Information Processing and Management 58 (2021) 102555

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A machine learning, bias-free approach for predicting business success using Crunchbase data

Kamil Żbikowski *, Piotr Antosiuk

Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland

ARTICLE INFO

Keywords:
Startups
Supervised learning
XGBoost
Crunchbase
Look-ahead bias

ABSTRACT

Predicting the success of a business venture has always been a struggle for both practitioners and researchers. However, thanks to companies that aggregate data about other firms, it has become possible to create and validate predictive models based on an unprecedented amount of real-world examples. In this study, we use data obtained from one of the largest platforms integrating business information – Crunchbase. Our final training set consisted of 213 171 companies.

This work aims to create a predictive model based on machine learning for the purpose of forecasting a company's success. Many similar attempts have been made in recent years. Plenty of those experiments, often conducted with the use of data gathered from several different sources, reported promising results. However, we found that very often they were significantly biased by their use of data containing information that was a direct consequence of a company reaching some level of success (or failure). Such an approach is a classic example of the look-ahead bias. It leads to very optimistic test results, but any attempt at using such an approach in a real-world scenario may result in dramatic consequences. We designed our experiments in a way that would prevent the leaking of any information unavailable at the decision moment to the training set.

We compared three algorithms – logistic regression, support vector machine, and the gradient boosting classifier. Despite the conscious decision to limit the number of predictors, we reached very promising results in terms of precision, recall, and F1 scores which, for the best model, were 57%, 34%, and 43% respectively. The best outcomes were obtained with the gradient boosting classifier. We give detailed information about the importance of different features, with the top three being country and region that the company operates in and the company's industry. Our model can be applied directly as a decision support system for different types of venture capital funds.

1. Introduction

The success of a business venture is a reason for founders and investors to feel proud. It is also strongly connected with a financial reward. Both founders and investors are actively looking for tools, methods, and advice that can give them an advantage over their competitors. It is debatable whether being a successful entrepreneur is associated with some intrinsic skills or whether those skills can be acquired (e.g. through formal business education). It is also very difficult to measure the significance of exogenous factors such as the industry that the company operates in, the area where the headquarters is located, or the level of competition in a particular sector and its sub-sectors.

* Corresponding author.
E-mail addresses: kamil.zbikowski@pw.edu.pl (K. Żbikowski), piotrek.antosiuk@gmail.com (P. Antosiuk).

https://doi.org/10.1016/j.ipm.2021.102555
Received 24 September 2020; Received in revised form 26 January 2021; Accepted 21 February 2021
Available online 6 March 2021

There is a long history of research that tries to determine what factors are crucial for a business to succeed. In Stuart and Abetti (1987), the authors stated that aligning an entrepreneur and the team with their technical and market experience is a key to success. They also mentioned some less trivial observations, like initial success coming more readily to firms in "more slowly growing or less dynamic markets". Another approach was to analyze the impact of both informal and formal information on the success of small and medium-sized companies in Shanghai (Vaughan, 1999). Vaughan (1999) used a linear mathematical model to examine the quantitative relationship between information and success using variables describing business environment, information usage and market situation.

Predicting business success has been an interesting challenge for management theory researchers, who analyzed the impact of management tools and theories on a company's success (Spyros Makridakis, 1996). Business research also focused on the challenges of predicting new firms' performance and the numerous, changing variables of the environment in which they operate. The alignment of the personal goals of the founder and the growth of the company was also studied by Cooper.

More recently, Huang et al. (2020) built a framework for assessing enterprise value using factors like number of patents, R&D employees, and share of owners among management. Enterprise valuation is an important metric that could be used to determine the success of a company. One of the popular milestones is the valuation of $1 billion. Privately-owned companies that reach this threshold are called *unicorns*, to signify the uniqueness of the achievement. The term was first used by venture capital investor Aileen Lee (2013) in her blog article and has since become popular in describing startups' success.

Venture capital (VC) funds invest in startups and small companies expecting the long-term growth potential, which would provide a considerable return on investment. A startup is "a human institution designed to create a new product or service under conditions of extreme uncertainty" (Ries, 2011). It is industry knowledge that 9 out of 10 startups fail. Even among the venture-backed firms, more than 75% companies fail or sustain a marginal existence (Picken, 2017). A business success prediction model could help improve the performance of VC funds. While they usually bring favorable return rates on investment, the study shows that in the 2000s, the venture capital funds underperformed compared to S&P 500 index (Harris et al., 2014). Finding businesses more likely to succeed is a challenge worth taking for venture capital funds.

2. Soft computing models for predicting business success

The machine learning approach has long been used to predict business success. Lussier (1995) used logistic regression to predict a young firm's success using the data collected through surveys on US companies. Other works explored the macroeconomic factors for predicting the success of Australian ICT companies. Tomy and Pardede (2018) successfully used *k*-nearest neighbors (*k*-NN), naive Bayes, and support vector machine (SVM) algorithms (Tomy & Pardede, 2018). However, the datasets used in those works were relatively small, having 216 and 250 instances, respectively.

Using data from Crunchbase allowed researchers to increase the size of the datasets to thousands of instances and use information about funding events in the models (Krishna et al., 2016). Xiang et al. (2012) used Crunchbase data and factual features from articles from TechCrunch to predict company acquisition using Bayesian networks Xiang et al. (2012). Bento (2018) used the Crunchbase data about startups located in the US to predict acquisition or an initial public offering (IPO) using logistic regression, SVM, and random forest algorithms. Sharchilev et al. (2018) built a gradient boosted decision tree model predicting series A funding in the next year for companies that had already acquired seed or angel funding. They used the dataset collected from Crunchbase in monthly snapshots and enriched it with data from the LinkedIn profiles of people working at the companies.

Crunchbase data was also used to predict investment behaviors modeled with graph methods. Yuxian and Yuan found that by using different link predictors like the shortest path in the graph or number of neighbors, it is possible to predict whether investors are going to invest (Yuxian & Yuan, 2013). Other alternative approaches to predicting business success include hybrid intelligence methods. Dellermann et al. proposed a framework that uses decisions made by machine learning algorithms using hard information (team size, entrepreneurial experience) and decisions made by a group of people. Both experts and non-experts would use their intuition, experience, and knowledge of the market to predict a startup's success. The results would be then aggregated to generate the classification output (Dellermann et al., 2017).

In many preceding studies that applied statistical methods, the dataset contained features that would not have been available at the time of the decision. Introducing this kind of look-ahead bias may lead to the final model not being useful in any real-world scenario. We observed such a vulnerability in the following works: Bento (2018), Dellermann et al. (2017), Krishna et al. (2016), Yuxian and Yuan (2013), and Xiang et al. (2012). It is also present in Sharchilev et al. (2018), a very thoroughly conducted study, where the authors emphasized their awareness of the potential harm that could be done through data leakage between different time periods. Nevertheless, they enriched the dataset with data gathered from Linkedin and web mentions dated long after the last sample from their original dataset. The negative impact of this operation on the performance of the final model is hard to predict.

Arroyo et al. (2019) proposed a time-aware approach in which the dataset was divided into warm-up and simulation periods. The warm-up period was only used to train models. It held the information posted in Crunchbase between the time of the company's founding and the beginning of the simulation window. The authors selected companies that had received series B and lower funding or no funding before the beginning of the simulation window and excluded companies that were acquired, had done an IPO, or received funding from Series C and above. As the median times between venture rounds A and B and between B and C are around 1.5 years, this may have led to the removal of a significant portion of successful companies from the training set. The implications of that are intuitively less problematic than making predictions based on the future data. However, it may have led the final model to be biased towards companies that have a positive value of a target variable postponed in time. A similar, but more subtle, bias can be seen when analyzing companies that started their operations at the end of the warm-up period. They simply might not have been

able to reach the requirements for assigning a positive value of a target value in a relatively short, three-year, simulation period. In our opinion, combining such an experimental setup with explicit information about current funding may result in implications on predictions that are hard to assess. Although we have not used the time-aware approach in our work, we have proposed how such a framework might look in Section 6.

The rest of the paper is organized as follows. Section 3 outlines our objectives and main contributions. In Section 4, we introduce the dataset that we use for the purpose of conducting experiments and present some statistical evaluations and insights using the data. In Section 5, we describe the experimental setup and discuss achieved results and important features. In Section 6, we summarize our work and sketch possible future research directions.

3. Objectives and contribution

In this study, we analyzed the problem of forecasting business ventures' success with the use of supervised machine learning methods. In the literature, one may find a wide range of examples that utilize modern machine learning models for that particular purpose. Those studies often employ different data sources, combining numerical features with data extracted from textual information crawled from the Internet. Such an approach is very tempting, but may easily lead to difficulties in applicability of the results. The main problem is adequately aligning the data on the timescale. Research often uses data gathered at a particular point in the past. This is rarely the point at which decision-makers would like to assess a particular venture in terms of its investment eligibility. Moreover, enriching such a dataset with data crawled at the point of starting the experiment leads to introducing the kind of bias described in the previous section. The main objective of this paper was to conduct an experiment that would lead to developing an information system that would not be flawed with the aforementioned biases – one that could be applied in practice to predict business success.

