A Machine Learning Approach to Detect Early Signs of of Startup Success

Supplementary Materials

1 DATA PREPARATION PIPELINE

The following section will describe in detail the first four steps of the data preparation pipeline as presented in Figure 1. The first four steps are the following: (1) Source data, (2) Extracted data, (3) Pre-preprocessed data, and (4) Merged data

1.1 Source and Extracted Data

U.S. federal small business R&D seed-fund

In May of 2020, we collected the names and descriptions of all ventures that received a Small Business Innovation Research (SBIR) grant or Small Business Technology Transfer Research (STTR) award. SBIR and STTR awards are granted through individual government agencies and provide small businesses (less than 500 employees) with zero-equity seed-funding to pursue R&D activities for technological innovations that have commercial potential and are relevant to each agency's mission. Approximately 5,000 SBIR/STTR awards are granted annually with individual award values ranging from $50,000 to $750,000 depending on the specific government agency and phase of the award. The SBIR/STTR program has an annual budget of $3.7 Billion, making it one of the largest seed-funding programs in the world.

We collected data from the public website of the SBIR[1] spanning 27,260 unique companies that were issued SBIR/STTR awards between the years 1983 to 2020. The data contained the name of the company, the company's physical address at the time of the award, a link to the company's website (if one existed), title of awarded project, year of award, and an abstract describing the proposed work to be undertaken during the award period.

Company Profiles

Using the unique company names extracted from the SBIR/STTR data, we used Google Custom Search JSON API[2] to search and collect the links to the profile pages of the companies on Crunchbase[3]. For companies without links, we searched for the Crunchbase link on the standard Google Search homepage. We were able to retrieve 23,691 Crunchbase links out of the 27,230 companies in the SBIR/STTR data. Using the Crunchbase links, the data from the Crunchbase profiles of the companies was collected using a custom-built scraping and parsing tool. The company name from the SBIR/STTR data was compared with the name on Crunchbase using the Levenshtein distance algorithm [2]; all companies with distances above 15 were discarded. As a result, we retained Crunchbase data for 16,880 companies. The data contained information such as industries the company operates in, founding date, company location, text description of the company, number of employees, lists featuring the company, web traffic statistics, current team, board members and advisors, and funding details.

The company-level data contained information on individuals affiliated with the company under the categories of 'Current Team' and 'Board Members and Advisors'. We collected additional information on these individuals from Crunchbase using our custom-built scraping and parsing tool. This resulted in data on 37,120 individuals from 9,413 companies. The data contained information such as the individual's job title, primary organization affiliation, investment history, board and advisor roles, gender, and education.

Word Wide Web

Utilizing the website links available in the SBIR/STTR data, we extracted the names of the companies and links to the company websites. 5,468 companies had links to their website, out of which 3,764 (69%) links were found to be active. We collected the raw HTML pages from these active websites.

1.2 Preprocessed Data

After removing sparse columns (greater than 90% of values missing) and columns with redundant or unnecessary information such as links to social media accounts, the data was reduced to 51 columns comprising of people-level information, 69 columns representing company-level data, and 1 column of website content.

People

Text columns (200 features): The Crunchbase description of each individual was transformed into a 200-dimensional vector representation using Doc2Vec [1].

Continuous columns (29 features): For each individual, we collected the total number of the following: organizations founded, events participated in, news articles appeared in, partner investments, past and current jobs, past and current 'Board & Advisor Roles', and advisor, board, investor, and other roles. We also collected the following continuous temporal data: duration of each job held and degree attained, years since graduation. Furthermore, we also collected the Crunchbase rank of the individual and each 'Related Hub'. Lastly, we collected the money raised in partner investments, and inflation-adjusted it to 2020 dollars. In total, 33 people-level continuous columns resulted in 29 continuous features.

