Key Characteristics of Successful Startups

By Joseph Higgins, Marin Petel and Roy Spencer

Motivation

This project is an attempt to identify key characteristics that make startups successful. We chose this topic because all three of us have an interest in financial services and hope to learn more about private equity and debt.

Ultimately our goal is to better inform the buyer-side of these firms. This project is intended to be a first step towards a machine learning business valuation tool that can help firms to better assess company values and pricing factors when conducting a merger or an acquisition.

Numerous researchers (see slide 11) have tried, with mixed results, to use machine learning techniques to identify success factors, mostly through supervised classification models.

Identifying a successful startup target is extremely challenging - according to Forbes, 90% of startups fail. We defined success for startups as startups that went through an IPO or were acquired. To note - a small number of startups manage to reach a valuation of over USD 1 billion and are generally referred to as 'unicorns'.

Potential concerns arose surrounding personal data. We obfuscated where possible, and didn't list any PII - but instead worked with abstracted identifiers.

Objective

Our main objective was to answer the below questions and publish our results in similar fashion to a market research and advisory firm.

- *Are there network effects across factors?*
- *Do the following variables impact the success of startups?*
  - *Geography*
  - *Industry*
  - *Size of company*
  - *Quantitative funding factors*
  - *Qualitative funding factors*
  - *Education*
  - *Exogenous factors*
    - *Country Demographics*
    - *Market Size*

Data Sources

Primary Dataset

Our main dataset is an extraction from the Crunchbase platform as of December 2013, made available on Kaggle.

Crunchbase is a platform that provides an intelligent prospecting software powered by live company data. Their content includes investment and funding data, founding members and individuals in leadership, positions, mergers and acquisitions, news, and industry trends.

To note - several research articles on the drivers of startups use Crunchbase as their primary source of information.

Our main dataset consists of 11 CSV tables that can be joined by identifiers.

- *acquisitions.csv, 9562 records, <1MB*
- *degrees.csv, 109610 records, 6.7MB*
- *funding_rounds.csv, 52928 records, 9.3MB*
- *funds.csv, 1564 records, <1MB*
- *investments.csv, 80902 records, 3.7MB*
- *ipos.csv, 1259 records, <1MB*
- *milestones.csv, 39456 records, 2.7MB*
- *objects.csv, 462651 records, 141.2MB.*
  - ***This is the main file***
- *offices.csv, 112718 records, 12.9MB*
- *people.csv, 226709 records, 10.4MB*
- *relationships.csv, 402878 records, 33.8MB*

Secondary Datasets

Our secondary datasets span across topics ranging from macroeconomic indicators to market capitalizations.

- *Wikipedia*
  - *To retrieve a list of unicorn companies*
- *The World Development Indicators (WDI) data set compiled by the World Bank*
  - *Zip archive of 200MB*
  - *Containing more than 1400 time series indicators*
  - *To retrieve macro-economic data and indicators*
- *Yahoo Finance*
  - *To retrieve a list of company's market capitalizations to compare with valuations for past ~10 years*
  - *Accessed via API calls to the yfinance module*

Data Cleaning, Manipulation and Methodology

Cleaning

- *We dropped unnecessary data (i.e. company logo information).*
- *We converted formats - such as when strings contained dates that needed to be formatted as timestamps.*
- *We handled missing data - our materiality assessment of missing data for some features led us to simply ignore absent values - which means our scope is never exactly the same, but successful startups always remain in strong minority.*
- *We handled abnormal values: while looking at descriptive statistics, logical consistency ruled that our dataset was not exclusively composed of startups. Some companies are very old and well established - this led us to adopt a rule to identify startups.*

Manipulation

- *Merge: we performed merges between csv files to join them together.*
- *Groupby: we used Pandas' groupby extensively through the project to compute summary statistics between startups.*
- *Crosstab: we used this function to compute the adjacency matrix for network representation.*
- *Qcut: discretized into bins quantitative features (eg: size of companies into 10 buckets corresponding to the deciles of its distribution)*
- *Apply: used throughout the project to create flags based on several conditions (eg: successful if acquired or ipo and/or unicorn)*