The main contributions of this paper are as follows:

1. To the best of our knowledge, this is the first study that strongly focuses on the applicability of its results by reducing the number of biases introduced in the dataset. We achieved it by purposefully limiting the set of predictors to information known at the beginning of the company's operations.
2. The dataset used for the purpose of this research is by far the largest one among similar studies. In our training set, there are 213 171 companies.
3. We provided statistical analysis and gathered insights from the dataset that might be of help for investors, policymakers, and founders. The scope of analysis includes main startup hubs as well as secondary locations and companies from multiple industries. The presented analysis benefits from the sample size in terms of its significance. We used data gathered by crawling over 700 000 websites to support our analysis and data selection process.
4. The definition of a target variable that combines information about IPO, acquisition, and subsequent funding rounds is one of the first such approaches.

4. Data collection and methodology

The research was conducted using data from the Crunchbase database (www.crunchbase.com). Crunchbase is a platform with business information about private and public companies, founders or people in leadership positions, investors, and funding rounds (Crunchbase Inc., 2020). We applied for access to the Crunchbase database for this research. After the positive response to the request, we got access to the daily snapshots of the Crunchbase database. The data used in the research and experiments was obtained on March 10, 2020. In this section, we will further describe preparing the dataset used in the training machine learning models.

4.1. Dataset from Crunchbase

The dataset provided by Crunchbase for research purposes consists of multiple tables that can be joined by unique identifiers. The simplified entity-relationship diagram (ERD) of Crunchbase tables is shown in Fig. 1.

The *organizations* table includes information about companies and investment funds. The table holds basic information such as name, HQ address, number of employees, website, social media links, email, and phone number. The summarized financial data include the number of funding rounds, the date of the last funding event, total funding, and the number of exits from investments. 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.

Additional information about funding events, acquisitions, IPOs, and investors is held in respective tables. They include data about dates of such events, amount of collected funds, and investment type (seed, angel funding, series A, B, C, etc.).

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.

Other tables, which were not used in the research, hold information about past jobs connecting organizations and people, parent organizations of those described in the *organizations* table, and industry events. There are also two tables containing descriptions of people and organizations.

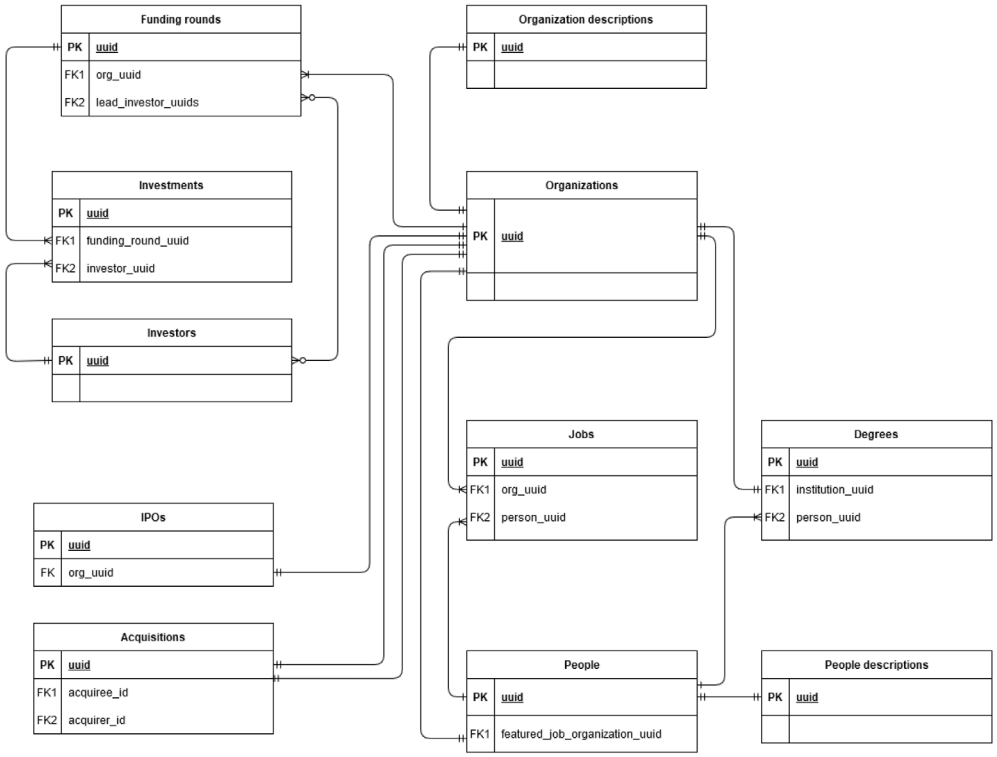


Fig. 1. Simplified ERD diagram of Crunchbase data.

4.2. Companies' homepage response

The Crunchbase dataset also features the companies' homepage URL. A homepage is a source of information about the company for investors and potential customers. An inactive homepage is a certain indicator of an organization's failure on the market. Such information could be useful in deciding whether a company is successful while creating a target variable.

We planned and conducted an experiment by crawling responses from the homepages of companies included in the *organizations* table with the status *operating*. The responses from the websites were collected using Python's `urllib` library. The crawling script collects the HTTP response code of the website or the message of the exception that was raised by `urllib` during the processing of the request. The output of the script is stored in *CSV* files. Each row is indexed by the organization's identifier in the Crunchbase dataset.

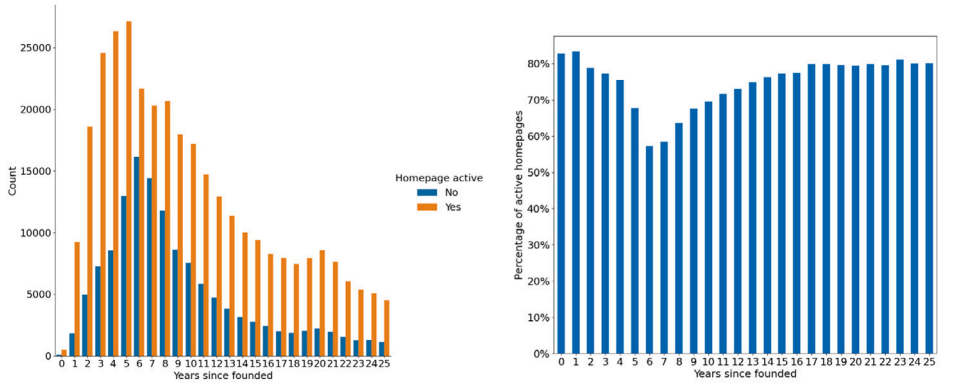
Based on data collected in the experiment, we added a new column to the *organizations* table with a flag indicating whether the homepage was active or not. Organizations with HTTP response code 200 were assigned with *1* (active) in this column; all the other organizations were assigned with *0* (inactive) (Fielding & Reschke, 2014). In Fig. 2(a), we can see the distribution of homepage activity in the years since the company's founding. The distributions follow each other with the maximum of inactive homepages for companies in the sixth year since founding and the maximum of active homepages in the fifth year.

Fig. 2(b) shows the ratio of companies with active and inactive homepages. We can see the minimum value for companies in the sixth year since the founding. After that, the ratio of companies with an active homepage rises. We can conclude that it takes five years for most companies to validate their idea on the market as shown in Fig. 2. The ratio of companies with inactive homepages reaches the minimum in the sixth year after founding. Then it returns to a value of over 70% and remains stable.

4.3. Statistical evaluation

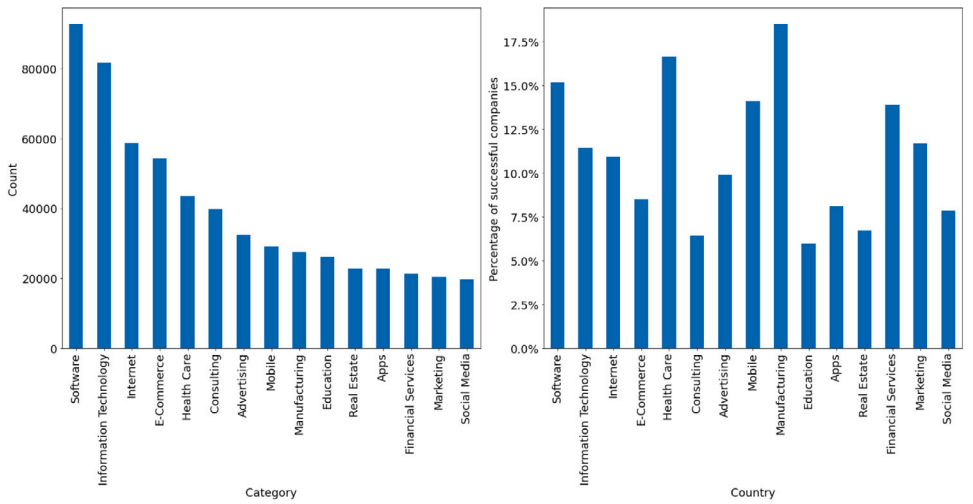
4.3.1. Most popular startup industries

In Fig. 3, we can see the fifteen most popular industries in which the analyzed companies operate. Four out of the five most popular categories are part of the information technology industry. Small and medium enterprises (SMEs) are the key part of the Internet economy because of their "high flexibility, adaptability, and inclination to implementing innovational products and business processes" (Sukhodolov et al., 2018). These industries are also a key part of the Fourth Industrial Revolution. It is set to change the world with the automation of manufacturing and services thanks to digital transformation. In the last decade, companies like Uber, AirBnB, and Amazon disrupted markets and grew at an unprecedented rate (Schwab & World Economic Forum, 2016). The example of these companies might also serve as an inspiration for entrepreneurs trying to repeat their success.



