_" (for Crunchbase) and "li_" (for LinkedIn). Positive SHAP values increase the probability of a startup being successful, while negative values decrease it. Red dots represent a high feature value, while blue dots represent a low feature value (for binary features: red represents 1 and blue represents 0).

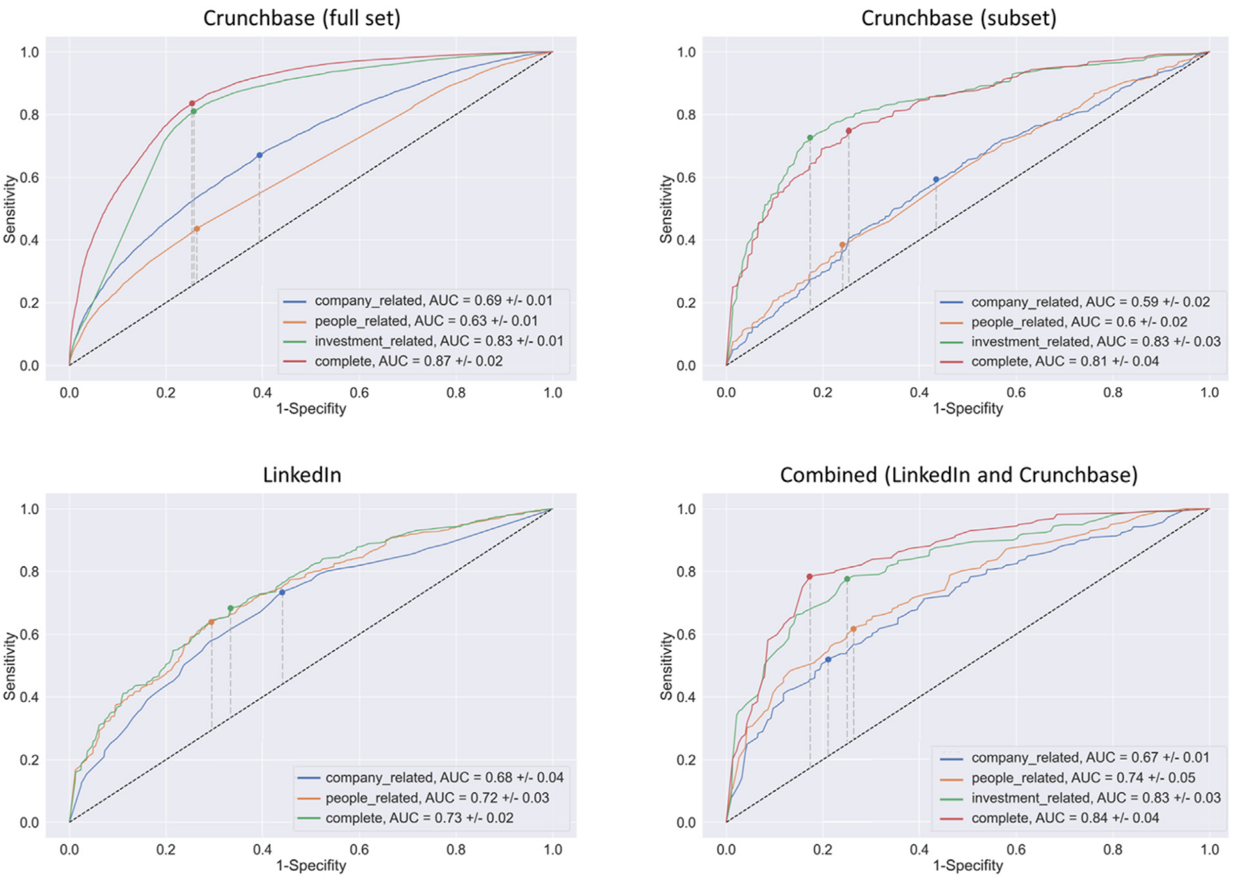


Fig. 11. ROC curves of model trained on Crunchbase full dataset (upper left), on Crunchbase subset (upper right), on LinkedIn data (lower left), and LinkedIn and Crunchbase combined (lower right). The point on the ROC curves represents the optimal threshold at which accuracy, precision, sensitivity, and specificity are indicated.

Table 2
Summary of model performances using AUC, accuracy (acc), precision (prec), sensitivity (sens), and specificity (spec). Two baseline models are trained based on Crunchbase data: a model using the full dataset (Crunchbase (full set)) and using a subset (Crunchbase (subset)). The models solely based on LinkedIn and combined data sources use the same subset of startups.