Methodology

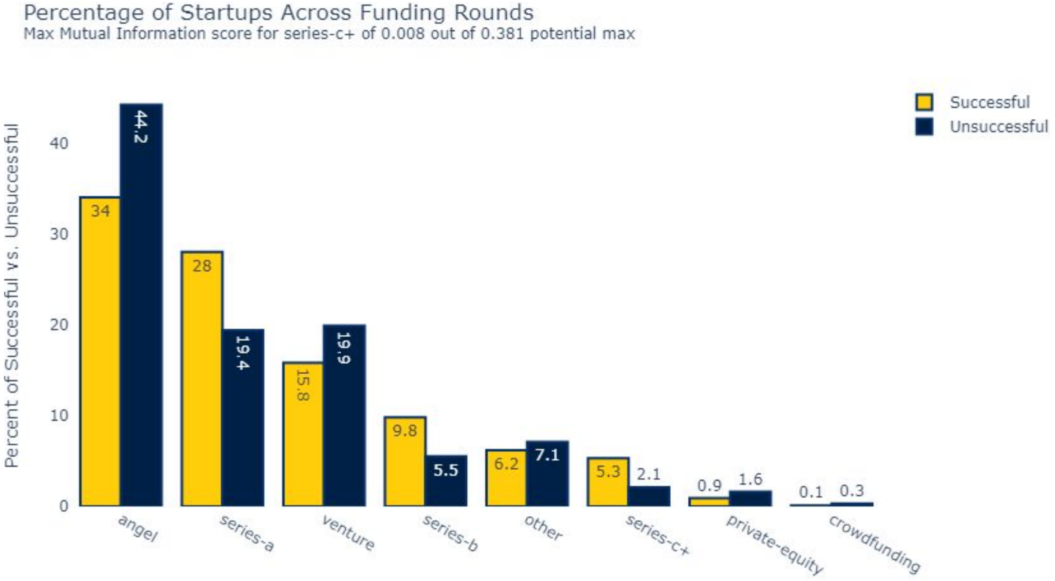
- *The most significant challenge was to identify startups as there is no clear definition of what a startup is, and there is generally a variety of criteria to consider. To simplify the process we flagged companies with (at most) 5 years of existence. This criteria was regularly mentioned in the various definitions we consulted and was straightforward to compute. However, this method has limitations - and we wrongfully flagged a small proportion of companies as startups (eg: General Motors is flagged as a startup because of a founding date in 2008 due to the many government bailouts that occurred around that time) although it is hard to quantify.*

#1 Analysis and Visualization: Funding Type and Round

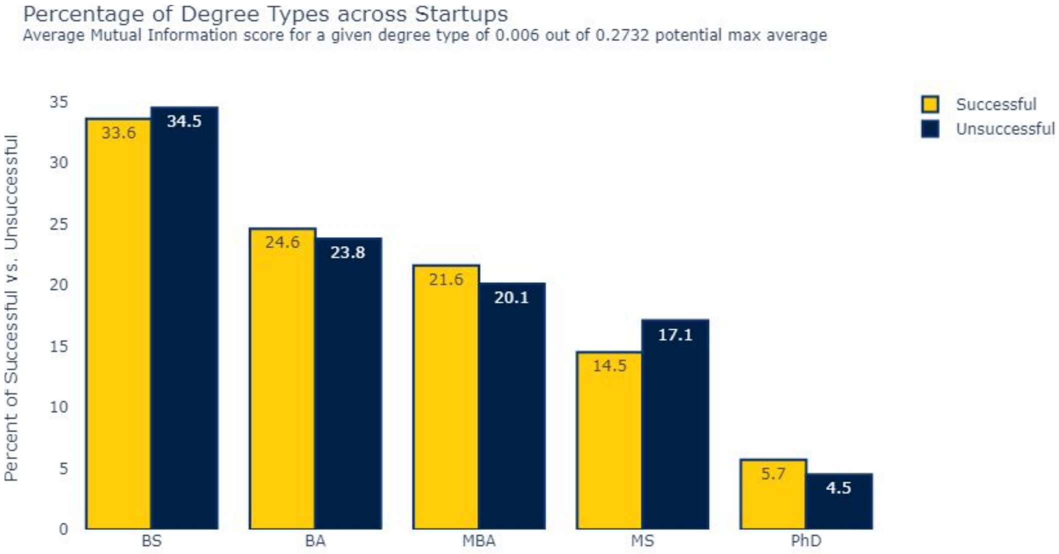
The majority of successful startups were funded by Angel investors. *About 34% of our startups to be exact. According to Investopedia, an Angel investor is, "...a high-net-worth individual who provides financial backing for small startups or entrepreneurs, typically in exchange for ownership equity in the company…".*