(a) Active and inactive homepage distribution (b) Ratio of companies with active homepage

Fig. 2. Active homepage distribution versus companies' age.



(a) Number of companies per category (b) Percentage of successful companies per category

Fig. 3. Most popular industry categories among analyzed companies.

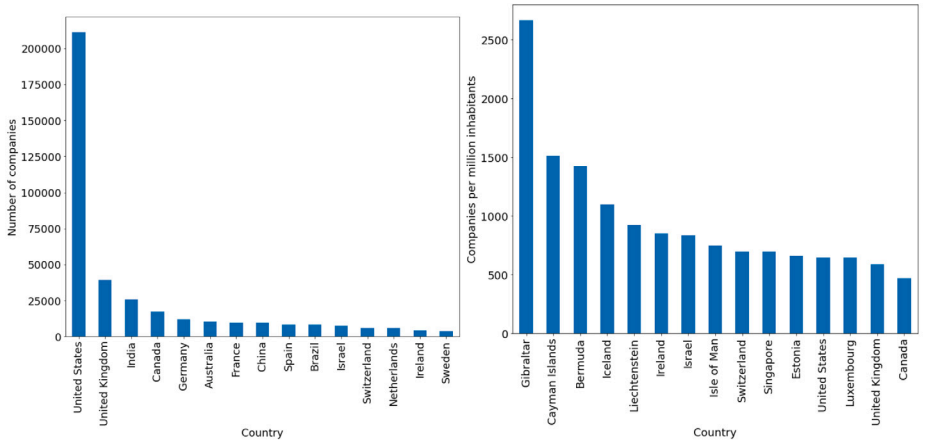
4.3.2. Top startup hubs

The United States is the country with the most startups in the dataset, as shown in Fig. 4(a). Most of the top startup hubs in 2018 (Silicon Valley, New York, Boston, and Los Angeles) are located in the United States (CB Insight, 2018). Startup hubs are located in places that meet the requirements to be a cluster of innovation. The components of a cluster of innovation include universities, enterprises, mature corporations, and large pools of private capital (Engel, 2014).

However, if we take into consideration the number of startups per million inhabitants, we can notice an almost completely different set of countries, as shown in Fig. 4(b). Some of the top countries such as Gibraltar, Cayman Islands, or Bermuda are tax havens with a favorable fiscal system. Enterprises choose to locate their financial centers in these countries for tax optimization (Palan et al., 2010).

We can also notice the presence of Estonia and Israel among countries with the most startups per million inhabitants. The former is called the Digital Republic because of the high digitalization of public services. Estonian citizens can use their identity card with an electronic chip for most administrative tasks. Programming is taught from the first grade, which might be another reason for the high number of startups per capita (Gat, 2018).

Dubbed the Startup Nation, Israel is also known for having a technology-oriented economy. It stands out as a country with a high number of startups founded by the former members of the military's cybersecurity unit 8200 (Fraiberg, 2017). Tel Aviv is also an established startup hub. While there are not as many startups operating in it compared to Silicon Valley, London, or New York,



(a) Number of companies (b) Number of companies per million inhabitants (2018)

Fig. 4. Countries with most companies in the dataset.

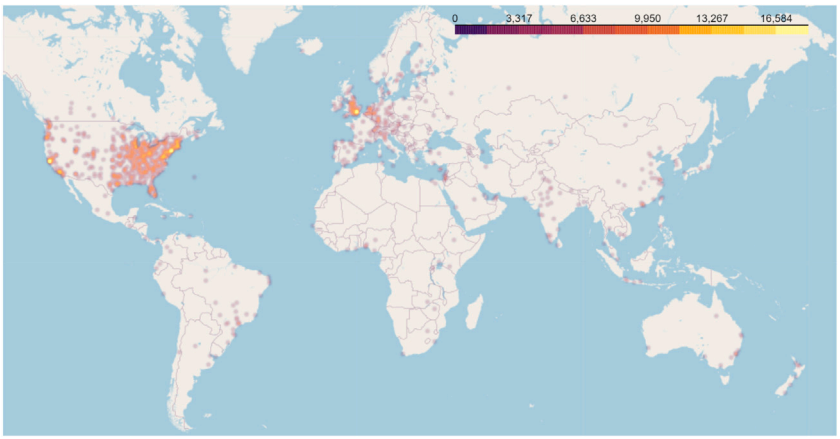


Fig. 5. Distribution and size of startup hubs in cities around the world. Map by © OpenStreetMap contributors under ODbL.

the investments provide high return. 14% of exits between 2012 and 2017 were valued over $100M. It is comparable only to Silicon Valley (CB Insight, 2018).

Fig. 5 shows the distribution of companies in the cities around the world. The size of the bubble indicates the number of companies operating in the city. We can notice a very high concentration of startup hubs in the United States. In Europe, the most prominent startup hub is London. It is the only city where the number of operating startups is comparable with hubs in the US.

Analyzed companies from Asia are concentrated in big cities like Singapore, Hong Kong, Beijing, Tokyo, and Seoul. In Europe and the US, we can notice the presence of startups operating from smaller cities. It is probable that the dataset features Asian companies that plan a global expansion or are looking for external sources of funding. In this hypothesis, their presence on Crunchbase would be to attract potential investors and customers. However, the analysis of the Crunchbase user base would be necessary to verify this hypothesis, and we do not have access to this data.

4.4. Feature engineering

4.4.1. Data selection

The dataset from Crunchbase consists of organizations founded between 1066 (HM Treasury) and 2020. Many other established organizations also have their profiles in Crunchbase and are listed as companies. Similar to previous works, we decided to select only the most recent companies that meet the definition of a startup (Bento, 2018; Krishna et al., 2016; Xiang et al., 2012). Our selected subset included companies founded between 1995 and 2015. We decided to exclude the companies founded between 2015

Table 1
List of final features in the dataset.

Feature name	Description	Type
category_list	List of organizations' subcategories	nominal
category_groups_list	List of organizations' categories	nominal
gender	Founder's gender	nominal
is_completed	Founder has completed a degree	boolean
has_multiple_degrees	Founder has more than one degree	boolean
region_org_size	Rank of region in number of startups	categorical
city_org_size	Rank of city in number of startups	categorical
years_between_graduation_and_founding	Number of years between founder's graduation and company's foundation	interval
years_of_studying	Number of years between founder's matriculation and graduation	interval

and 2020 from the dataset because they are at the initial stage of operation. We would not be able to accurately assign ground truth labels to those companies. The analysis of homepage responses presented in 4.2 in Fig. 2 shows that after five years, the number of companies with active status and an inactive homepage declines. It then becomes possible to determine successful companies. Companies founded before 1995 were mostly pioneers in the software and Internet industry. The popularity of the Internet grew in the late 1990s as the number of households in the United States owning computers rose from 22% in 1993 to 51% in 2000 (Ryan & Lewis, 2017). It led to the introduction of new business models based on the Internet (Litan & Rivlin, 2001). A lot of new companies were founded to provide their services and build their products around the Internet — new and rapidly expanding technology at that time. Many of them adopted the *.com* prefix in their name (hence *dot-com* companies). Excessive speculation in such companies led to the NASDAQ Composite Index rising 450% between January 1995 and February 2000 (MacroTrends LLC, 2020). In 2000, the *dot-com* bubble burst, which was one of the reasons for the 2001 recession in the United States (Kraay & Ventura, 2005). We selected 1995 as the lower bound of analysis to include the companies that operated during the *dot-com* bubble.

When predicting business success, we may face a similar phenomenon to the one present in measuring the performance of hedge funds (Amin & Kat, 2003). Our dataset most likely does not include companies that operated in the past but have not survived. The oldest samples are biased towards companies that persisted through difficult times (possibly through several economic downturns). In the real distribution, the fraction of successful companies should be much smaller than in the one obtained from Crunchbase. This will result in overestimating the sensitivity of trained models. The magnitude of this effect is difficult to estimate as, to our knowledge, there is no dataset publicly available that would gather information about unsuccessful business ventures. However, it is reasonable to assume that after Crunchbase gained popularity, this bias reduced over time. The company mission is to be a "master database for companies". However, even with the assumption of almost full coverage of companies that have ever existed, we have to address the problem of data not being up-to-date. In particular, this is a very common situation for companies that have an "active" status in the database while not operating anymore. This issue is addressed in more detail in the following sections, especially in 4.4.3.