Data source	Attribute Group			
	Organization	Individual	Investment	Complete
Crunchbase (full set)	AUC = 0.69 (±0.01)	AUC = 0.63 (±0.01)	AUC = 0.83 (±0.01)	AUC = 0.87 (±0.02)
	Acc = 0.64 (±0.02)	Acc = 0.61 (±0.01)	Acc = 0.78 (±0.01)	Acc = 0.79 (±0.01)
	Prec = 0.63 (±0.02)	Prec = 0.59 (±0.02)	Prec = 0.77 (±0.02)	Prec = 0.78 (±0.02)
	Sens = 0.64 (±0.02)	Sens = 0.47 (±0.02)	Sens = 0.78 (±0.02)	Sens = 0.79 (±0.01)
	Spec = 0.64 (±0.02)	Spec = 0.71 (±0.02)	Spec = 0.78 (±0.02)	Spec = 0.79 (±0.02)
Crunchbase (subset)	AUC = 0.59 (±0.02)	AUC = 0.60 (±0.02)	AUC = 0.83 (±0.03)	AUC = 0.81 (±0.04)
	Acc = 0.58 (±0.02)	Acc = 0.55 (±0.03)	Acc = 0.77 (±0.04)	Acc = 0.75 (±0.04)
	Prec = 0.58 (±0.02)	Prec = 0.56 (±0.03)	Prec = 0.77 (±0.04)	Prec = 0.75 (±0.05)
	Sens = 0.57 (±0.02)	Sens = 0.41 (±0.02)	Sens = 0.77 (±0.04)	Sens = 0.74 (±0.04)
	Spec = 0.58 (±0.02)	Spec = 0.71 (±0.04)	Spec = 0.78 (±0.03)	Spec = 0.76 (±0.04)
LinkedIn	AUC = 0.68 (±0.04)	AUC = 0.72 (±0.03)	–	AUC = 0.73 (±0.02)
	Acc = 0.63 (±0.02)	Acc = 0.67 (±0.02)	–	Acc = 0.67 (±0.01)
	Prec = 0.64 (±0.02)	Prec = 0.67 (±0.02)	–	Prec = 0.67 (±0.02)
	Sens = 0.58 (±0.03)	Sens = 0.67 (±0.02)	–	Sens = 0.67 (±0.01)
	Spec = 0.70 (±0.03)	Spec = 0.68 (±0.02)	–	Spec = 0.68 (±0.02)
Combined (LinkedIn & Crunchbase)	AUC = 0.67 (±0.01)	AUC = 0.74 (±0.05)	AUC = 0.83 (±0.03)	AUC = 0.84 (±0.04)
	Acc = 0.63 (±0.01)	Acc = 0.70 (±0.05)	Acc = 0.77 (±0.04)	Acc = 0.77 (±0.05)
	Prec = 0.63 (±0.01)	Prec = 0.69 (±0.05)	Prec = 0.77 (±0.04)	Prec = 0.77 (±0.04)
	Sens = 0.62 (±0.01)	Sens = 0.69 (±0.05)	Sens = 0.77 (±0.04)	Sens = 0.77 (±0.04)
	Spec = 0.63 (±0.02)	Spec = 0.70 (±0.05)	Spec = 0.78 (±0.03)	Spec = 0.77 (±0.05)

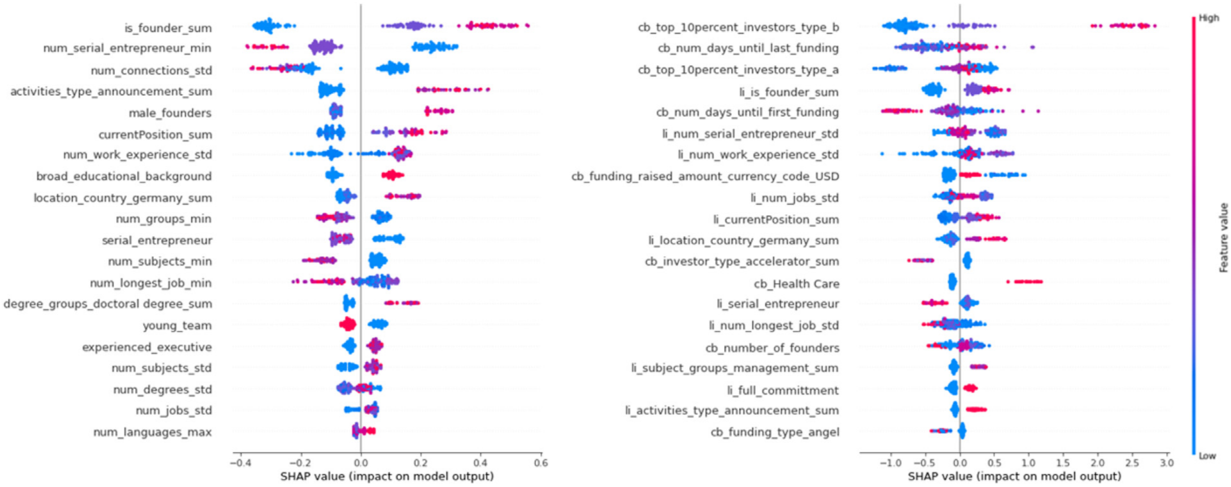


Fig. 12. Shapley summary plots combining feature importance with their local Shapley values on a sample of the dataset. Summary plot for Li-model (left) and Cb-li model (right).