The second highest group of successful startups (~30%) reached the 'Series-A' round of funding - which is a fundamental step in the funding process.

Lastly - our third most notable group of successful startups (~16%) were funded by Venture capital. Again - according to Investopedia, "Venture capital (VC) is a form of private equity and a type of financing that investors provide to startup companies and small businesses that are believed to have long-term growth potential. Venture capital generally comes from well-off investors [and] investment banks…".



Analysis and Visualization: Degree Type #2



The majority of successful startups are staffed by associates who hold a Bachelor of Science degree. *About 34% of our startups to be exact. This degree typically takes three to five years to complete and is generally selected by students who wish to continue their chosen discipline throughout graduate school. An interesting fact about Bachelor of Science degrees uncovered while researching this topic - the University of Michigan was the second university ever to issue a B.S. degree.*

Surprisingly, a minority of successful startups are staffed by associates with PHDs (~5.7%). A number of confounding variables could be causing this low count of PHDs - for example, the particular rigor of PHD programs most likely means there are lesser of these degree types in circulation - also, many PHD holders may choose academia instead of industry.

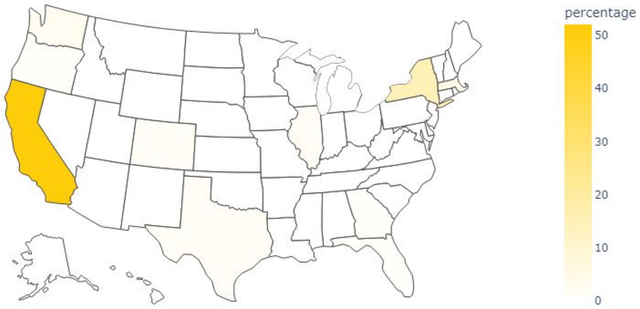
#3 Analysis and Visualization: Geography

Successful startups are mostly concentrated in the US. *In a global context, the countries for which we see the highest concentration of success (US, UK, Canada, Germany, India, France, Israel) are the ones deemed the most "startup friendly" according to this ranking from CEOWORLD magazine. This ranking is an annual assessment of a country's competitiveness in "scientific and technical-focused" economies.*

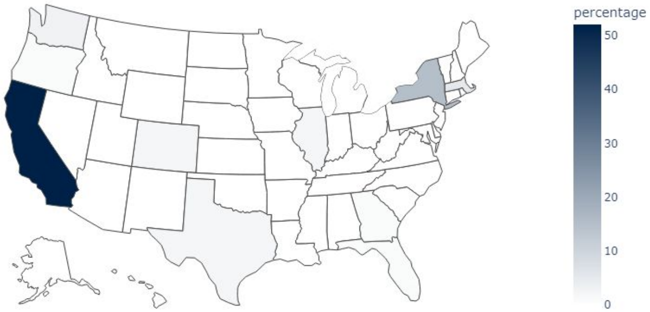
According to StartupRanking, ~54% of startups in the world are located in the US, whereas in our dataset only ~36% of startups are in the US. ***The high concentration of successes in the US for our dataset (76%) is therefore not due to selection bias*** *- which would be a possible reason for over-representation in the data.*