4.4.2. Dataset creation

The actual dataset used in the experiment was created using information from the *organizations*, *people*, *degrees*, and *funding_rounds* tables. The rest of the tables were not used because they hold text data (*people_description* and *organization_descriptions* tables) or could result in the look-ahead bias (*investments*, *investors*, *ipos*, and *funds* tables). People and organization descriptions are usually filled by those particular people and members of those organizations.

The selected tables were joined using identifiers. The *people* and *degrees* tables were merged, and the chronologically first degree was left in the resulting table. The resulting table was then merged with the *organizations* table. The founder or the executive officer was selected in case there were many people assigned to one organization. The *funding_rounds* table was used to create the target variable. Information from the *funding_rounds* table was not used as features in the dataset.

The following features were removed from the dataset after analyzing value distribution and counting the missing values ratio: state code (*organizations* table), subject, institution name, degree type, city (*people* table), region (*people* table), country code (*people* table).

The selected features belong to three categories: nominal, interval, and binary. Many nominal attributes hold multiple unique values, which might make it difficult to build an accurate predictive model. They require additional transformation steps before encoding.

The final list of features before encoding is shown in Table 1. The number of selected features is intentionally smaller than the number in previous works that used Crunchbase data. Previous works included information known in the later stages of the company's life. An example might be information about funding events or Venture Capital (VC) backing (Bento, 2018; Krishna et al., 2016; Xiang et al., 2012). We decided to leave only information that would have been known at the beginning of the company's operation. We believe that this approach makes instances comparable and reduces bias against younger organizations.

Some of the features would also be subject to one-hot encoding, which increases the number of columns. See 4.4.7 for details on attribute encoding. The dataset used in model training holds 213171 instances, which are described by 244 features. Each instance is also assigned the value of the target variable.

4.4.3. Target variable

Defining the target variable requires defining what success in a business venture means. Surviving in the market is not enough to label a company as successful. Startups often rely on external funding to grow their operations, which is also a sign of investor confidence in the company's potential. Business funding is usually split into rounds. Each subsequent round is usually larger than the previous one, and engaged investors become more institutionalized. Receiving each funding series means reaching a certain milestone for the company. Reaching a certain milestone in funding was a success metric in some of the past works (Sharchilev et al., 2018; Spiegel et al., 2016).

We decided to choose completing series B as a milestone for labeling the company as successful. Completing a series B round usually means that the startup proved that it had a stable user base that generates profit. The median of the raised amount in series B among the analyzed companies is $11.2M compared to $5M raised in series A. Receiving series B means going through the selection process of the investment funds twice, which is a strong indicator of a company's success.

Another popular startup success metric is undertaking an IPO. Completing an IPO means that the company proved to be interesting for investors in the stock exchange market. Studies show that while VC-backed companies have a lower mortality rate in the pre-IPO stage, the post-IPO performance of companies with VC funding and without it is similar (Ber & Yafeh, 2007). Undertaking an IPO is a sign that the company is well established in the market and very likely to survive.

The acquisition of the company is another indication of its value on the market. The acquirer takes control of the idea, product, and employees of the acquired company. Being acquired might be considered as a growth strategy for some startups. For example, studies show that university spin-offs are more likely to be acquired than to go public. The reason for this might be that academics are more focused on the technical part of the product and less on growing the company. A more established company might fill these needs while acquiring expertise. The acquisition might be considered a "win-win" strategy for both parties (Bonardo et al., 2010).

The target variable is binary, with the positive class encompassing companies that proved to be successful and the negative class grouping all the other ones. The criteria for the positive class are meant to single out with certainty the instances that are successful. Undertaking an IPO means that the company went through the preparation process, which includes issuing a prospectus with the publicly available valuation. The acquisition of the company is also a certain indicator of success, as the founders receive the return on the investment they made developing the company. We also label the companies that received series B funding and are still operating in the market as successful. We cannot determine that they will not fail in the future, but receiving series B funding is a sign of trust from the investors and a significant milestone for the company. There is high certainty that the companies that we assign to the positive class are successful in the market. The target variable is defined as follows:

$$y = \begin{cases} 1, & \text{if } x_{status} = \text{"acquired"} \vee x_{status} = \text{"ipo"} \\ & \vee (x_{status} = \text{"operating"} \wedge x_{investment\_type} = \text{"series\_b"}) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

87.8% of the instances in the dataset belong to the negative class, and 12.2% belong to the positive class. This imbalanced distribution is due to the nature of the problem: most founded companies do not succeed in the market.

Fig. 6 shows the percentage of instances belonging to the positive class in fifteen countries with the highest number of organizations in the dataset. We can notice China standing out, with over 30% of companies in the positive class. On the other hand, India, Spain, and Brazil have the lowest percentage of successful companies in the analyzed group of countries.

4.4.4. Grouping values below frequency threshold

Based on the frequency distribution over values, we selected a certain threshold that would cut off the least frequent ones. This approach was applied to the following columns: country code, category list, and gender. Values in the gender column were grouped into three labels: male, female, and other.

The *other* label groups the country codes with a frequency below the median. We selected the median (53.5) for the threshold because it kept country codes of smaller technologically advanced nations like Estonia (875 instances) and Iceland (388 instances) in the dataset.

The category list column is a multi-valued field representing the industries in which the company operates. The field was split into multiple rows to count values and group those below a threshold. The column holds many unique values, but most of them appear rarely. The mean count of instances for category value is 2154.9, but the median is equal to 407 and the upper quartile starts from 1232.5. Values below the upper quartile were grouped in the *other* label.

4.4.5. Binning values based on frequency

The dataset holds numerous cities and regions in which the startups operate. The number of unique values in those columns is, respectively, 21159 and 1755. The cutoff value would have to be very high to lower the number of unique values to a processable value. However, the frequency distribution over values shows that grouping them in tiers could provide information to the model while reducing the complexity.

The values were binned based on their frequency in the dataset into five equal-width bins. This resulted in lowly populated top bins but separated the top startup hubs from the rest. For example, California is in a separate bin to New York or England, which reflects its unique place among startup hubs.

The equal-depth binning approach was also tried, but the result did not provide the desired separation between top hubs and the rest. We believe that using equal-depth binning would result in the loss of information.

The city and region features were replaced with ordinal attributes of city size and region size based on the binning described above.

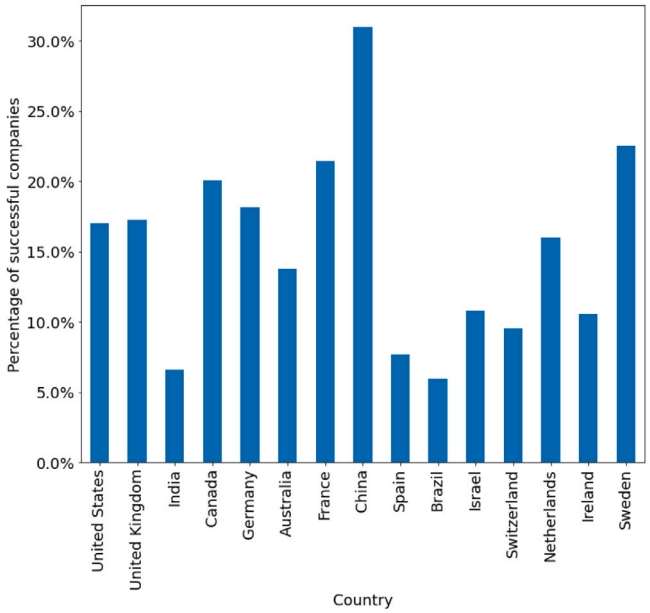


Fig. 6. Percentage of instances in positive class in countries with most instances in the dataset.

4.4.6. Transforming dates into time ranges

Providing dates explicitly to the model could result in introducing a bias. For example, the longer a company operates, the higher the chance of being successful on the market. Using the date of company foundation in the dataset would result in a model favoring older organizations when predicting success. Our initial experiments confirmed that dates should be transformed into relative time ranges in order to avoid this scenario.

The year of the company's foundation was replaced with the number of years between the founder's graduation and the company's foundation. The new feature provides explicit information about the founder's professional experience when founding the company. Dates of the founder's matriculation and graduation were replaced with the number of years spent at the university. The missing values were filled with 0.

4.4.7. Encoding attributes

After reducing the number of unique values in the dataset, the resulting set of attributes should be encoded before using it as a dataset in model training. Gender, category, and category group attributes were transformed using one-hot encoding. Region size, city size, country code, and boolean flags were transformed using ordinal encoding with missing values as a separate label.

5. Experiments

5.1. Supervised learning

Defining the target variable company's success with two possible values: 1 (successful) or 0 (unsuccessful) reduces the problem of predicting business success to binary classification. This problem can be solved using a supervised learning approach. In supervised learning, the dataset is split into a subset of features (matrix $\boldsymbol{X}$) and the target variable (vector $\boldsymbol{y}$). In the training process, models try to predict the value of y when given a set of features x. The predicted value $\hat{y}$ is compared with the ground truth y. The model's parameters are changed in the iterative process to maximize the number of correct predictions.