6.1. *Explanation of the Li model*

The top 20 features of Li model are characterized by a mixture of features reflecting factors mainly related to demographic features, work experience, and education. Notably, among the top 20 features are also features created by calculating the standard deviation (indicated by "_std"). The rationale for these features is not clear, as their SHAP values show a mixed pattern. It can be assumed that depending on the features, a balanced team (i.e., low standard deviation) or a diverse team (i.e., high standard deviation) is desirable.

Demographic features include *is_founder_sum* (the number of founders in the startup), *male_founders* (number of male founders), and *young_team* (the average age of the team is less than 35 years). The Li model suggests that a large number of founders and especially male founders in a startup (i.e., a high value for *is_founder_sum* and *male_founders*) significantly increases the probability of success. To our surprise, the probability of success decreases if the average age of the team is below 35 years (recognizable by the negative SHAP values for red dots in *young_team*).

Work experience related features include *num_serial_entrepreneur_min* (number of startups previously founded by the founder with the fewest startups established), *currentPosition_sum* (number of jobs executed in parallel), *num_longest_job_min* (number of days spent on professional activity by the founder with the least work experience), and *experienced_executive* (number of founders with prior leadership experience). The Li model suggests that a startup with founders who have founded a few startups before is more likely to succeed. Moreover, both *currentPosition_sum* and *experienced_executive* are positively correlated with startup success. For *num_longest_job_min*, the model provides mixed results, i.e. both high and low values contribute to success.

Education related features include *broad_educational_background* (founders have broad subject knowledge), *num_subjects_min* (number of studies the founder has completed with the fewest number of degrees), and *degree_groups_doctoral_degree_sum* (number of founders with a doctoral degree). The Li model suggests that a startup with a broad educational background increases the probability of success. However, founders who have studied many subjects do not increase the likelihood of success; on the contrary, they decrease it. Moreover, a doctorate among founders further increases the probability of success of a startup.

6.2. *Explanation of the Cb-li model*

While the top 20 features of the Cb-li model largely cover the same features identified by the Li model (e.g., *is_founder_sum*, *num_jobs_std*, *currentPosition_sum*), features related to investments and investors are predominant in the top 10 features.

Investment related features include *cb_num_days_until_last_funding* (number of days elapsed from the startup's inception to the last round of funding before Series A funding, if any occurred), *cb_num_days_until_first_funding* (number of days elapsed from the company's inception to the first round of funding), and *cb_funding_raised_amount_currency_code_USD* (total amount of funds raised before Series A funding, if any occurred). While *cb_num_days_until_last_funding* does not demonstrate a clear tendency, *cb_num_days_until_first_funding* reveals that receiving the first funding later than early correlates negatively with the likelihood of success (i.e. negative SHAP values for high value of *cb_num_days_until_first_funding*). Furthermore, receiving a large amount of funding does not necessarily lead to a successful Series A funding.

Investor related features include *cb_top_10percent_investors_type_a* (investor among the top 10% in terms of investment volume or number of investments, only including micro VC, incubator, accelerator, and angel investors) and *cb_top_10percent_investors_type_b* (investor among the top 10% in terms of investment volume or number of investments, only including corporate VC, investment bank, hedge fund, pension fund, and private equity firm), which are among the top three features. *cb_top_10percent_investors_type_b* clearly indicates that backing from a medium to large investment firm vastly increases the likelihood of Series A funding. A similar but weaker pattern can be observed for *cb_top_10percent_investors_type_a*.

