VC-Robot

Mar Hershenson marita@stanford.edu December 10, 2010

1. Abstract

In the start-up investing community, many think that investing is an art more than a science. Investing decisions from angels and venture capitalists are based many times in intuition (first impressions), simple rules of thumb (e.g., if founder is a graduate student from Stanford I will invest), herd mentality (e.g., if Kleiner-Perkins has invested in this company I should too despite the price), etc. Few investment decisions are based on quantifiable arguments. In fact, most investors do not believe it is possible to come up with a quantitative model for venture investing. The goal of this project is to build a quant "VC Robot" that predicts success of a start-up company based on some measurable features. We show in section 4, that the robot built has significantly higher hit rate than an average venture capitalist.

2. Data Collection

Obtaining a good and large enough data is one of the biggest challenges in this project. Data gathering is complex for various reasons: easily accessible data is skewed (much more data available on successful companies), data can only be retrieved manually (free online databases are not complete) and private data like is either not available or available via expensive subscriptions.

2.1. Training data set

Rather than choosing from any available start-up data, I imposed the following restrictions on the training set:

- 1. Company must belong to the semiconductor industry. The company must be shipping integrated circuits or intellectual property directly dropped in integrated circuits (ICs). I have excluded some neighboring industries such as the electronic design automation industry (EDA) that provides software tools for the design of ICs, semiconductor equipment companies, semiconductor manufacturing companies, etc. Although I plan to lift this restriction in the near future, for now it allows me to have a consistent set of data with similar macroeconomic expectations.
- 2. The companies had to be founded in 1999 or later. This restriction was due to the fact that there is little consistent data online for companies that were founded earlier.
- 3. The company must have raised some money from venture capitalists at some time during their lifetime. Again, this restriction was also due to the fact that there is barely any data accessible for companies that are angel financing, or that died before raising any venture money.
- 4. If the company is still alive and has not had a liquidity event, I have excluded it from the data set. My plan is to use our model to predict whether these companies will be successful.

With these restrictions I have been able to find data on 115 companies. As I created the model, I found that after 70 companies, the training and test error were similar so no more companies were needed. At the moment, I have used the following feature set.

- 1. Total funding received
- 2. Location measured as distance of company headquarters to Santa Clara, CA (Intel headquarters)
- 3. Year it was founded
- 4. Year it exited (either IPO, acquisition or bankruptcy)
- 5. Number of Silicon Valley "Tier 1" venture investors. I define "Tier 1" venture investors that are at least fifteen years old and have more than \$1B in assets under management.

- 6. Number of corporate investors
- 7. One of the founders was CEO at the time of exit
- 8. One of the founders was CTO at the time of exit
- 9. The CEO had a technical background (at least MS or PhD in a technical field and had held some technical role in a previous company)
- 10. The CEO had been at a start-up before
- 11. One of the founders holds a degree from a top university
- 12. Total number of founders
- 13. Total number of investors
- 14. Amount of initial series A investment.
- 15. Time of series A investment.
- 16. Amount of initial series B investment.
- 17. Time of series B investment.
- 18. Amount of initial series C investment.
- 19. Time of series C investment.
- 20. Amount of initial series D investment.
- 21. Time of series D investment.

The quantity we are interested in measuring is the return on investment. I have recorded the total exit price for each company. From that we can create several criteria of success (return on investment larger than some threshold, return compounded yearly, etc.).

2.2. Data Collection Process

Collecting the data was an extremely manual and slow process that has taken me tens of hours. Although I contacted several private database holders and well-known individual investors, no one wanted to provide me access to the data. I also worked with the Stanford Business School Library but unfortunately they also don't have access to data I could use. Towards the end of the project, I got access to VentureSource[1] database which has helped collect some of the data. Additional data was obtained:

- 1. From free online databases Crunchbase[2], VentureBeat[3] and semi-free SiliconTap[4]. The maintenance of these databases is crowd-sourced, i.e., the companies themselves provide the information. As a result, they are fairly incomplete (not all companies are there, not all data of a given company is there, etc.).
- 2. From "Silicon Times" magazine[5]. I found a few free numbers online of a publication that tracks startups in the semiconductor and telecom industries. This was extremely useful as it allowed me to uncover many names of failed companies.
- 3. Using "Wayback machine"[6], a web archive that allowed me to dial back in time and view pages of companies I was interested in.
- 4. Using results of various other internet searches.

3. Data Analysis

Before proceeding to construct a model I took a look at the data. I define success of a company (the "ones") as a return over 5 for a series A investor (27 successes out of 115 data). It assumes that series A investment buys 30% of the company and that that ownership gets diluted 30% by a liquidity event. Exact numbers on internal round valuations are practically impossible to get so we have to make this estimate. This would help an investor decide whether he should invest in the series A of a company. A series A return of 5 is a low return rate over the typical 7-8 years lifetime of a fund but if I increase it to 10, we would have very few successes (only 17 out of 115 data points).