The binary classification can be solved with multiple algorithms as it is a very well covered problem in machine learning. We decided to select three simple models whose implementation is available in popular machine learning packages for Python like *scikit-learn* and *XGBoost*. These models were logistic regression, Support Vector Machine (SVM), and XGBoost. Logistic regression and SVM are established algorithms used in many previous works on Crunchbase data (Bento, 2018; Krishna et al., 2016; Xiang et al., 2012). The XGBoost algorithm, which uses decision trees and boosting techniques, has recently gained popularity because of its performance in Kaggle competitions (Kaggle Inc., 2019).

Table 2
Results of 10-fold cross validation on the training set. Precision, recall, and F1 score are reported for the positive class.

Algorithm	Accuracy	Precision	Recall	F1
Logistic regression	0.86	0.70	0.21	0.33
SVM (RBF kernel)	0.87	0.86	0.20	0.32
XGBoost	0.86	0.90	0.17	0.28

5.2. Train/test split

The dataset was split into train and test subsets in a stratified fashion after shuffling (Sechidis et al., 2011). The stratified split guarantees that the distribution of the classes of the target variable is the same in both training and test subsets. The test set is then a representative sample of the dataset in terms of target variable. The size of the test set is fixed and equal to 10,000. It is around 5% of instances in the dataset. The final results of the model's performance were reported on the test set. We used the test set with fixed size rather than the popular 80–20 split to provide more instances in the training process.

We did not use a separate validation test but performed cross-validation on the training set to make sure that models do not overfit to the validation set. Cross-validation is recommended in hyperparameter tuning to reduce the problem of selection bias and overfitting (Cawley & Talbot, 2010). The split in each fold of cross-validation was stratified – the distribution of the target variable was the same in a validation subset and the rest of the training set used in the fold.

5.3. Experiment setup

Fig. 7 shows the experiment schema of model selection process. The dataset was split as described earlier into the training and test sets. Data was preprocessed using feature scaling methods in experiments with logistic regression (minmax normalization) and SVM (standardization). Feature scaling used with these algorithms improves the performance of the algorithms and reduces the time an algorithm takes to converge. Scaling inputs is necessary with logistic regression because the algorithm uses lasso and ridge regularization (see 5.5.1 for explanation and results) (Hastie et al., 2013a). We used minmax normalization, which maps feature values to the [0, 1] range. SVM training was performed with data standardization as a preprocessing step. Standardization changes the distribution of each feature to have zero-mean and unit-variance. The standardization of features is said to reduce the error in SVM classification (Juszczak et al., 2002). Experiments with the *XGBoost* algorithm did not require the scaling of variables. This algorithm uses classification tree internally as a base model. The classification tree looks for the optimal split of instances based on the selected variable and value. This algorithm is invariant to feature scaling (Hastie et al., 2013b).

The models were selected in the hyperparameter tuning process, which is described in detail in Section 5.5. The best performing models were then trained on the entire training set and tested on the test set. The reported results come from the test set, which was not used in the model selection process to provide an objective benchmark for the models' performance.

5.4. Initial results

The first experiments used the default set of parameters of *scikit-learn* and *XGBoost* packages. The classifiers were trained using 10-fold cross-validations. Accuracy, precision, recall, and F1 score (the harmonic mean of precision and recall) were measured on the validation subset in each fold. The metrics presented in Table 2 are means of values measured in 10-fold cross-validation for the positive class. All of the classifiers scored high accuracy values close to 90%. SVM and XGBoost algorithms stood out for high precision also close to 90%. The logistic regression classifier performed best in terms of the F1 score. However, the recall metric for all three classifiers was very low. This means that models misclassified 70% of successful startups as unsuccessful. We would like the algorithm to discover more successful startups and not only those which follow the most popular patterns.

5.5. Hyperparameter tuning

Initial results showed that models could improve performance in terms of recall. The hyperparameter tuning process aims to find the optimal set of the models' parameters. Searching the entire space of possible values for parameters is computationally expensive, so we used two methods to find and test different sets of parameters of the models: grid search and randomized search. We selected values of parameters for experiments in both of those methods. An exhaustive grid search was applied for logistic regression and SVM because of the limited number of parameters to tune. It was possible for us to test all combinations of the selected values. We tuned the XGBoost model's performance using a randomized search because of the high number of parameters. An exhaustive search was not possible in this case because of computational limits.

In both methods, the performance of different variations of models was measured using 5-fold cross-validation. The validated models were ranked according to the value of the F1 score. Maximizing the F1 score balances the trade-off between precision and recall in the classifiers' performance.

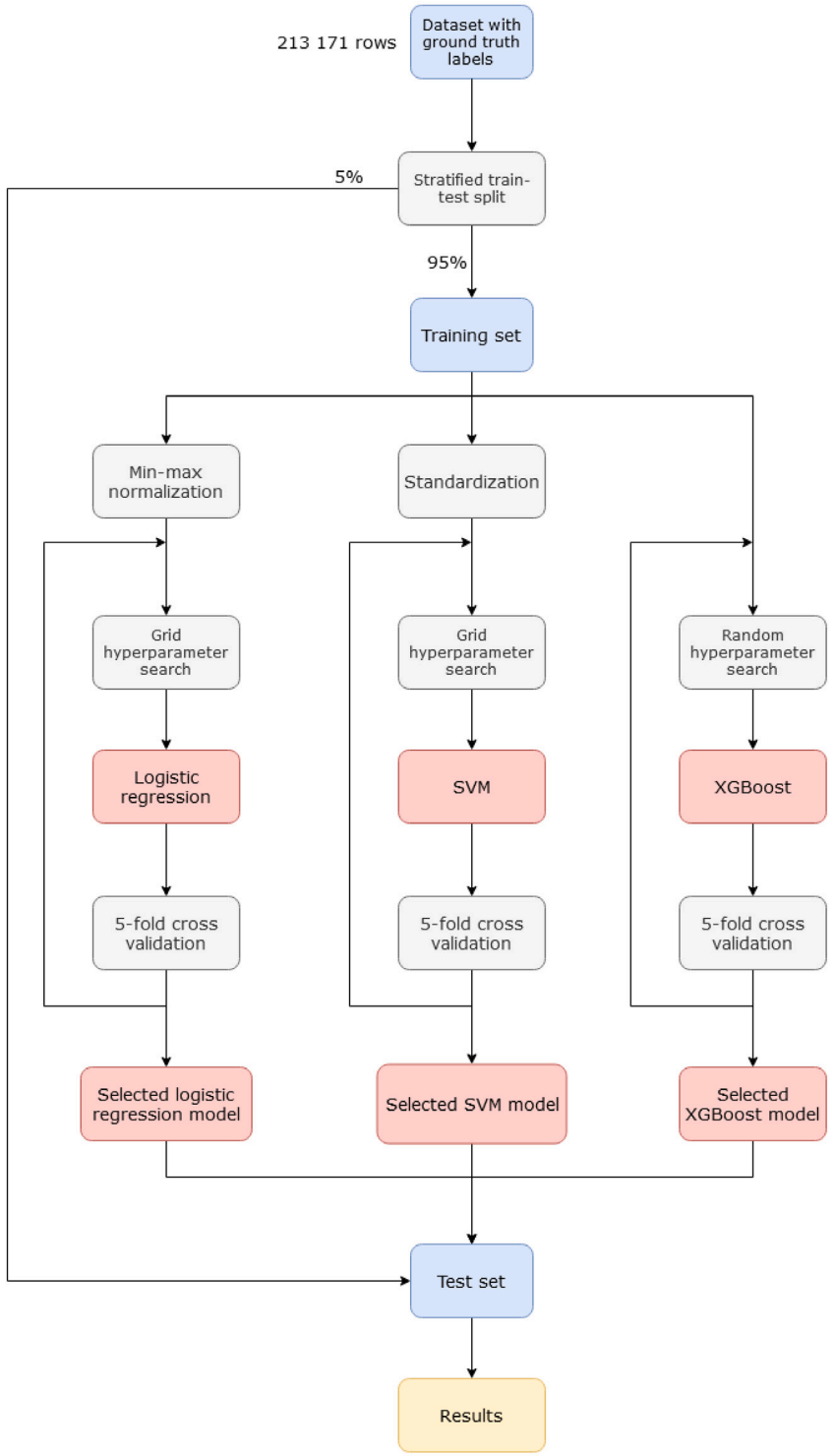


Fig. 7. Experiment setup.