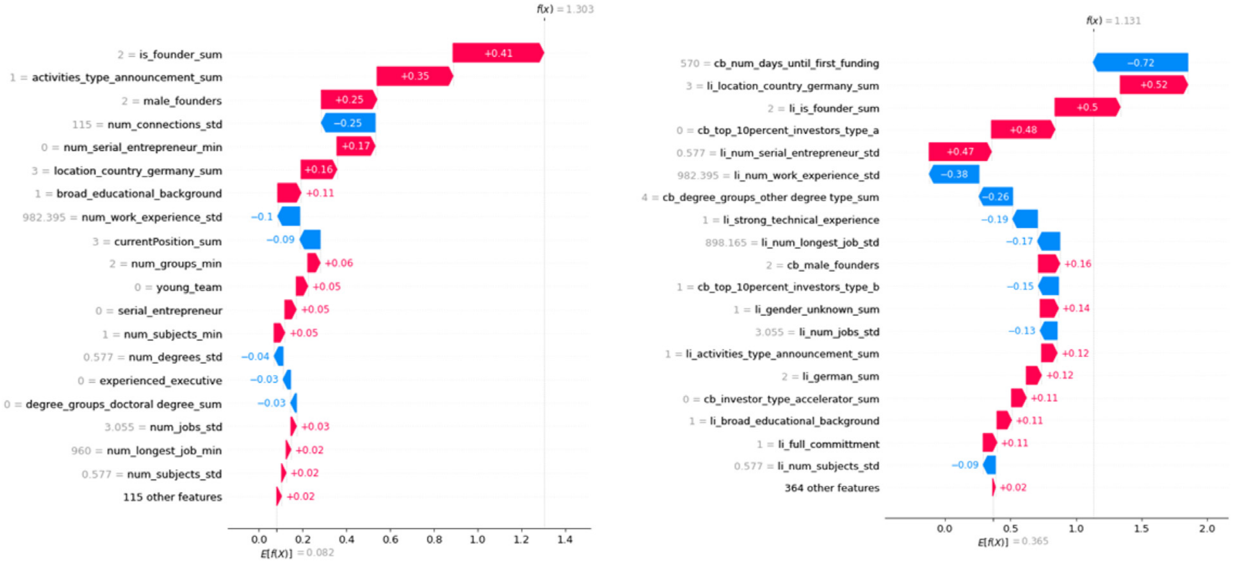


Fig. 13. Shapley waterfall diagrams showing explanations for individual predictions. Waterfall diagram for the Li model (left) and the Cb-li model (right) for an identical prediction.

6.3. Explanation of individual predictions

Fig. 13 shows the SHAP feature importance for Li model (left) and Cb-li model (right) of a single prediction including the feature values. Red arrows indicate a positive contribution to success, while blue arrows indicate a negative contribution. The illustrated example concerns a start-up company based in Germany that provides efficient wireless charging systems for industrial and mobile robotic applications. Both models predicted that the startup is successful in raising Series A funding. However, the decision-making of the models are considerably different.

In the Li model, the number and gender of founders (i.e., *is_founder_sum* and *male_founders*) contribute most significantly to the probability of success, along with features related to announcements posted on LinkedIn (*activities_type_announcement_sum*) and the number of startups previously founded by the team (*num_serial_entrepreneur_min*).

In the Cb-li model, on the other hand, the number of days that have elapsed from the founding of the company to the first round of financing (*cb_num_days_till_first_funding*), number of founders (*li_is_founder_sum*), whether there are top 10% investors in the company (*cb_top_10percent_investors_type_a*), and a feature relating to startups that have founded previously (*li_num_serial_entrepreneur_std*) have a strongest impact on the prediction.

7. Conclusion and application

In this study, we explore the potential of using publicly available LinkedIn profiles as an alternative and as an additional source of data for predicting startup success. First, we provide a comprehensive review of the existing literature on the factors that influence startup success. We then create a comprehensive feature set based on factors identified in prior research and our own considerations using Crunchbase and LinkedIn data. We then train two success prediction models that use LinkedIn data (1) as a standalone and (2) as a complementary data source to Crunchbase, and compare them to two baseline models based solely on Crunchbase data: Baseline 1 uses the full Crunchbase dataset with 54,475 startups, while Baseline 2 uses a smaller dataset with 1247 startups that matches the dataset used to train the two candidate models.

The experiments suggest that using LinkedIn profile data as an alternative data source to Crunchbase for building a success prediction model (Li-model) results in worse prediction performance than using Crunchbase. The Cb-li model which uses both Crunchbase and LinkedIn data outperforms both the Li-model and the Baseline 2 model. However, Cb-li model still performs slightly worse than Baseline 1. This is somehow expected, as Baseline 1 is trained on approximately 40 times more data than the Cb-li model. The results suggest that using publicly accessible LinkedIn profiles in addition to Crunchbase is a promising and viable approach to achieve greater model performance for success prediction. Considering the fact that only a small dataset is collected from LinkedIn in the present work, the true potential of LinkedIn data unfolds if large volumes of web content is collected.

In addition, we apply SHAP to gain a deeper understanding of the reasons behind the model's decision making. The summaries of the characteristics of the Li model and the Cb-li model show that a mixture of characteristics related to investment, demographic factors, work experience, and education play a crucial role in prediction, highlighting the importance of including a wide range of factors when modeling the success of startups.