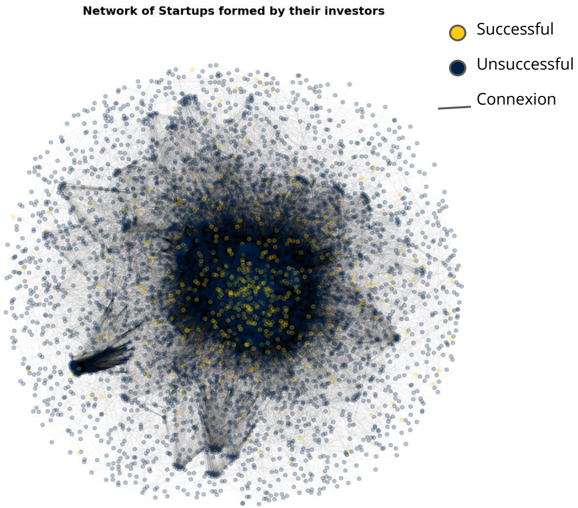
A startup's home country has relatively low discriminatory power with a lower mutual information score of .005. Zooming in on the US, we can look at the spread (to the right). There are minimal differences between successful and unsuccessful- California and New York concentrate both categories of startups.

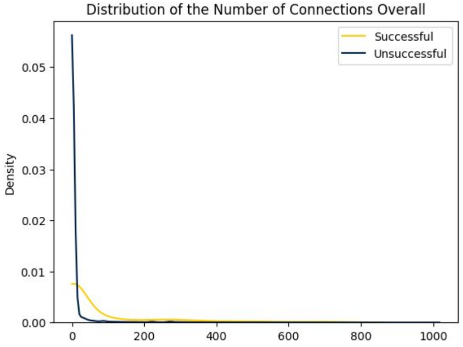
US successful startups by state



US unsuccessful startups by state







Analysis and Visualization: Network Effect #4

Successful startups tend to have more investors in common with other startups.

We wanted to visualize the startups as a network - so we formed connections between startups if they had at least one investor in common. Most of the startups appear isolated, having no connections (hence both densities peak at 0 on the distribution graph). However, a significant connected network of 6,574 startups does appear (~12% of our dataset). ***An interesting finding is that 43% of our successful firms are in this network.***

The network uses a 'spring' layout in which startups, represented by nodes, are put as far apart from each other as possible, but the connections between them, represented by straight lines, act a spring and hold them closer together. We see that success tends to be mostly in the center, meaning that successful firms have more connections than unsuccessful ones, which tend to be more scattered.

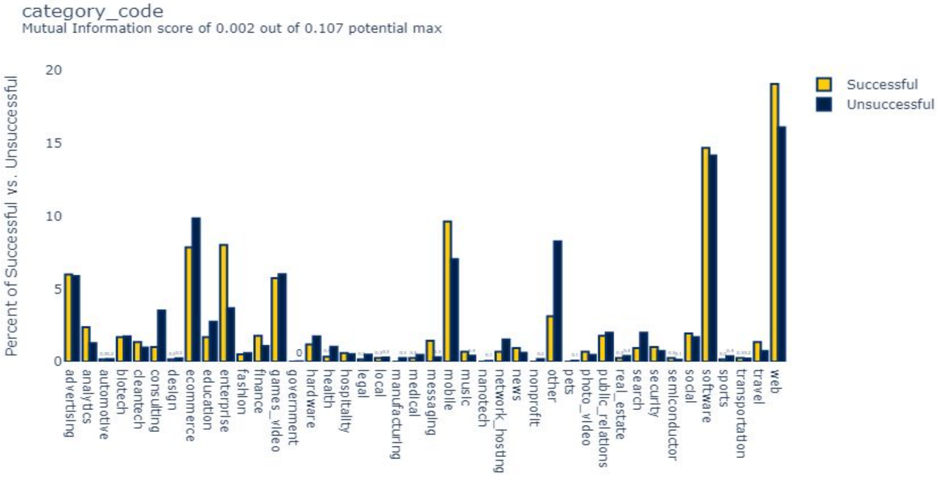
The number of connections for a startup has the highest discriminatory power *(with a mutal information score of 0.015) out of all our analysis (see slide 10), while* ***preserving the most our original scope as we have almost no missing data****. For unicorns only - we could not check if investors appeared after reaching success (as opposed to prior). If this was the case, using the number of connections to bet on success might be less informative than originally thought, as their number would increase but partly post-success.*

#5 Analysis and Visualization: Industry and ICT

Startups (both successful and unsuccessful) are mostly concentrated in the web, software, mobile and e-commerce industries.