5.5.1. Logistic regression

The logistic regression algorithm usually overfits training data when the values of coefficients values are too high (Jurafsky & Martin, 2014). The coefficients are adjusted in each training step to minimize the cost function of the algorithm. We used the mean

squared error (MSE) as a cost function in logistic regression. Regularization is a technique used to stop models from overfitting by adding a term to the cost function that is proportional to values of coefficients. The regularization term might be proportional to the sum of magnitudes of all the coefficients (L1 or Lasso regularization). Let $\hat{y}_i$ be the prediction of the *i-th* instance, y_i the ground truth value of the *i-th* instance, n the number of instances, β_j the value of the jth coefficient, and p the number of coefficients. The strength of regularization is controlled by the C parameter. The formula for the cost function is then equal to:

$$\sum_{i}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{C}\sum_{j}^{p}\|\beta_j\| \tag{2}$$

Another option is introducing a regularization term that is proportional to the sum of squared coefficients in the model (L2 or ridge regression):

$$\sum_{i}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{C}\sum_{j}^{p}\beta_j^2 \tag{3}$$

Both options were tested in the experiments along different values of the C parameter. Note that smaller values of C enforce stronger regularization. The following set of parameters was used in the experiments:

- penalty: L1, L2
- C: 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100

5.5.2. Support vector machine

SVM is, similarly to logistic regression, a linear classifier. The algorithm looks for the optimal decision hyperplane to separate classes. The C parameter of the SVM algorithm controls the regularization of the decision function. Its value is inversely proportional to the strength of regularization (Pedregosa et al., 2011). Lower C forces a higher margin of the classifier's decision function and a simpler decision surface. This might lead to a model underfitting the data. A model with higher values of the C parameter tends to classify all training samples correctly. The side effect of this might be overfitting training data and poor generalization on the test set (Cortes & Vapnik, 1995).

SVM can be used as a non-linear classifier with the use of special functions, called kernels (Smola & Schölkopf, 1998). One of them is the radial basis function kernel, which is used in this study. The parameter responsible for controlling its behavior is γ. High values of γ force the model to create a decision surface based on instances that are located close to each other. Low values of γ allow more distantly located instances to influence the decision surface (Pedregosa et al., 2011). We chose two values suggested in the scikit-learn's application programming interface (API) because of the computational complexity of SVM's hyperparameter tuning:

- γ: $\frac{1}{n_{features}}$, $\frac{1}{n_{features} * Var(X)}$
- C: 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100

5.5.3. XGBoost

We used the XGBoost algorithm with CART as the base model (Chen & Guestrin, 2016). We decided to tune the number of estimators used in the ensemble. Intuitively, the more trees in the ensemble, the more options are explored, because each tree uses only a subset of features. The ratio of used features to total number of features is controlled by the *colsample bytree* parameter. We can also control how many levels are allowed when building a tree by controlling its maximum depth. Deeper trees tend to overfit the data. The minimum child weight parameter sets the minimum value for the sum of the instance's weight to perform a split. Higher values might stop building a tree earlier and prevent overfitting. The gamma parameter is another way of controlling whether there should be a split. It sets the minimum value of the loss reduction. The learning rate parameter controls the impact of further boosting steps to prevent overfitting (xgboost developers, 2020).

XGBoost offers more parameters to tune than the other used algorithms. A full grid search of all combinations was not possible due to computational constraints. The randomized search method allows some of the combinations to be checked. While this method does not guarantee finding the optimal set of parameters, it is said to find models that might be as good as those found using a grid search (Bergstra & Bengio, 2012). 25 randomly chosen sets of the parameters described below were tested to find the best performing model:

- number of estimators: 100, 250, 500, 750, 1000
- maximum depth of a tree: 3, 5, 7, 10, 12, 15, 17, 20, 25
- minimum child weight: 1, 3, 5, 7
- gamma: 0.0, 0.1, 0.2, 0.3, 0.4
- colsample bytree: 0.3, 0.4, 0.5, 0.7
- learning rate: 0.05, 0.10, 0.15, 0.20, 0.25, 0.30

A randomized search helped to find the best performing set of parameters out of the tested combinations. The selected algorithm had the following parameters: *colsample bytree*=0.5, *gamma*=0.1, *learning rate*=0.25, *maximum depth*=25, *minimum child weight*=5, *number of estimators*=750. The full results of hyperparameter tuning showed that performance could be further improved by increasing the number of estimators and depth of trees. Using the model found in the random search procedure, the following set of parameters was tested in an exhaustive grid search:

Table 3
Results of 5-fold cross validation on training set of selected models with highest F1 score after hyperparameter tuning. Precision, recall, and F1 score are reported for the positive class.

Algorithm	Accuracy	Precision	Recall	F1
Logistic regression	0.86	0.69	0.22	0.33
SVM (RBF kernel)	0.84	0.52	0.33	0.40
XGBoost	0.86	0.60	0.33	0.43

Table 4
Results on test set of models selected during hyperparameter tuning. Precision, recall, and F1 score are reported for the positive class.

Algorithm	Accuracy	Precision	Recall	F1
Logistic regression	0.86	0.67	0.21	0.32
SVM (RBF kernel)	0.84	0.49	0.31	0.38
XGBoost	0.85	0.57	0.34	0.43

- number of estimators: 1000, 1250, 1500
- maximum depth of a tree: 20, 30, 35

The results were slightly better, which led to choosing the model with 1250 trees of 30 levels of maximum depth.

5.5.4. Results of cross-validation on the training set

The models were selected in the cross-validation process based on the performance on the validation subsets used in each fold. The results of models with the highest F1 scores are presented in Table 3.

The highest score in the logistic regression algorithm came from the model with the L2 penalty and a C value equal to 100. The results are close to those from the model with default parameters (see Table 2). The improvement in F1 score is not significant. The logistic regression model does not offer more parameters to tune its performance, but it is simple and offers a baseline for more complex models.

The SVM model with a C value of 100 and γ equal to $\frac{1}{n_{features}}$ scored the highest F1 score. However, the increase of the F1 score and recall came at the expense of precision and accuracy, as shown in Table 3. The drop of precision and accuracy is higher than for other algorithms. The SVM model created a decision function that fits the training data better than the default one, but it is not as precise in classification (see Table 2).

Increasing depth and number of trees in the ensemble gave the best results for the XGBoost classifier.

5.5.5. Verification on the test set

We then tested models selected in the cross-validation process on the test set of 10,000 representative samples. Table 4 shows the results of the selected models on the test set. We can notice an only slight decrease in performance compared to the results from cross-validation in Table 3. The difference is relatively small, which shows the robustness of the models. The models did not overfit training data in the process of model validation and hyperparameter tuning.

The final results showed a decrease in precision of all tested models at the cost of increased recall. The logistic regression was the least affected in the process. Performance before and after hyperparameter tuning is similar. The SVM algorithm had the largest decrease in the precision metric among the classifiers while achieving a similar recall to XGBoost. The XGBoost algorithm managed to increase the recall while keeping precision above 50%. It also performed best among classifiers when the F1 score is considered.

In the hyperparameter tuning process, we managed to increase the recall and F1 score of classifiers. As a result, the classifiers were able to find more successful startups among instances previously misclassified as unsuccessful. It came at the price of reduced precision, which means that the classifiers were more likely to misclassify an unsuccessful startup as successful.

5.5.6. Feature importance

As the XGBoost model uses a classification tree as a base estimator, we can look into the decision process of the classifier. While it is possible to plot a single decision tree, it would be difficult to draw conclusions from 1250 trees in the ensemble. However, we can use the aggregated data of the ensemble to visualize factors that impacted the decision-making process. In Fig. 8, we can see the 10 most important features in the best performing XGBoost model trained on the entire training set. The F-score is the number of times a feature appears in a tree.

Based on Fig. 8, we are not able to determine the actual values of the features that led to the classification decisions. We know which features were used most frequently to split instances in leaves when building a tree, and this measure is used as a features' importance estimator. The geographical features (country code, region size, and city size) were among the most important ones. As we showed in 4.3.2, governments try to build startup ecosystems that aim to increase the number of successful startups in the global market.