This work has both theoretical and practical implications. It contributes to the existing literature of startup success research by reinforcing previous findings in a data-driven and model-based manner through supervised machine learning. The approach in the present study can be used to further explore new success factors, for example, based on the feature importance identified by applying the Shapley approach, thus adding to the empirical body of knowledge. In addition, the research results can be used to develop an information system to identify promising startups at scale using Crunchbase and LinkedIn data. The proposed information system enables automated collection and data aggregation of investment-related information in a structured representation of web data, which facilitates the screening and due diligence process for VCs. In addition, VCs can use the proposed information system to monitor startups' performance by regularly checking for changes in Crunchbase or LinkedIn, serving as an "early detection system" for future opportunities for success. Finally, for startups, the information system can be used to evaluate the characteristics of companies based on the information available in Crunchbase and LinkedIn. The absence of important success factors can be pointed out to companies, thus serving as an advisory program.

8. Limitations and future directions

The explainability of machine learning models is an important issue for both scientists and practitioners (Raff and Sylvester, 2018). For example, VCs need to understand why they should invest in and promote a particular startup. This could be due to the fact that they have already invested in similar companies or that the startup has certain value propositions that are of interest to the investor. While SHAP provides a solid understanding of the feature relevance and the reasoning behind the model's decision-making, correlation does not imply causality: it is therefore difficult to draw such inferences from non-linear and highly complex machine learning models such as LGBM. Therefore, recent advances in learning interpretable models will be explored, which will pave the way for learning fair models and representations that are invariant to sensitive attributes such as gender, race, etc. Causal models aim to capture the underlying mechanism driving the decision-making process while ignoring other domain-specific factors. In the future, we aim to train independent models and detect anomalies and strong deviations from the model that may indicate new trends.

Furthermore, the completeness of the used data can be biased towards successful and large companies, as other studies have shown (Retterath and Braun, 2020). In addition, Retterath and Braun (2020) found that greater financing rounds are more likely to be reported than lower ones. The size of funding rounds is also more likely to be reported for larger funding rounds than for smaller ones. Moreover, data quality depends heavily not only on the update schedule of the Crunchbase team, but also on the willingness of users to frequently update their information on Crunchbase and LinkedIn. This fact may cause the predictive power of our models to differ from the actual predictive power. Therefore, multiple or more reliable sources such as Pitchbook and VentureSource are desirable as ground truth to validate the present study (Retterath and Braun, 2020).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Penny Schiffer is co-founder of Raized.ai, which provides venture capitalists promising startups for deal sourcing. Findings of this work are used in the data pipeline of Raized.ai Ltd.

References

Agmon, Tamir, Sjögre, Stefan, 2016. The future of VC funds: the effects of technology and globalization. In: Venture Capital and the Inventive Process. Springer, pp. 107–113.

Ahmad, Noor, Seet, Pi-Shen, 2006. Financial and Non-financial Indicators of Business Success: a Study of Australian and Malaysian SME Entrepeneurs.

Arroyo, Javier, Corea, Francesco, Jimenez-Diaz, Guillermo, Recio-Garcia, Juan A., 2019. Assessment of machine learning performance for decision support in venture capital investments. IEEE Access 7, 124233–124243.

Aryoubi, Abdullah, Hildebrand, Marie, Meser, Michael, 2020. Quantitative easing and its implications on private equity in the euro area. J. Int. Bus. Econom. 8, 1–10.

Balfer, Jenny, Bajorath, Jürgen, 2015. Visualization and interpretation of support vector machine activity predictions. J. Chem. Inf. Model. 55, 1136–1147.

Bantis, Leonidas E., Nakas, Christos T., Benjamin, Reiser, 2019. Construction of confidence intervals for the maximum of the Youden index and the corresponding cutoff point of a continuous biomarker. Biom. J. 61, 138–156.

Barreira, Jose Celestino Dias, 2005. The Influence of Business Knowledge and Work Experience, as Antecedents to Entrepreneurial Success. others. University of Pretoria. PhD thesis.

Batista, Gustavo EAPA., Monard, Maria Carolina, 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17, 519–533.

Baum, J Robert, Locke, Edwin A., 2004. The relationship of entrepreneurial traits, skill, and motivation to subsequent venture growth. J. Appl. Psychol. 89, 587.

Bento, Francisco Ramadas da Silva Ribeiro, 2018. Predicting Start-Up Success with Machine Learning. PhD thesis.

Bergstra, James, Bengio, Yoshua, 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13.

Bhat, Harish S., Zaelit, Daniel, 2011. Predicting private company exits using qualitative data. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 399–410.

Cassion, Christopher, Qian, Yuhang, Bossou, Constant, Ackerman, Margareta, 2020. Investors embrace gender diversity, not female CEOs: the role of gender in startup fundraising. In: International Conference on Intelligent Technologies for Interactive Entertainment. Springer, pp. 145–164.

Chemmanur, Thomas J., Krishnan, Karthik, Nandy, Debarshi K., 2011. How does venture capital financing improve efficiency in private firms? A look beneath the surface. Rev. Financ. Stud. 24, 4037–4090.

Cockburn, Iain M., MacGarvie, Megan J., 2009. Patents, thickets and the financing of early-stage firms: evidence from the software industry. J. Econ. Manag. Strat. 18, 729–773.

