3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11633

Setting Crunchbase for Data Science: Preprocessing, Data Integration and Feature Engineering

Francesco Ferrati, Moreno Muffatto
School of Entrepreneurship (SCENT), Department of Industrial Engineering, University of Padova, Italy.

Abstract

In order to support equity investors in their decision-making process, researchers are exploring the potential of machine learning algorithms to predict the financial success of startup ventures. In this context, a key role is played by the significance of the data used, which should reflect most of the variables considered by investors in their screening and evaluation activity. This paper provides a detailed description of the data management process that can be followed to obtain such a dataset. Using Crunchbase as the main data source, other databases have been integrated to enrich the information content and support the feature engineering process. Specifically, the following sources has been considered: USPTO PatentsView, Kauffman Indicators of Entrepreneurship, Academic Ranking of World Universities, CB Insights ranking of top-investors. The final dataset contains the profiles of 138,637 US-based ventures founded between 2000 and 2019. For each company the elements assessed by equity investors have been analyzed. Among others, the following specific areas were considered for each company: location, industry, founding team, intellectual property and funding round history. Data related to each area have been formalized in a series of features ready to be used in a machine learning context.

Keywords: *Crunchbase; startup; investments; feature engineering; data mining; machine learning.*

1. Introduction

The large amount of business-related data available today, allows researchers in entrepreneurship, economics and social sciences to investigate complex phenomena using innovative approaches. In an economic system where entrepreneurship activities are considered as a driver for growth and social improvement, a very topical issue concerns the possibility of predicting to some extent the probability of success of early stage technology-driven ventures. Due to their inherently high level of innovation, startup companies are considered to be highly uncertain and risky business activities and the statistics on their failure rate are still very high (Gage, 2012). From an academic point of view, since the 1980s researchers have been analyzing the equity investors’ decision-making process, questioning about its effectiveness and wondering whether it could be improved (MacMillan, Siegel & Narasimha, 1985). The assumption underlying this research stream is that the use of more effective assessment criteria could lead to the identification of the best entrepreneurial projects, which might in turn contribute to the success of an investment portfolio (Zacharakis & Meyer, 1998).

Leveraging the growing amount of available data as well as the increasing accessibility of advanced data mining frameworks, in recent years researchers have started exploring new approaches to the so called “picking winners” problem. Specifically, they have begun investigating the potential of machine learning algorithms as a tool to support venture capitalist in their screening and evaluation processes. Given the complexity of the task, retrieving and processing data that can be used to properly model an early stage venture has a huge impact on the performance of the final models. In this regard, Crunchbase is an innovative online platform collecting and providing business information about technology-driven companies, investors, funding rounds and key people involved in the entrepreneurial network (Ferrati & Muffatto, 2020). Thanks to the quantity and quality of their data, it is effectively used not only by practitioners (e.g., entrepreneurs, investors or policy makers), but also by academic researchers who intend to apply quantitative approaches to the research on entrepreneurship and innovation. In this context, a key problem that can be addressed by applying machine learning algorithms to Crunchbase data concerns the prediction of a company's exit event, commonly considered as the critical milestone defining a company’s financial success. (Krishna, Agrawal & Choudhary, 2016).

Although previous works on this topic have usually described the data modeling process in detail, pre-processing, data integration and feature engineering activities have not always been covered in depth. Specifically, the logic used for the identification and selection of the features used in the models as well as the steps followed to obtain the considered samples have generally not been fully described. As a result, the considered datasets have not always been clear in their content and the models’ results could be therefore hard to interpret. In addition, in previous contributions Crunchbase has generally been used as the only data source, and no activity has been carried out to integrate other databases to enrich the information content. The present work aims to fill these two research gaps by providing a full example of how the database can be prepared according to the well established-steps of the data science workflow.

This paper is organized as follow. Section 2.1 defines the research goal for which the database can be used. Section 2.2 reports the relevant data sources that have been integrated into Crunchbase in order to add some key features. Section 2.3 describes in full detail all the steps that have been followed to create the final dataset. Finally, Section 3 discuss the obtained results and presents some elements for future research.

2. Steps of the data setting process

The setting of the dataset was carried out following the main steps of the data science process. After carrying out a literature review of the assessment criteria used by equity investors (Ferrati & Muffatto, 2019), and identifying the key information to evaluate a startup company, we defined the purpose for which the database can be used. Considering Crunchbase as the main data source, we then search for the most useful publicly available data to enrich its information content. We then moved on to the data preparation phase, going to combine, clean and transform the available data. Once aggregated the different datasets, we then made an analysis of the available variables, performing a feature selection and feature engineering activity. Figure 1 shows in bold the steps presented here in the context of a data science workflow.