[1] https://www.sbir.gov/sbirsearch/award/all

[2] https://developers.google.com/custom-search/v1/overview

[3] https://www.crunchbase.com/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICAIF 2021, November 3–5, 2021, Online
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

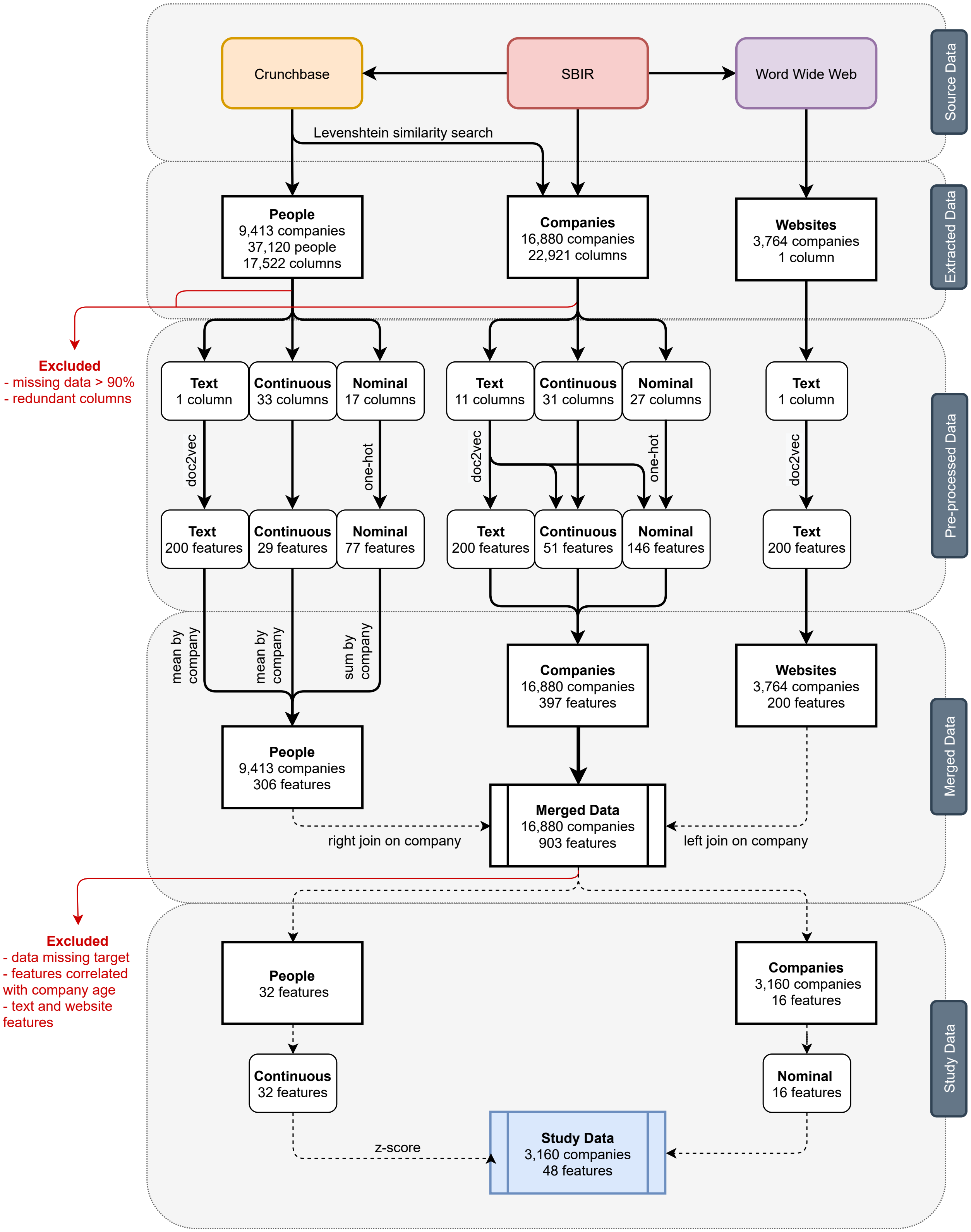

Figure 1: *Data Preparation Pipeline*: Pictorial representation of the five steps of the data preparation pipeline for this work: source data, extracted data, pre-processed data, merged data, and study data.