Cohen, Adam, 2011. Fuzzywuzzy: fuzzy string matching in python. ChairNerd Blog 22, 51.

Cole, Rebel A., Sokolyk, Tatyana, 2018. Debt financing, survival, and growth of start-up firms. J. Corp. Finance 50, 609–625.

Corea, Francesco, 2019. In: AI and Venture Capital in An Introduction to Data. Springer, pp. 101–110.

Cumming, Douglas, Johan, Sofia Atiqah, 2008. Preplanned exit strategies in venture capital. Eur. Econ. Rev. 52, 1209–1241.

Dalle, Jean-Michel, Den Besten, Matthijs, Menon, Carlo, 2017. Using Crunchbase for Economic and Managerial Research.

Das, Sanjiv R., Jo, Hoje, Kim, Yongtae, 2011. Polishing diamonds in the rough: the sources of syndicated venture performance. J. Financ. Intermediation 20, 199–230.

Datta, Anupam, Sen, Shayak, Zick, Yair, 2016. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: 2016 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 598–617.

Davila, Antonio, Foster, George, Gupta, Mahendra, 2003. Venture capital financing and the growth of startup firms. J. Bus. Ventur. 18, 689–708.

Fan, Junliang, Ma, Xin, Wu, Lifeng, Zhang, Fucang, Xiang, Yu, Zeng, Wenzhi, 2019. Light Gradient Boosting Machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agric. Water Manag. 225, 105758.

Ferrati, Francesco, Muffatto, Moreno, 2020. Using Crunchbase for research in Entrepreneurship: data content and structure in. In: 19th European Conference on Research Methodology for Business and Management Studies (ECRM), Aveiro, Portugal (18-19 June), pp. 342–351.

Ferrati, Francesco, Muffatto, Moreno, 2021. Entrepreneurial finance: emerging approaches using machine learning and big data. others Found. Trend. Entrep. 17, 232–329.

Fluss, Ronen, Faraggi, David, Reiser, Benjamin, 2005. Estimation of the Youden Index and its associated cutoff point. Biom. J.: J. Math. Meth. Biosci. 47, 458–472.

Folli, Gabriely S., Nascimento, Márcia HC., Paulo, Ellisson H., Cunha, Pedro HP., Romão, Wanderson, Filgueiras, Paulo R., 2020. Variable selection in support vector regression using angular search algorithm and variance inflation factor. J. Chemometr. 34, e3282.

Fragkiskos, Apollon, Krasotkina, Olga, Spilker III, Harold D., 2021. Wermers Russ. Modeling Private Equity: A Machine-Learning Approach. *Available at SSRN 3367079*.

Giudici, Giancarlo, Moncayo, Giancarlo Giuffra, Martinazzi, Stefano, 2020. The role of advisors' centrality in the success of Initial Coin Offerings. J. Econ. Bus. 112, 105932.

Gloor, Peter A., Colladon, Andrea Fronzetti, Grippa, Francesca, Hadley, Beth Marie, Woerner, Stephanie, 2020. The impact of social media presence and board member composition on new venture success: evidences from VC-backed US startups. Technol. Forecast. Soc. Change 157, 120098.

Gompers, Paul, Kovner, Anna, Lerner, Josh, 2009. Specialization and success: evidence from venture capital. J. Econ. Manag. Strat. 18, 817–844.

Gottschalk, Sandra, Niefert, Michaela, 2011. Gender differences in business success of German start-up firms. In: ZEW-Centre for European Economic Research Discussion Paper.

Greenberg, Gili, 2013. Small firms, big patents? Estimating patent value using data on israeli start-ups' financing rounds. Eur. Manag. Rev. 10, 183–196.

Groenewegen, Gerard, Langen, Frank, 2012. Critical success factors of the survival of start-ups with a radical innovation. J. Appl. Econ. Bus. Res. 2, 155–171.

Gupta, Heena, Asha, V., 2020. Impact of encoding of high cardinality categorical data to solve prediction problems. J. Comput. Theor. Nanosci. 17, 4197–4201.

Hajian-Tilaki, Karimolla, 2018. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. Stat. Methods Med. Res. 27, 2374–2383.

Hambrick, Donald C., 1987. The top management team: key to strategic success. Calif. Manag. Rev. 30, 88–108.

Harris, Robert S., Jenkinson, Tim, Kaplan, Steven N., 2014. Private equity performance: what do we know? J. Finance 69, 1851–1882.

Hoenig, Daniel, Henkel, Joachim, 2015. Quality signals? The role of patents, alliances, and team experience in venture capital financing. Res. Pol. 44, 1049–1064.

Hsu, David H., 2007. Experienced entrepreneurial founders, organizational capital, and venture capital funding. Res. Pol. 36, 722–741.

