THE HOLY GRAIL OF VC

Final Report of DATA-X

Alex Nakagawa

Anny You

Dimitrios Hytiroglou

Mert Gurkan

Lorenzo Ong

Surya Sendyl

INTRODUCTION

In 2016, there were $57.4 million in capital flows of Venture Capital financing, going towards 3,718 companies. However, based on previous data, the mean period for a start up's Initial Public Offering (IPO) is 7 years, but there were only 39 companies that went public last year. With VCs faced with this 1% chance that an investment will actually be successful, choosing the right founder and startup to bet on can be quite a random process, and is usually done based on quantitative metrics. Our team approaches this by combining both quantitative and qualitative aspects - from personality surveys to personal information and company data - to ultimately predict a founder's likelihood of success. In this case, we will define success as the startup being able to raise a Series B.

PROBLEM

Applying machine learning for Venture Capital is a hot topic, but few people takes action due to its uncertainty. During this project, we faced quite a few problems.

First of all, we were met with a dilemma to choose between predicting success of a potential future-founder OR a current founder+startup. Then, we seeked help from our mentors Patrick Chung (a VC) and Kevin Liu (angel investor) to give us some business requirement in their industry. They advised that the latter is more of a problem for a VC/angel.

The second problem: how can we get the data? To begin, we reached out to ScrapingHub, and received a JSON dump of 1.5k Crunchbase profiles. The only problem was that the dataset was too small to perform a classification on. Next, we tried scraping/downloading AngelList, LinkedIn, Pitchbook, Mattermark data, but we decided that we were wasting lots of time to cross-reference the rows person-by-person. Finally, we settled on Crunchbase as our dataset, and professor Ikhlaq Sidhu gave us Pro access to all the features Crunchbase has to offer us. From there, we manually exported 100k rows of founder/company data. Thanks to a little manual labor done by each team member, we finally had our data to begin our modeling.

Thirdly, although Crunchbase provided many useful data points, we wanted to avoid deferring to solely to the Crunchbase data and explore non-objective, non-third-party data sources. We decided to send out a short survey to roughly 2000+ founders in the Bay Area to capture first-person psychological traits of successful founders, and mapping this data to our initial dataset. We only received 50 responses, and on top of that, only 15 people who responded were in our training data.

Finally, Crunchbase data was a mess to wrangle with in a CSV format (different data types, strings in int columns, ints/floats in string columns, missing/unicode values, duplicates), so it's
common in data preparation to do data cleaning using a combination of pandas, numpy and excel to join and clean data. Our group had to go back and repeat this step continuously to account for the many compatibility issues popped up.

SYSTEM SOLUTION

Our system, as with most machine learning systems, starts with the data. After pursuing various avenues to get data for our problem, including speaking with companies and VCs as well as attempting scraping, we settled on manually exporting data from Crunchbase. This data comprises of 100K rows of people data and 100K rows of company data, both sorted by Crunchbase rank. We then joined these two datasets on the 'Primary Organization' column, which lead to about 60K rows of data. This dataset had to be filtered further for only founders, both successful and not, so we filtered the 'Job Title' column for those that contained only 'founder' or 'ceo'. We then dropped NaN values for the columns left over, of which there were many, which lead us to have 25K rows of data remaining.

Once this process is completed, we selected various features that we wanted (can be found here and below) and recoded them to categorical or numerical. We also create an independent output variable, which is binary for whether the startup and founder raised at least a Series B.

To augment our Crunchbase data, we also have the survey data (survey found here), which, due to the small number of respondents (n=50), we use in isolation from the Crunchbase features.

To our features we apply a train-test-split (80/20) and then test out various models, including XGBoost, logistic regression, kNN, SVM and random forest. Random forest results in the highest accuracy of classification on the test set, at 77.45%.

The remainder of our system deals with evaluation and presentation of the models, where we plot out ROC and Evaluation curves and tabulate the accuracy scores for each model. We then append predictions to our original dataset and order them to make the data usable by the VC

customer. In our intended UI, we also show the top contributing features for each founder, most similar founders, and also links to Crunchbase and LinkedIn profiles.

In the intended system, a VC would log into our website and link to a particular founder's Crunchbase profile. Our system would then scrape that profile for the needed features, and apply those to our best model. The output would be a rank for that particular founder, along with top contributing features and a list of similar founders. If the founder fills out our survey, we would also provide a correlation table of which psychological features, if any, are likely to be the top contributors to success for that founder.