We can also notice that trees were often split by industry. The split was performed on the industries from the extremes of popularity. Software and Internet services were the most popular industries among companies in the dataset, as we can see in

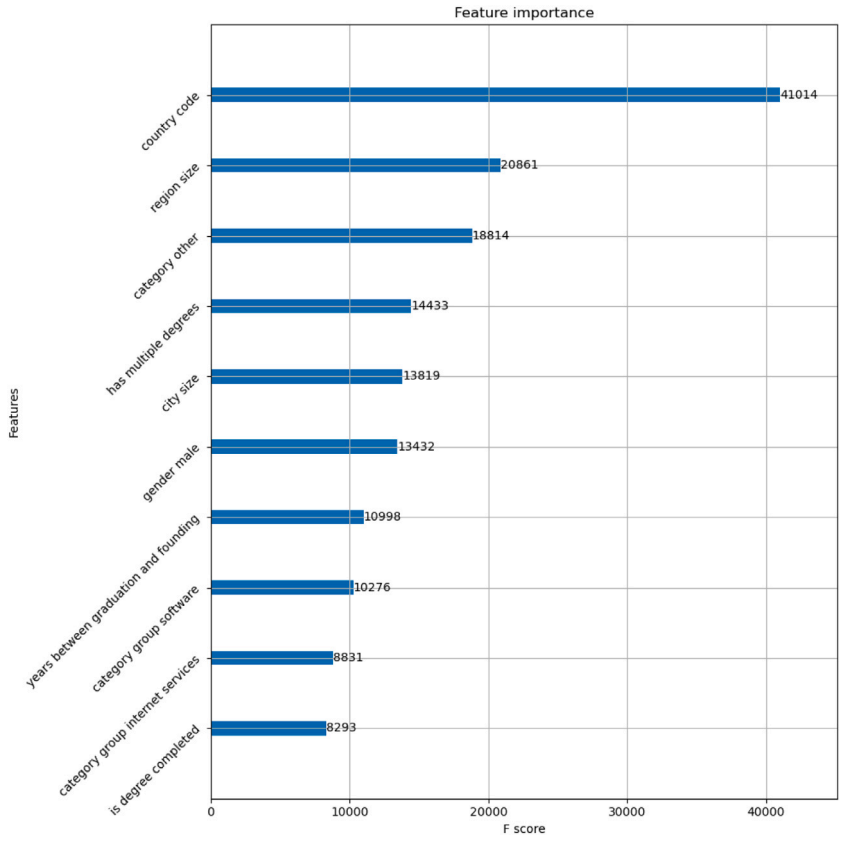


Fig. 8. 10 most important features in the selected XGBoost model. F-score is the number of times a feature appears in a tree.

Fig. 3(a), while the Other category grouped least frequent industries. It is interesting that important features did not include the industries that have the highest percentage of successful companies, like healthcare or manufacturing (see Fig. 3(b)).

Information about the founder's education also played a key role in building the trees. Features like multiple degrees, completing a degree, or the number of years between graduation and founding shows that getting or not getting a degree by the founder is an important decision-making factor. We can also notice that the founder's gender is one of the most important features. It would be necessary to make sure that the classifier is not biased against any group of people if it were used to support the decision-making process (Garfinkel et al., 2017).

5.6. Comparison with previous work

Previous works on predicting business success used different definitions of success. Some of the definitions partially cover the conditions we used for the positive class, which makes it difficult to directly compare the results. As our work uses fewer predictors than previous studies, we did not expect to outrank the performance of their models. However, we can compare our results to see the effect of reducing the variables to the set, which would be applicable in real-world situations.

Xiang et al. (2012) used data from Crunchbase such as the location of the company, number of products, as well as information about received funding and past managerial experience. The authors enriched the dataset with analysis of news articles on TechCrunch – a portal with information about startups and investments. The model used the Bayesian network to predict the acquisition of a company, which is one of the conditions we used to define the positive class. The model's performed recall was between 56.5% and 59.9% for companies with no matching articles on TechCrunch.

In Krishna et al. (2016), several models using data from different rounds of financing were used along with custom factors ranking the product-market fit, uniqueness of the idea, or go-to-market strategy based on authors analyzing the business information available about a company on websites like Forbes or TechCrunch. For the model not including the seed fund amounts, Krishna et al. (2016) received 78.3% recall with 73.3% precision. This approach uses hard information from Crunchbase with custom features based on the authors' assessment of a company. Using human input might bring bias to the model because the criteria for company assessment are not clearly stated.

The results of the study performed by Arroyo et al. (2019) are comparable to ours. However, as we discussed in 2, the approach presented in that work is still not free of bias. The first approach of the authors was to perform a multi-class classification. The global

accuracy results range from 74.6% for the decision trees classifier to 82.2% for gradient tree boosting classifier. We were able to achieve accuracy between 84% for SVM and 86% for the logistic regression classifier. However, accuracy is not an appropriate metric for problems with imbalanced classes. Arroyo et al. (2019) also presented aggregated results from positive and negative classes. The aggregated precision for positive classes was between 45% and 68%. The results are similar to the precision values for the positive class in our work, which ranged from 49% to 67% (see Table 4). Our results are better in terms of accuracy and comparable in terms of precision, while some of the biases discussed in Section 2 were successfully avoided.

5.7. Experiment implications

We confirmed the significant impact of the look-ahead bias, which is introduced by including the variables not known at the time of the company assessment. Our results are lower than those presented in previous works, which shows how much the models might be affected by overlooking the look-ahead bias. We believe that future experiments should be designed in any way that would use only the data known at the time of the decision.

The practical implication of the experiments is the feasibility of a bias-free approach in predicting business success. Reducing the number of features to those available in the early stage of a startup's life makes the models applicable in real-world situations. We provided an approach that reduces the look-ahead bias without the need to gather historical data. Although we purposefully limited the set of predictors, the models' performance remained high enough for them to be applicable in practice.

The models could support investment decisions in venture capital funds by assessing the company's chances of succeeding in the market. As shown earlier, predicting business success is a crucial challenge for VC funds, which affects their performance and return rates. Our model can serve as a decision-support system in an investment fund. The trained model provides a classification decision about whether the company will undertake an IPO, become acquired, or complete a series B round. The dataset used for model training consists of more than 200 thousands of observations of real-world companies.

The general approach of the proposed framework applies to other similar datasets. In particular, a VC fund with its database of companies could build the bias-free model following the guidelines presented in this paper. Funds that do not have their data management systems or databases could use the provided models and Crunchbase dataset as a starting point for creating such systems.

Most of the VC funds do not keep track of such a large number of companies as Crunchbase. The venture capital institutions usually focus their analysis on a smaller part of the market. Our work is based on the data from companies located in hubs from the whole world and founded over 30 years, offering a broad approach. We believe that combining these two approaches in VC funds decision support systems could potentially increase their performance.

Our research could also be of use for potential founders looking for information about conditions in which companies thrive. We believe that the findings of our work may help to make them an informed decision. We cannot measure the impact of the financial or managerial factors on the business' success. However, we believe that conclusions based on the analyzed factors would help founders identify opportunities to increase chances for success.

6. Conclusions

Predicting business success is a challenging task, but it is crucial to many public and private stakeholders who shape economics, make funding and investment decisions, and found companies. Intuitively, the task becomes easier as the company matures, tests its product-market fit, and goes through the selection processes of angel investors and VC funds.

We proposed a machine learning approach for predicting business success at the early stage, narrowing down the set of features to geographical, demographic, and basic information about the companies. Unlike previous works, we did not use any information about external funding even if it was available. The main advantage of the research is the reproducibility of its outcomes for real-world scenarios. We achieved it by carefully and consequently reducing the dataset to factors that would only be available at the decision time. We believe that such a decision-support system can be useful in venture capital funds to help find potentially successful companies that otherwise could be unnoticed. However, the presented approach does not take into account the most recent data for prediction purposes. One could argue that companies founded recently are very different in terms of patterns that led to their success. The solution here is straightforward and would require reducing the dataset to the companies created more recently or assigning them larger weights in the learning process.

We analyzed selected data from the Crunchbase database and proved how the country's fiscal policy, investment in innovation, and its creation of a system supporting the education of new technologies increase the number of startups per million inhabitants. The examples of Estonia and Israel could serve as inspiration for policymakers who would like to increase the impact of the new technologies sector in the economies. We also showed that the most popular startup industries, such as software and Internet services, are not the ones with the highest share of successful companies. Creating a company operating in healthcare or manufacturing might require more significant upfront resources compared to the software industry, but the chances of succeeding are the highest among the most popular industries. However, it is possible that a relatively low entrance barrier may be the reason for the lower success rate among software companies compared to other industries. Another way of looking at this problem is analyzing the competition in particular industries. In this sense, our research confirms the hypothesis formulated by Stuart and Abetti (1987) about the superiority of slowly growing and less dynamic markets for initial business success.

To predict business success, we built machine learning models and compared the performance of three algorithms: logistic regression, SVM, and XGBoost. To the best of our knowledge, this is the first study on Crunchbase data using the XGBoost algorithm,

Table A.5
Columns in *organizations* table.

Name	Type
uuid	string
name	string
type	string
permalink	string
cb_url	string
rank	float
created_at	timestamp
updated_at	timestamp
legal_name	string
roles	string
domain	string
homepage_url	string
country_code	string
state_code	string
region	string
city	string
address	string
postal_code	string
status	string
short_description	string
category_list	string
category_groups_list	string
num_funding_rounds	float
total_funding_usd	float
total_funding	float
total_funding_currency_code	string
founded_on	string
last_funding_on	date
closed_on	date
employee_count	string
email	string
phone	string
facebook_url	string
linkedin_url	string
twitter_url	string
logo_url	string
alias1	string
alias2	string
alias3	string
primary_role	string
num_exits	float

which recently gained popularity because of high performance in machine learning competitions. We were able to increase the recall and F1 score compared to initial experiments by tuning hyperparameters. The trained models are robust and generalize well, as can be seen when comparing the results of the cross-validation and the test set. The XGBoost model scored the highest values of precision, recall, and F1 score out of the tested algorithms, which, for the best model, were equal to 57%, 34%, and 43%, respectively. The results prove the XGBoost algorithm's usefulness in future applications of predicting business success.