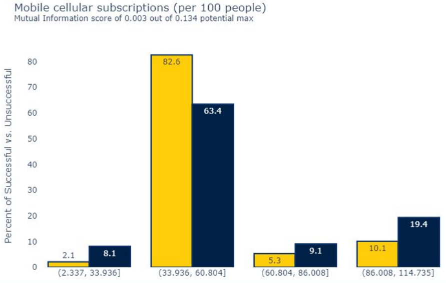
There doesn't seem to be a meaningful difference between successful and unsuccessful startups in terms of industries repartition. The few industries for which we can see bigger gaps (eg: consulting, enterprise, etc.) are typically less frequent. This low discriminatory power is reflected in a very low mutual information score. ***An interesting element here is that the most important industries are all related to Information and Communication Technologies (ICT).***

Therefore we explored potential links between success of a startup and the quality of the ICT infrastructure of the country it is located in, so we computed an average score for different ICT indicators between 2000 and 2007. The results can be seen to the right, but are not conclusive as mutual information scores remain very low at 0.002.



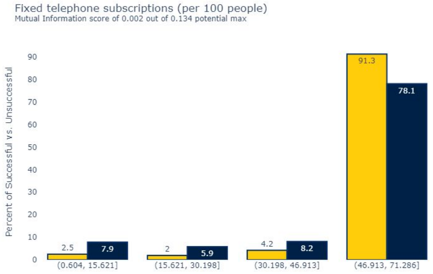
category_code

Mutual Information score of 0.002 out of 0.107 potential max



Mobile cellular subscriptions (per 100 people)

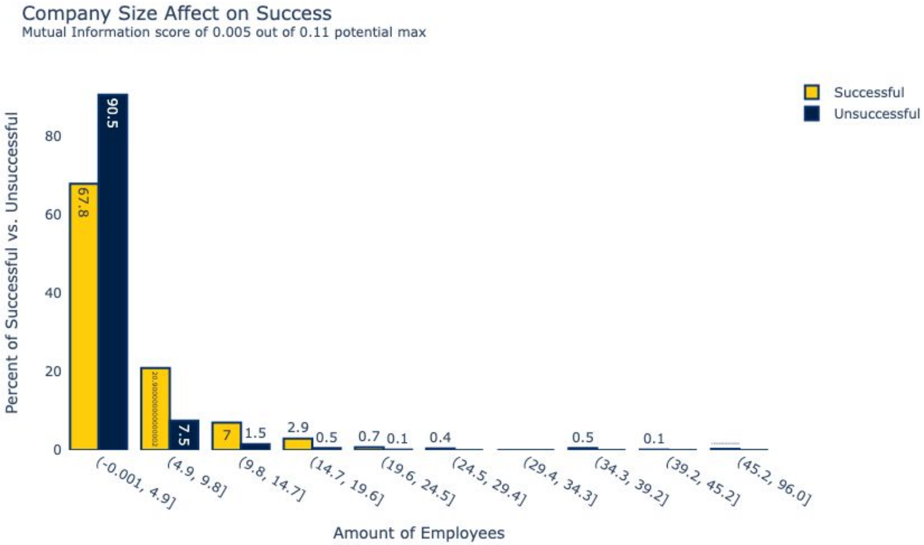
Mutual Information score of 0.003 out of 0.134 potential max



Fixed telephone subscriptions (per 100 people)

Mutual Information score of 0.002 out of 0.134 potential max

Analysis and Visualization: Company Size #6



Company Size Affect on Success

Mutual Information score of 0.005 out of 0.11 potential max

Startups with more employees tend to be more successful than startups with fewer employees.