FUTURE PLAN

DISTINGUISH INDUSTRY CATEGORIES

After talking with our mentor, Patrick Chung was really focused on one area: technology. As a VC in the Silicon Valley, this is to be expected. Our future goal is to make this program accessible to VC firms that are interested in funding companies that are in more traditional fields. Our next step would be to allow for subcategorization in different areas of interest, and have a more personalized categorization system based on different features that may seem more important.

IMPROVE SUCCESS REPRESENTATION

As a VC, defining success is not that easy: many companies might declare bankruptcy past a Series B round, losing firms money. Therefore, want to build a new model to better clarify what success means for a company based on what stage of funding they are in. That way, VC's can rely more heavily on our algorithm to display higher projection accuracies for companies that they fund/prospectively will fund until the company hits IPO.

GET MORE NON-OBJECTIVE DATA

Currently, we found little to no highly correlated characteristic features to a founder's success. It might have been the small size of our survey data, or asking the wrong questions in our survey. From the limited survey results we possess, we found that many people highly evaluate

themselves for success, making the survey results null from an issue in self-confidence. To make more meaningful characteristic data, we plan to redesign our survey questions or extend the size of our data.

CONTRIBUTIONS

Team Member	Contributions Done
Alex Nakagawa	Data sourcing, testing various independent variables, feature engineering, formed web UI through Flask, cleaned and commented final code, edited and redrafted final report
Anny You	Data scraping from Crunchbase, Merge Data Set, Cleaning Data, Feature Engineering, Building and evaluate different models' accuracy, wrote the final report and edited slides
Dimitrios Hytiroglou	Data scraping from Crunchbase, Crunchbase data merging script, Cleaning Data, Data Manipulation, Feature Engineering, Build/Evaluate various models, Precision-Recall and ROC Curve Plots, AutoSKLearn on AWS, Cleaning and Commenting code
Mert Gurkan	Data scraping (17k profiles) from Crunchbase, Cleaning Data, Feature Engineering, Data Manipulation, Building and evaluating the models, Cleaning and commenting the final code, getting emails of people from hunter.io, Typing the report.
Lorenzo Ong	Design intended UI (with photoshop); create 15 hunter.io profiles to source 1.5k emails. Emailed 1.2k founders. Scraped 17k founder profiles from crunchbase. Merged survey data with crunchbase profiles. Data analysis of survey results.
Surya Sendyl	Data sourcing (Scrapinghub, Crunchbase), Data scraping from Crunchbase, Building/testing various independent variables (Built 'success' variable), Research and create survey, Incorporated survey data into model, Build/evaluate various models

APPENDIX

LIST OF FEATURES AND WHAT WE DID TO EACH OF THEM

Feature Name	Description	What we have one with it
Full Name	Full Name of the person	Dropped
Primary Job Title	It is the titles of the person in that company (Founder & CTO & Product Manager etc)	Since there are lots of different titles (and title combinations) here, we only check whether he/she is founder or not
Gender	Gender of the person	1 for Male, 2 for Female
Number of News Articles	Number of news articles published about this person	Normalization applied
Number of Founded Organizations	Number of organizations this person has founded	Normalization applied
Number of Portfolio Companies	Number of portfolio organizations this person has	Normalization applied
Number of Investments_x	Number of investments this person has	Normalization applied
Number of Partner Investments	Number of partner investments this person has	Normalization applied
Number of Lead Investments_x	Number of lead investments this person has	Normalization applied
Number of Exits_x	Number of exits this person performed	Normalization applied
Number of Events_x	The number of events that person appeared in	Normalization applied
Categories	Category of the company	Since there are lots of different categories here, it was not very helpful for the algorithms. Dropped
Headquarters Location	Headquarter location of the company	Since it is a categorical variable with so many levels, we found the frequencies of each location and change the locations with these values (California being the highest)

Operating Status	Whether the company is operational or not	1 if operational, 0 if not
Founded Date	Founding date of the company	Gives the year of foundation of the company
Closed Date	Closing date of the company (if closed)	Since it essentially gives very similar info with operating status, we dropped it
Company Type	If the company is for profit or non-profit	1 if for profit, 0 is for non-profit
Number of Founders	Number of founders the company has	Normalization applied
Success *(output variable)*	Whether the company has found success or not	1 if successful, 0 if not (successful if raised Series B or higher)

FOUNDER PSYCHOLOGY SURVEY
FIND OUR GITHUB LINK HERE