All models performed with higher precision than recall. This means that they were able to find only the organizations that fit the most popular patterns of success. Analysis of feature importance showed that the best performing XGBoost model used a startup's location and operation in industries such as software or Internet services frequently while building decision trees.

Future works could increase the recall of the models by enriching the dataset. More detailed data about the founder's prior experience and the company's product or service could improve the performance of models. This would require adding other sources of data about companies and their founders than Crunchbase. Although we decided not to use the descriptions of people and companies because they were not objective, the text data could be explored as an additional source of features for the dataset. Another approach would be to gather snapshots of the Crunchbase database at regular intervals. Such a dataset would provide potentially valuable information about the dynamics of the company's growth. It could be modeled using time-series techniques. We believe that our work gives more insight into the startup industry and provides robust machine learning models for predicting business success.

CRediT authorship contribution statement

Kamil Żbikowski: Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing, Supervision. **Piotr Antosiuk:** Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Supervision.

Table A.6
Columns in *people* table.

Name	Type
uuid	string
name	string
type	string
permalink	string
cb_url	string
rank	float
created_at	timestamp
updated_at	timestamp
first_name	string
last_name	string
gender	string
country_code	string
state_code	string
region	string
city	string
featured_job_organization_uuid	string
featured_job_organization_name	string
featured_job_title	string
facebook_url	string
linkedin_url	string
twitter_url	string
logo_url	string

Table A.7
Columns in *degrees* table.

Name	Type
uuid	string
name	string
type	string
permalink	float
cb_url	float
rank	float
created_at	timestamp
updated_at	timestamp
person_uuid	string
person_name	string
institution_uuid	string
institution_name	string
degree_type	string
subject	string
started_on	date
completed_on	date
is_completed	bool

Appendix. Tables used to build the dataset

See Tables A.5–A.7.

References

Amin, G. S., & Kat, H. M. (2003). Welcome to the dark side: Hedge fund attrition and survivorship bias over the period 1994–2001. *The Journal of Alternative Investments*, *6*(1), 57–73.

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.

Bento, F. R. d. S. R. (2018). Predicting start-up success with machine learning. Universidade Nova de Lisboa.

Ber, H., & Yafeh, Y. (2007). Can venture capital funds pick winners? Evidence from pre-IPO survival rates and post-IPO performance. *Israel Economic Review*, *5*(1), 23–46.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bonardo, D., Paleari, S., & Vismara, S. (2010). The M&A dynamics of European science-based entrepreneurial firms. *The Journal of Technology Transfer*, *35*(1), 141–180.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

CB Insight (2018). Global tech hubs report. https://www.cbinsights.com/research/report/global-tech-hubs/ (Accessed 05 July 2020).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. (pp. 785–794).

Cooper, A. C. (1993). Challenges in predicting new firm performance. *Journal of Business Venturing*, *8*(3), 241–253. http://dx.doi.org/10.1016/0883-9026(93)90030-9, http://www.sciencedirect.com/science/article/pii/0883902693900309 Special Theoretical Issue.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Crunchbase Inc. (2020). Crunchbase profile. https://www.crunchbase.com/organization/crunchbase (Accessed 04 July 2020).

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2017). Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. In *International Conference on Information Systems*.

xgboost developers (2020). Xgboost parameters. https://xgboost.readthedocs.io/en/latest/parameter.html (Accessed 26 July 2020).

Engel, J. (2014). Global clusters of innovation: Entrepreneurial engines of economic growth around the world. Edward Elgar.

Fielding, R., & Reschke, J. (2014). Hypertext transfer protocol (HTTP/1.1): semantics and content. *RFC 7231*, RFC Editor, http://www.rfc-editor.org/rfc/rfc7231.txt (Accessed on: 04 July 2020).

Fraiberg, S. (2017). Start-up nation: Studying transnational entrepreneurial practices in Israel's start-up ecosystem. *Journal of Business and Technical Communication*, *31*(3), 350–388, https://doi.org/10.1177/1050651917695541.

Garfinkel, S., Matthews, J., Shapiro, S. S., & Smith, J. M. (2017). Toward algorithmic transparency and accountability. *Communications of the ACM*, *60*(9), 5–5.

Gat, O. (2018). Estonia goes digital: Residents of the tiny baltic nation are going all in on techno-governance. *World Policy Journal*, *35*(1), 108–113.

Harris, R. S., Jenkinson, T., & Kaplan, S. N. (2014). Private equity performance: What do we know?. *The Journal of Finance*, *69*(5), 1851–1882.

Hastie, T., Tibshirani, R., & Friedman, J. (2013a). *Springer series in statistics*, *The elements of statistical learning: data mining, inference, and prediction* (pp. 61–79). Springer New York.

Hastie, T., Tibshirani, R., & Friedman, J. (2013b). The elements of statistical learning: Data mining, inference, and prediction. In *Springer series in statistics*, (pp. 305–317). Springer New York.

Huang, W.-B., Liu, J., Bai, H., & Zhang, P. (2020). Value assessment of companies by using an enterprise value assessment system based on their public transfer specification. *Information Processing & Management*, *57*(5), Article 102254. http://dx.doi.org/10.1016/j.ipm.2020.102254, http://www.sciencedirect.com/science/article/pii/S0306457319307976.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing. vol. 3*. Pearson London London.

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

Kaggle Inc. (2019). Kaggle's state of data science and machine learning 2019. https://www.docdroid.com/qzyxCr4/kaggle-state-of-data-science-and-machine-learning-2019-pdf (Accessed 31 July 2020).

Kraay, A., & Ventura, J. (2005). The dot-com bubble, the bush deficits, and the US current account. The World Bank.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th international conference on data mining workshops* (pp. 798–805). IEEE.

Lee, A. (2013). Welcome to the unicorn club: Learning from billion-dollar startups. *Cowboy Ventures (blog)*, https://techcrunch.com/2013/11/02/welcome-to-the-unicorn-club/ (Accessed on: 12 September 2020).

Litan, R. E., & Rivlin, A. M. (2001). Beyond the dot.coms: The economic promise of the internet. Brookings Institution Press.

Lussier, R. N. (1995). A nonfinancial business success versus failure prediction model for young firms. *Journal of Small Business Management*, *33*(1), 8, https://search-1proquest-1com-18e5j5izx069a.eczyt.bg.pw.edu.pl/docview/221008471?accountid=27375.

MacroTrends LLC (2020). NASDAQ Composite - 45 year historical chart. https://www.macrotrends.net/1320/nasdaq-historical-chart (Accessed: 20 June 2020).

Palan, R., Murphy, R., & Chavagneux, C. (2010). Tax havens: How globalization really works. In *Cornell studies in money*, Cornell University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Picken, J. C. (2017). From startup to scalable enterprise: Laying the foundation. *Business Horizons*, *60*(5), 587–595. http://dx.doi.org/10.1016/j.bushor.2017.05.002.

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

Ryan, C. L., & Lewis, J. M. (2017). Computer and internet use in the United States: 2015. US Department of Commerce, Economics and Statistics Administration.

Schwab, K., & World Economic Forum (2016). The fourth industrial revolution. World Economic Forum.

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine Learning and knowledge discovery in databases* (pp. 145–158). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based startup success prediction. In *CIKM '18*, *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 2283–2291). New York, NY, USA: Association for Computing Machinery, https://doi.org/10.1145/3269206.3272011.

Smola, A. J., & Schölkopf, B. (1998). Learning with kernels, vol. 4. Citeseer.

Spiegel, O., Abbassi, P., Zylka, M. P., Schlagwein, D., Fischbach, K., & Schoder, D. (2016). Business model development, founders' social capital and the success of early stage internet start-ups: a mixed-method study. *Information Systems Journal*, *26*(5), 421–449.

Spyros Makridakis (1996). Factors affecting success in business: Management theories/tools versus predicting changes. *European Management Journal*, *14*(1), 1–20. http://dx.doi.org/10.1016/0263-2373(95)00043-7, http://www.sciencedirect.com/science/article/pii/0263237395000437.

Stuart, R., & Abetti, P. A. (1987). Start-up ventures: Towards the prediction of initial success. *Journal of business venturing*, *2*(3), 215–230.

Sukhodolov, A. P., Popkova, E. G., & Kuzlaeva, I. M. (2018). Modern foundations of internet economy. In *Internet economy vs classic economy: struggle of contradictions* (pp. 43–52). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-319-60273-8_4.

Tomy, S., & Pardede, E. (2018). From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship. *Sustainability*, *10*(3), 602.

Vaughan, L. Q. (1999). The contribution of information to business success: a LISREL model analysis of manufacturers in shanghai. *Information Processing & Management*, *35*(2), 193–208. http://dx.doi.org/10.1016/S0306-4573(98)00048-X, http://www.sciencedirect.com/science/article/pii/S030645739800048X.

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Yuxian, E. L., & Yuan, S.-T. D. (2013). Investors are social animals: Predicting investor behavior using social network features via supervised learning approach.