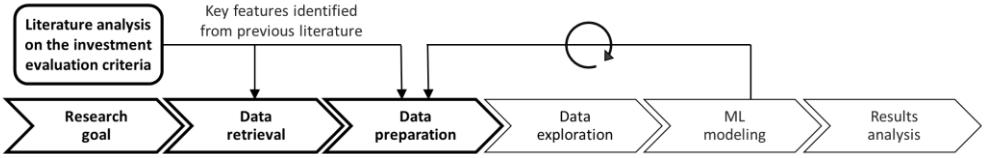
Literature analysis on the investment evaluation criteria
Key features identified from previous literature
Research goal
Data retrieval
Data preparation
Data exploration
ML modeling
Results analysis

Figure 1. The six steps of the data science workflow.

2.1. Research goal definition

In order to prepare data effectively, the first step is to clearly define the research goal for which the dataset will be used. The dataset we prepared can be used in a supervised machine learning environment to predict the financial success of a startup company. In this context, a venture's financial success is defined by the occurrence of an exit event, both in the form of merger and acquisition (M&A) and an Initial Public Offering (IPO). Crunchbase provides information about the status of a company through a specific categorical variable that can assume four different values: operating, closed, acquired or IPO. The “status” can be effectively used as a target variable in a machine learning classification problem.

2.2. Data retrieval

Since a machine learning model should support investors in their screening and evaluation process, the largest number of significant variables should be identified in the data retrieval phase to cover most of the aspects generally considered by investors. Crunchbase itself provides many useful information. The database is organized in seventeen .csv files: *Organizations, Organization descriptions, Category groups, Funding rounds, Investors, Investments, Investment partners, Funds, People, People descriptions, Jobs, Degrees, Acquisitions, Ipos, Organization parents, Event appearances* and finally *Events.* By grouping the information contained in each individual dataset, the complete database covers five macro information areas respectively related to organizations, investment activities, people, exits and public events. In order to map the content and give a representation of how the different datasets are linked together, we started by inferring the Crunchbase relationships scheme as shown in Figure 2. After an accurate exploration of the individual files, we select the datasets (colored in black) that could provide the most relevant information according to the considered research goal. Despite the large amount of information already provided by Crunchbase, we decided to go further and integrate it with other additional data sources in order to enrich the information content. Specifically, as shown in Figure 2, four additional data sources have been considered.

- **United States Patents**. Because of the high level of innovation experienced by technology-driven startups, a key competitive advantage concerns their intellectual property portfolio. Since Crunchbase does not directly provide this kind of information, we used the public data collected by the USPTO PatentsView platform to search for patents assigned to each company. A similar process has been reported also in previous literature considering the PATSTAT dataset (Menon & Tarasconi, 2017).
- **Kauffman Indicators of Entrepreneurship**. For each company in the database, Crunchbase provides its location in terms of country, state, region and city. Since the entrepreneurial ecosystem has a strong impact on a company's performance (Sheriff & Muffatto, 2018), the values of the Kauffman Indicators of Entrepreneurship have been integrated, providing some key metrics for each US state.
- **Academic Ranking of World Universities**: In order to assess the educational background of each company’s team members, we integrated the ARWU 2018 ranking to identify the founders who graduated from a worldwide top university.
- **CB Insights top investors list**. The fact that a successful investor decides to support a startup has a strong impact in terms of credibility and can facilitate the occurrence of subsequent funding rounds. In order to identify the companies backed by top investors, we merged the investments’ data with the ranking made by CB Insights

in 2019. This list provides the 48 top-investors who have backed the most unicorn companies.