Indeed: Career Guide, 2020. 215 Job Titles for Your Resume. (Accessed 15 November 2021).

Islam, Mazhar, Fremeth, Adam, Marcus, Alfred, 2018. Signaling by early stage startups: US government research grants and venture capital funding. J. Bus. Ventur. 33, 35–51.

Jain, Chahat, 2018. Artificial Intelligence in Venture Capital Industry: Opportunities and Risks. PhD thesis. Massachusetts Institute of Technology.

Jin, Linlin, Madison, Kristen, Kraiczy, Nils D., Kellermanns, Franz W., Russell, Crook T., Xi, Jing, 2017. Entrepreneurial team composition characteristics and new venture performance: a meta–analysis. Entrep. Theory Pract. 41, 743–771.

Judge, Timothy A., Cable, Daniel M., Boudreau, John W., Bretz Jr., Robert D., 1995. An empirical investigation of the predictors of executive career success. Person. Psychol. 48, 485–519.

Kerr, William R., Lerner, Josh, Schoar, Antoinette, 2014. The consequences of entrepreneurial finance: evidence from angel financings. Rev. Financ. Stud. 27, 20–55.

Kotane, Inta, Kuzmina-Merlino, Irina, 2012. Assessment of financial indicators for evaluation of business performance. Eur. Integrat. Stud. 6.

Lencioni, Henri, 2020. Assessing Venture Capital Investor Performance Drivers–A Practical Application of Machine Learning. others.

Li, Jinze, 2020. Prediction of the success of startup companies based on support vector machine and random forset. In: 2020 2nd International Workshop on Artificial Intelligence and Education, pp. 5–11.

Li, Fei, Zhang, Li, Chen, Bin, et al., 2018. A light gradient boosting machine for remaining useful life estimation of aircraft engines. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 3562–3567.

Lundberg, Scott M., Lee, Su-In, 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30.

Lussier, Robert N., Corman, Joel, 1995. There are few differences between successful and failed small businesses. J. Small Bus. Strat. 6, 21–34.

Mann, Ronald J., Sager, Thomas W., 2007. Patents, venture capital, and software start-ups. Res. Pol. 36, 193–208.

Matusik, Sharon F., George, Jennifer M., Heeley, Michael B., 2008. Values and judgment under uncertainty: evidence from venture capitalist assessments of founders. Strateg. Entrep. J. 2, 95–115.

McKenzie, David J., Sansone, Dario, 2017. Man vs. machine in predicting successful entrepreneurs: evidence from a business plan competition in Nigeria. In: Machine in Predicting Successful Entrepreneurs: Evidence from a Business Plan Competition in Nigeria (December 2017).

Merrick, Luke, Taly, Ankur, 2020. The explanation game: explaining machine learning models using shapley values. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer, pp. 17–38.

Miettinen, Marika Rosanna, Littunen, Hannu, 2013. Factors contributing to the success of start-up firms using two-point or multiple-point scale models. Enterpren. Res. J. 3, 449–481.

Miloud, Tarek, Aspelund, Arild, Cabrol, Mathieu, 2012. Startup valuation by venture capitalists: an empirical study. Ventur. Cap. 14, 151–174.

Morgan, Neil A., 2012. Marketing and business performance. J. Acad. Market. Sci. 40, 102–119.

Mueller, Brandon A., Wolfe, Marcus T., Imran, Syed, 2017. Passion and grit: an exploration of the pathways leading to venture success. J. Bus. Ventur. 32, 260–279.

Müller, Bettina, Murmann, Martin, 2016. The workforce composition of young firms and product innovation: complementarities in the skills of founders and their early employees tech. rep. In: ZEW Discussion Papers.

Nanda, Ramana, Samila, Sampsa, Sorenson, Olav, 2020. The persistent effect of initial success: evidence from venture capital. J. Financ. Econ. 137, 231–248.

Nann, Stefan, Krauss, Jonas S., Schober, Michael, Gloor, Peter A., Fischbach, Kai, Führes, Hauke, 2010. The Power of Alumni Networks-Success of Startup Companies Correlates with Online Social Network Structure of its Founders.

Ng, Weiyi, Stuart, Toby E., 2016. Of hobos and highfliers: disentangling the classes and careers of technology-based entrepreneurs. In: Unpublished Working Paper.

Ozmel, Umit, Robinson, David T., Stuart, Toby E., 2013a. Strategic alliances, venture capital, and exit decisions in early stage high-tech firms. J. Financ. Econ. 107, 655–670.

Ozmel, Umit, Reuer, Jeffrey J., Gulati, Ranjay, 2013b. Signals across multiple networks: how venture capital and alliance networks affect interorganizational collaboration. Acad. Manag. J. 56, 852–866.

Peneder, Michael, 2010. The impact of venture capital on innovation behaviour and firm growth. Ventur. Cap. 12, 83–107.