Nominal columns (77 features): For each individual, we one-hot encoded the following columns: gender (agender, male, female, non-binary, other, and prefer not to identify), regions (East Coast, Midwestern US, etc.), current position at the company (current team, and board members and advisors), academic degrees (bachelors, masters, MBA, PhD, other), academic majors (engineering, business, physical/life science, social science, mathematics, law, art, other), and title of the job previously held (entrepreneur, manager, assistant, intern, marketer, engineer, scientist, and other). We also one-hot encoded academic background (prestigious, and non-prestigious academic institution), and partner investment investor (prestigious investor and prestigious lead investor). Specifically, an individual's academic background was one-hot encoded according to the prestige of the university they attended. A university was considered prestigious if it is an Ivy League school and/or Top 10 in the Wall Street Journal Times Higher Education 2021 College Ranking List[4]. For each individual's investment that they partnered in, the investor was one-hot encoded based on the prestige of the investor and if that investor was also a lead investor. A investor was considered prestigious if they were in the 2019 Forbes Midas List[5] and resulted in two features: prestigious investor and prestigious lead investor. In total, 17 people-level nominal columns resulted in 77 features.

Companies

Text features (200 features): Columns with one-line summary text sentences were pre-processed as follows: interest signals of the employees, company tech and website stack, and patents and trademarks were extracted from the sentences and one-hot encoded while values of projected IT expenditure, number of technologies used in company tech and website stack, number of patents and trademarks, and global rank and monthly visitors of the website were extracted and recorded as continuous features. Number of organizations, total funding amount, and number of investors were extracted as continuous features for each list under the 'Lists Featuring This Company'. Finally, the text description of each company was transformed into a 200-dimensional vector using Doc2Vec [1].

Continuous features (51 features): A company's Crunchbase Rank and the number of organizations in each of the 'Related Hubs' were collected. We also gathered the total numbers of the following: employees (lower and upper bound), events, news articles, funding rounds, investors, and lead investors. Particularly, we collected the number of investors in each funding round and the total number of investors and lead investors of the company. Years since founding, duration between founding year and acquisition year, and funding round since the founding year were calculated by subtracting the founding year from 2020, acquisition year from founding year, and funding round year from founding year, respectively. Money raised in each funding round was inflation-adjusted to 2020 dollars. In total, 51 company-level continuous features were computed.

Nominal features (146 features): A company's type (1 indicating "for-profit", 0 for "non-profit") and operating status (1 indicating "active", 0 for "inactive") were encoded as a binary variable. The following columns were one-hot encoded: headquarters regions, funding status, and last funding status. A company's investors and partners were encoded as binary variables based on the prestige of the investor and if that investor was also a lead investor. An investor and partner was considered prestigious if they were in the 2019 Forbes Midas List[6]. 91 industries that companies operated in and occurred more than 100 times in the data were classified into 10 sectors (health care, financials, information technology, real estate, communication services, industrials, energy, consumer discretionary, consumer staples, materials) based on the Global Industry Classification Standard (GICS)[7], while the remaining industries were classified as 'Other'. In total, 146 company-level nominal features were generated.

Websites

Text features (200 features): To capture informational content contained within company websites in a structured format, we first extracted all text content within '$< p >$' tags of the websites and then trained a 200-dimensional representation for a given company's website using Doc2Vec [1].

1.3 Merged Data

The following subsection describes the process used to merge the pre-processed data for the companies, people, and websites.

Companies

The pre-processed data constituting information on people and websites were joined with the company data, values for missing people and website data were set to Null.

People

Before joining, rows of data on people had to be reduced to a single row to capture company-level information, hence, features representing information on people were aggregated using two techniques : for a given company (1) the mean values were calculated for both the 200 text features and 29 continuous features, and (2) the sum of the binary values of the 77 nominal features were calculated. Once the features of people for a given company were summarized, the data was joined with the company pre-processed data.

Websites

The 200-dimensional representations of the companies website were appended to the companies data by performing a join on the company pre-processed data.

REFERENCES

[1] Andrew M Dai, Christopher Olah, and Quoc V Le. 2015. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998* (2015).

[2] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.