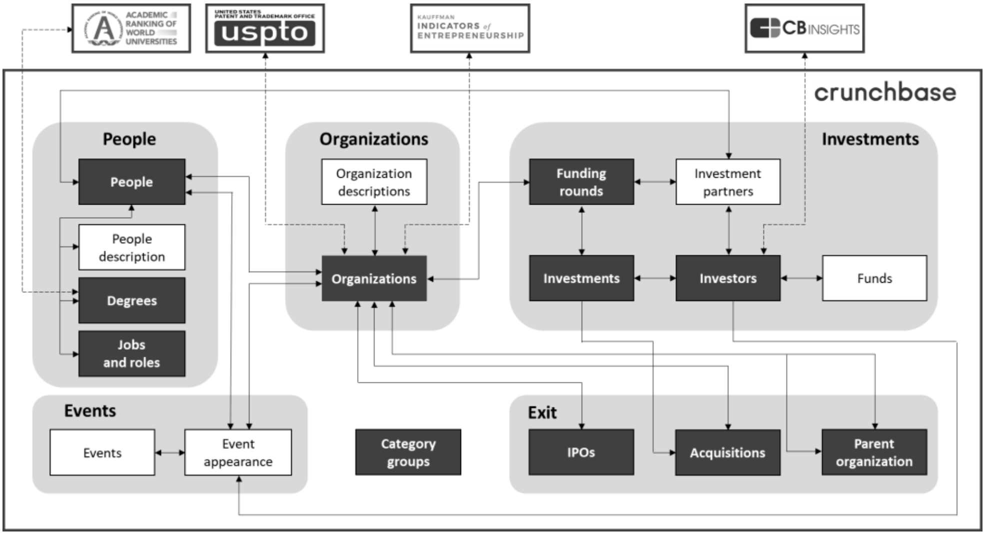

Figure 2. Scheme of the relationships between the Crunchbase datasets and the additional integrated sources.

2.3. Data preparation

The version of Crunchbase considered in this work is dated May 21, 2019 and contains information about 760,590 organizations, 121,509 investors (e.g., venture capital, angel investors, etc.), 263,426 funding rounds, 890,429 people, 1,346,357 jobs, 17,068 IPOs and 89,959 acquisitions. Starting from the raw data, the final dataset consists in a single file and considers the company as unit of analysis. Each row corresponds to a different company and more than 130 columns provide the features that can be used for machine learning. All the steps that led to the creation of the final dataset are described below in sequential order.

1. **Add exit info to organizations**. Starting from the "organizations" dataset, the information about exit events was integrated from the "acquisitions" and "ipos" datasets. Whereas an organization was acquired several times or reported more than one IPO, only the first exit event was considered. Among the 760,590 organizations, 85,099 have been acquired and 16,457 went public. A boolean value was used to label the organizations that made an exit and the target variable was so defined.
2. **Add lifetime info to organizations**. In order to take into account the phase in the business life cycle of each company, a variable has been added for the computation of the company's lifespan (in number of months). For the organizations with status equal to "acquired", "ipo" or "closed", the difference between the date of these events and the foundation date was considered. For the still "operating" companies,

the difference between 2018-12-31 and the date of foundation was considered. Organizations with missing dates were excluded. After this step, 518,661 organizations remained.

3. **Filtering companies**. Since Crunchbase considers as organizations both companies, investment firms and universities/schools, the dataset was filtered considering only companies (489,710 remaining companies). To focus on a specific entrepreneurial ecosystem, only US-based companies were then considered (195,542 remaining companies) and in order to not suffer the effects of “dot-com bubble”, only companies founded between 2000 and 2019 have been finally considered (138,637 remaining companies).
4. **Founding team analysis**. Since the quality of the founding team is a key element for the success of a startup, for each company information about its founders has been integrated. In order to identify the people related to each company, the "jobs" and “people” datasets were used, selecting only the jobs related to the considered companies. Of the 1,346,357 records in the "jobs" dataset, 458,429 were related to the considered companies. On the other hand, 45,153 companies didn’t report any job information, resulting in a lack of data about their founding team. In order to identify the founders of each company, only "founder" or "co-founder" job types were considered (84,834 unique founders were identified, some of whom had founded more than one company). In order to get a measure of the founders’ work experience, all the jobs carried out in their career have been identified (195,993 jobs carried out by the founders have been identified). To understand the roles covered by founders, for each company the presence of some key chief-roles (e.g. CEO, CTO, COO, CFO, etc.) has been verified. As the "degrees" dataset provides information about the education of the registered persons, for each founder an analysis of the educational background was carried out. Of all the 335,414 degrees collected in the dataset, only those related to founders were extracted, resulting in 100,366 degrees. Each degree was classified according both to the type (e.g., bachelor degrees, master degrees, doctoral degrees, etc.) and to fifteen subject areas in order to associate to each company the team’s areas of knowledge (e.g. business, engineering, computer science, science, etc.). For each degree, it was also verified whether the title was obtained from a university in the top 25 of the ARWU 2018 ranking. In total 11,783 degrees were obtained from a top university. Finally, for each company the gender of the founders has been considered (73,853 males, 10,586 females, 7 other and 46 not declared). In summary, the team analysis allowed to define for each company the following groups of features: number of founders (N.F.), N.F. per chief role, N.F. per type of degree, N.F. per subject degree, N.F. from top universities, N.F. per gender, average number of companies founded by each founder (serial entrepreneurs).