The more employees a startup has - the better its chances of success. Startups that have between 0 and 4 employees tend to be more unsuccessful than successful, percentage-wise. If we double the number of employees (between 5 and 9) we see that statistic flip and there are more successful startups than unsuccessful startups, percentage-wise. Zooming in on startups that staff between 35 and 39 employees we see that there are almost no unsuccessful companies.

One thing to note is that this graph is created with the qcut function (from Pandas), thus the right side of the range is a partial employee estimate, but we can't have companies with a percent of an employee.

One limitation - the estimated size of the company is a proxy for employee count, and some companies reported a size zero - which isn't possible.

Conclusion and Next Steps

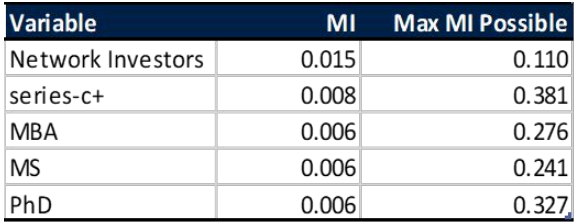
Conclusion

- *The majority of successful startups were funded by Angel investors.*
- *The majority of successful startups are staffed by associates who hold a Bachelor of Science degree.*
- *Successful startups are mostly concentrated in the US.*
- *Successful startups tend to have more investors in common with other startups.*
- *Startups (both successful and unsuccessful) are mostly concentrated in the web, software, mobile and e-commerce industries.*
- *Startups with more employees tend to be more successful than startups with fewer employees.*

Next Steps

- ***Funding seems to be a key element**, as this dimension accounts for our two top MI scores. It could be explored further by looking at: how quickly startups get access to funding after creation, what combination of type of funding matters the most. Additionally, the network of investors could have been weighted to reflect that sharing several investors with another startup to ensure more meaningful insights.*
- *We could have explored the number of degrees held by board/c-suite members, the diversity of majors, or perhaps analyzed the network effect of universities. However, **the strong limitation with analysis related to workforce is that the sourcing of this data is not complete (and the timing of acquisition process unclear)**.*

Mutual Information Table (Top Five)



Variable	MI	Max MI Possible
Network Investors	0.015	0.110
series-c+	0.008	0.381
MBA	0.006	0.276
MS	0.006	0.241
PhD	0.006	0.327

Statement of Work and Sources

Statement of Work

Joey Higgins

- *Cleaned - degrees.csv, funding_rounds.csv, ipos.csv and the yahoo finance data*
- *Methodology - Established product vision, importation functions and set up project workspace in DeepNote*
- *Analysis - Analyzed degree, funding round, ipo and market cap data*
- *Project Management - Coordinated weekly Zoom chats and sent meeting minutes*

Roy Spencer

- *Cleaned - funds.csv, milestone.csv,* investments.csv
- *Methodology - Research coordinator for source creation; aimed at understanding key valuation methods and macro trends in the private equity markets*
- *Analysis - Reviewed final versions of the files to ensure understanding and quality*

Marin Petel

- *Cleaned - objects.csv, people.csv, relationships.csv, offices.csv*
- *Methodology - defined startup and success criteria, retrieved unicorn from Wikipedia, computed all mutual information scores*
- *Analysis - countries/state, ICT, network of investors*

Collaboration was productive both in terms of content ideas and coding

- *How to improve in future work? Conducting review and challenge sessions from external peers during the analysis could be a powerful tool.*

Sources

- *J. Arroyo et al.: Assessment of Machine Learning Performance for Decision Support in VC Investments, September 2019*
- *K. Żbikowski, P. Antosiuk: A machine learning, bias-free approach for predicting business success using Crunchbase data, February 2021*
- *Ganti, A. (2022). Angel Investor Definition and How It Works. Investopedia. https://www.investopedia.com/terms/a/angelinvestor.asp*
- *Hayes, A. (2022). Venture Capital: What Is VC and How Does It Work? Investopedia. https://www.investopedia.com/terms/v/venturecapital.asp*