[4] https://www.wsj.com/articles/best-colleges-2021-explore-the-full-wsj-the-college-ranking-list-11600383830?st=o78haqf5zxhl4is

[5] https://www.forbes.com/sites/forbespr/2020/04/14/forbes-publishes-19th-annual-midas-list-package-featuring-exclusive-content-regarding-vc-investing-in-the-current-environments/?sh=5b2ff74662da

[6] https://www.forbes.com/sites/forbespr/2020/04/14/forbes-publishes-19th-annual-midas-list-package-featuring-exclusive-content-regarding-vc-investing-in-the-current-environments/?sh=5b2ff74662da

[7] https://www.msci.com/gics

Table S1: Grouping of specific academic majors under general major categories

Major Category	Specific Academic Majors
Engineering	aeronautics, aerospace, artificial intelligence, astronautics, bioengineering, bioinformatics, biomedical, biotechnology, circuits, civil, computer science, computation, computational, computer networking, computer vision, computing, data mining, database, ee, electrical, electromagnetics, electronics, engineering, informatics, information, machine learning, mechanical, mechatronics, microelectrical, microelectromagnetics, microelectronics, microprocessor, nanomaterials, networking, networks, neural systems, programming, robotics, semiconductors, systems, signal processing, software, software systems, structural, system design, technology, telecommunications
Business	account, accountancy, accounting, acquisitions, administration, advertising, banking, business, capital, capital markets, commerce, communications, corporate, development, employee relations, entrepreneurship, equity, executive, finance, human resources, innovation, investment, leadership, logistics, management, managerial, marketing, mass communications, operations, organizational, public relations, real estate, sales, strategic, strategy, trade, venture capital
Physical/Life Science	sciences, allergy, anatomy, bacteriology, biochemistry, biological, biology, biomedical sciences, biophysics, biosciences, biosynthesis, botany, brain, cancer, cellular, chemical, chemistry, clinical, cognitive science, dental, dermatology, diagnostics, earth, ecology, embryology, enzymology, epidemiology, genetics, genomics, geochemical, geochemistry, geology, geomechanics, geophysics, health, health science, hematology, hydrogeology, hydrology, immunology, life, material sciences, medic, medical, medicine, microbiochemistry, microbiology, microbiosciences, microbiosynthesis, molecular, nephrology, neuroanatomy, neuropharmacology, neuroscience, nutrition sciences, oncology, pathobiochemistry, pathobiology, pathobiosciences, pathobiosynthesis, pathology, pharmacoepidemiology, pharmacology, physics, physiology, plant, premedical, protein, psychiatry, psychophysics, surgery, toxicology, virology, zoology
Social Science	policy, studies, american institutions, american studies, behavior, economics, government, history, humanities, international affairs, international policy, international relations, international studies, middle eastern, pharmacoeconomics, politics, political, psychology, public policy, security studies, social, sociology
Mathematics	actuarial, biostatistics, econometrics, mathematics, quantitative, statistics
Law	law, j.d., legal, copyright, intellectual property
Art	arts, cinema, film, guitar, music, theatre, violin, vocal

Table S2: Grouping of specific industries under general sectors based on the Global Industry Classification Standard (GICS)

Sector	Specific Industries
Health Care	Wellness, Biotechnology, Medical Device, Pharmaceutical, Therapeutics, Health Diagnostics, Clinical Trials, Biopharma, Hospital, Health Care, Fitness, Life Science, Medical, Genetics
Financials	Financial Services, FinTech, Finance, Venture Capital, Insurance
Information Technology	Developer Platform, Mobile, CRM, Internet of Things, Virtual Reality, Web Development, Cloud Computing, Enterprise Software, Mobile Apps, Machine Learning, Web Design, iOS, Software, Apps, Big Data, Internet, Artificial Intelligence, Semiconductor, Computer, Consumer Electronics, Robotics, Electronics, SaaS, Information Technology
Real Estate	Real Estate
Communication Services	Marketing, Advertising, Wireless, Video, Social Media, Telecommunications
Industrials	Commercial, Analytics, Product Design, Sensor, Cyber Security, 3D Printing, Legal, Transportation, Logistics, Government, Security, Manufacturing, Business Intelligence, Market Research, Product Research, Industrial, Industrial Engineering, Industrial Automation, Service Industry, Consulting, Professional Services, Information Services
Energy	Renewable Energy, Solar, Energy, Oil and Gas
Consumer Discretionary	Hardware, Education, Aerospace, Construction, Consumer Goods, Automotive, Retail, E-Commerce
Consumer Staples	Food and Beverage
Materials	Advanced Materials, Nanotechnology, Chemical, Agriculture
Other	Communities, Other