5. **Location analysis**. In order to enrich the information about the entrepreneurial ecosystem in which each company operates, we integrated Crunchbase with the five Kauffman indicators of entrepreneurship (i.e., Kauffman early-stage entrepreneurship index, rate of new entrepreneurs, opportunity share of new entrepreneurs, startup early job creation, startup early survival rate). Since each index assumes different values depending on the year and the US state, the association was made considering the state in which the company is located and the average (and median) value of the indicators calculated over the years of the company's lifespan.
6. **Sector analysis**. Crunchbase provides two variables, called “categories” and “category groups”, for the classification of companies’ activities. The variable "categories" can take one or more labels (related to industries, technologies, business models, etc,) from 680 possible options. To make these values more consistent with each other, we introduced a new classification scheme in order to reduce the 680 categories to 64 main areas. Then, we reclassified each company according to the 64 areas so identified.
7. **Investments analysis**. One of the most important information provided by Crunchbase regards the funding rounds collected by each company. The “funding rounds” dataset reports 263,426 rounds and the “investments” dataset collects 400,432 investments (the term *investment* refers to the participation of a single investor in a specific funding round). The information has been filtered considering only the rounds raised by the selected companies. As a result, we identified 52,037 companies with at least one funding round, 113,572 rounds related to them and 30,975 unique investors involved. For each company the number of rounds, the total amount collected (in USD) and the number of total as well as unique and serial investors were calculated. Finally, considering the top-investors ranking by CB Insights, for each company the number of top investors in portfolio was computed.
8. **Acquisitions analysis**. The Crunchbase “acquisitions” dataset collects information about 89,959 acquisitions. Starting from this data, the number of acquisitions made by the selected companies have been derived. As a result, we identified 12,223 acquisitions made by 6,369 companies among the considered one.
9. **Patent analysis**. Since the companies in the sample are all based in the USA, their patent portfolio was analyzed using the data collected by the USPTO PatentsView platform. Since both Crunchbase and PatentsView provide information about the location of each company and assignee, the match between the two datasets was made using a concatenate string between the company name and the U.S. state where it is based. In this phase special attention was paid to make the companies names homogeneous between the two datasets (e.g. by removing all the legal entity types abbreviations) and to manage homonymy cases. All the patents dated after a

company acquisition or IPO were excluded from the computation. The merge process identified 9,025 companies in the sample with at least one patent. For each of the identified companies, all the patents registered in the "patent" dataset were searched and 69,537 patents were identified. Patent data has been filtered considering only utility and design patents and patent applications were dropped from the computation. For each company the number of patents (utility and design) was reported.

Table 1 summarizes the content of the final dataset.

Table 1. Content of the final dataset.

	Number
Companies (based in USA and founded between 2000 and 2019)	138,637
Founders	84,834
Funding rounds	113,572
Companies with at least one funding round	52,037
Investors involved in the considered funding rounds	30,975
Companies having made at least one acquisition	6,369
Acquisitions made by the companies	12,223
Patents granted to the companies	69,537
Companies with at least one patent	9,025

3. Conclusion and future research

In this paper we presented the steps that could be followed to prepare Crunchbase to be used in machine learning to predict a startup’s exit event. A series of filters and operations were applied to make the dataset as consistent as possible, and to integrate other information not considered in previous contributions. For future research, other data sources could be integrated to cover aspects related to products or services, business models, competitors and financials. Feature importance could be analyzed by applying logistic regression algorithm.

Acknowledgment

This research was made possible thanks to the support of Crunchbase Inc. http://www.crunchbase.com

References

Ferrati, F., & Muffatto, M. (2019). A Systematic Literature Review of the Assessment Criteria Applied by Equity Investors. *In 14th European Conference on Innovation and Entrepreneurship* 304-312

Ferrati, F., & Muffatto, M. (2020). Using Crunchbase for research in Entrepreneurship: data content and structure. *In 19th European Conference on Research Methodology for Business and Management Studies*

Gage, D. (2012). The venture capital secret: 3 out of 4 start-ups fail [online]. *The Wall Street Journal U.S. Edition*, updated Sept. 20, 2012 12:01 am ET.

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 (ICDMW)* 798-805.

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

Menon, C., & Tarasconi, G. (2017). Matching Crunchbase with patent data. *OECD Science, Technology and Industry Working Papers*, 2017(7), 1-21.

Sheriff, M., & Muffatto, M. (2018). High-tech entrepreneurial ecosystems: using a complex adaptive systems framework. *International Journal of Entrepreneurship and Innovation Management*, 22(6), 615-634.

Zacharakis, A. L., & Meyer, G. D. (1998). A lack of insight: do venture capitalists really understand their own decision process? *Journal of Business Venturing*, 13(1), 57-76.