Plummer, Lawrence A., Allison, Thomas H., Connelly, Brian L., 2016. Better together? Signaling interactions in new venture pursuit of initial external capital. Acad. Manag. J. 59, 1585–1604.

QS World University Rankings 2021, 2021. Top Global Universities. Top Universities. (Accessed 15 November 2021).

QS World University Rankings by Subject 2021, 2021. Top Global Universities. Top Universities. (Accessed 15 November 2021).

Raff, Edward, Sylvester, Jared, 2018. Gradient reversal against discrimination: a fair neural network learning approach in 2018. In: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 189–198.

Ramanath, Rohan, Inan, Hakan, Polatkan, Gungor, et al., 2018. Towards deep and representation learning for talent search at LinkedIn. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 2253–2261.

Retterath, Andre, 2020. Essays on Machine Learning and the Value of Data in Venture Capital. PhD thesis. Universitätsbibliothek der TU München.

Retterath, Andre, Braun, Reiner, 2020. Benchmarking Venture Capital Databases. *Available at SSRN 3706108*.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2016. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Roth, Alvin E., 1988. The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge University Press.

Schmidt, Christina Maria, 2019. The Impact of Artificial Intelligence on Decision-Making in Venture Capital Firms. PhD thesis.

Shane, Scott, Stuart, Toby, 2002. Organizational endowments and the performance of university start-ups. Manag. Sci. 48, 154–170.

Sharchilev, Boris, Roizner, Michael, Rumyantsev, Andrey, Ozornin, Denis, Serdyukov, Pavel, Rijke, Maarten, 2018. Web-based startup success prediction. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 2283–2291.

Sharma, Vipul, Naaz Mir, Roohie, 2019. An enhanced time efficient technique for image watermarking using ant colony optimization and light gradient boosting algorithm. J. Kind Saud Univ. Comput. Inf. Sci. 34 (3), 615–626.

Simpson, Mike, Tuck, Nicki, Bellamy, Sarah, 2004. Small business success factors: the role of education and training. Educ + Train 46 (8/9), 481–491.

Soriano, Domingo Ribeiro, Castrogiovanni, Gary J., 2012. The impact of education, experience and inner circle advisors on SME performance: insights from a study of public development centers. Small Bus. Econ. 38, 333–349.

Stekhoven, Daniel J., Bühlmann, Peter, 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118.

Štrumbelj, Erik, Kononenko, Igor, 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 647–665.

Sundararajan, Mukund, Najmi, Amir, 2020. The many Shapley values for model explanation. In: International Conference on Machine Learning. PMLR, pp. 9269–9278.

Taha, Altyeb Altaher, Malebary, Sharaf Jameel, 2020. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access 8, 25579–25587.

Tanyel, Faruk, Mitchell, Mark A., McAlum, Harry G., 1999. The skill set for success of new business school graduates: do prospective employers and university faculty agree? J. Educ. Bus. 75, 33–37.

Tomy, Sarath, Pardede, Eric, 2018. From uncertainties to successful start ups: a data analytic approach to predict success in technological entrepreneurship. Sustainability 10, 602.

Tykvová, Tereza, 2018. Venture capital and private equity financing: an overview of recent literature and an agenda for future research. J. Bus. Econ. 88, 325–362.

Walker, Elizabeth, Brown, Alan, 2004. What success factors are important to small business owners? Int. Small Bus. J. 22, 577–594.

Wang, H., Lei, Z., Zhang, X., Zhou, B., Peng, J., 2016. Machine learning basics. Deep Learn. 98–164.

Winter, Eyal, 2002. The shapley value. Handb. Game Theor. Econ. Appl. 3, 2025–2054.

Worthington, Ian, Britton, Chris, 2009. The Business Environment. Pearson education.

Wu, Veronica, Gnanasambandam, Chandra, 2017. A machine-learning approach to venture capital. McKinsey Q. 27.

Xiang, Guang, Zheng, Zeyu, Wen, Miaomiao, Hong, Jason, Carolyn, Rose, Liu, Chao, 2012. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 6, pp. 607–610.

Zafar, Muhammad Rehman, Khan, Naimul, 2021. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach. Learn. Knowl. Extract. 3, 525–541.

Zarutskie, Rebecca, 2010. The role of top management team human capital in venture capital markets: evidence from first-time funds. J. Bus. Ventur. 25, 155–172.

Żbikowski, Kamil, Antosiuk, Piotr, 2021. A machine learning, bias-free approach for predicting business success using Crunchbase data. Inf. Process. Manag. 58, 102555.

Zhang, Qizhen, Ye, Tengyuan, Essaidi, Meryem, Agarwal, Shivani, Liu, Vincent, Loo, Boon Thau, 2017. Predicting startup crowdfunding success through longitudinal social engagement analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1937–1946.