I was able to see some trends on the data that allowed me then to refine the features I collected. Figure 1 shows a set of bar plots analyzing success factors.

- Subplot (1,1) shows successes versus the quality of the CEO at time of liquidity. Quality is defined as the sum of the following features: CEO is founder, CEO holds a graduate technical degree and CEO was at a startup before.
- Subplot(1,2) shows successes versus the quality of the CTO/VP Engineering at time of liquidity. Quality
 is defined as the sum of the following features: CTO is founder, CTO holds a graduate technical degree
 from a top school.
- Subplot(2,1) shows successes versus the number of founders occupying either the CTO or the CEO position at the time of exit. Obviously having the founders remain at the company is critical.
- Subplot(2,2) shows success versus combined CEO and CTO quality.
- Subplot(3,1) shows successes versus qualities of CEO and CTO known at the time of funding (education).
- Subplot(3,2) shows successes versus the number of founders. It appears that one founder only is not as desirable (perhaps suggesting that one needs support starting a company). Also three founders is not ideal (perhaps suggesting odd team dynamics are not desirable).

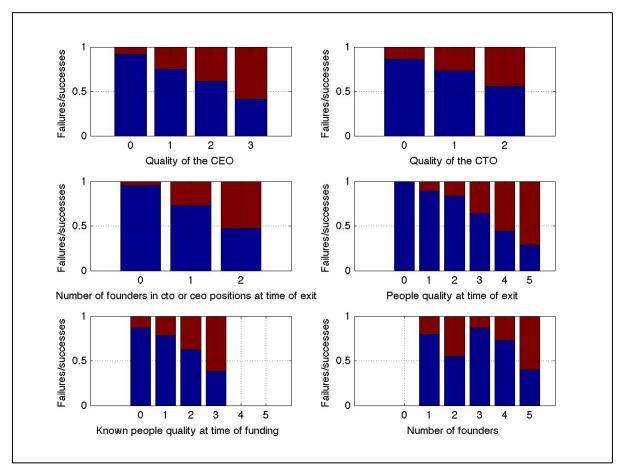


Figure 1. Bar plots analyzing different success factors. Red indicates success, blue failure.

I uncovered other interesting data relations such as low spending rate is desirable, smaller number of investors is desirable etc. I used several x/y plots to see if data was separable in any features. However, the data does not seem separable at first sight so it was hard to extract any information from that.

4. Model

I built a logistic regression model that classifies data into success/failures (whether a return produces a

significant return). Due to the space limitation, I am not showing other experiments such as the return for a later investor.

I trained the model with 80% of the data and I measured the performance on 100% of the data using the feature set in Table 2. I used cross-validation on different 80% subsets of the data and selected the one that provided the best test error. I tried several of the features combination by adding features one at a time and understanding the impact on the overall model error. Also based on some of my initial results, I retook some of the data (collected more info on CEO and CTO as well as the number of founders).

x ₀	Constant term	-1.501
X ₁	Spending rate/mean(spending rate)	3.227
X ₂	Founding date -1999	-0.538
X ₃	Quality of CEO and CTO at funding	1.218
X ₄	Series A investment amount	-0.189
X ₅	Number of founders	1.508
X ₆	Number of founders is even	-2.687

Table 1. Logistic regression model parameters for a series A investor. Note that spending rate is an estimate based on plan and may not be accurately known at time of funding.

The test error was as follows for a logistic regression and SVM are shown in Table 2.

_	Logistic	SVM	
Success Guessed Correctly	(20/27) 74%	(20/27) 74%	
Success Guessed Incorrectly	(7/27) 26%	(7/27) 26%	
Failure guessed correctly	(77/88) 88%	(80/88) 90%	
Failure guessed incorrectly	(11/88) 12%	(8/88) 10%	
Total guessed correctly	(97/115) 84%	(100/115) 87%	
Total guessed incorrectly	(18/115) 16%	(15/115) 13%	

Table 2. Quality of model for series A investor

The model suffers of high bias and more features are needed to generate lower error model. Unfortunately, the time involved in collecting these features is large and I was not able to collect more information.

I also created a regression model for the ROI as function of the variables in Table 1. I obtained a large residual when I fit the ROI number directly. Based on the residuals I obtained, I realized that if I fit the ROI^ α , where 0< α <1, I would obtain a better fit. With a simple optimization loop, I found that α =0.4 gave me the mostnormal like residuals. I made a QQ-plot of the residuals in Figure 2. As one can see, residuals follow a normal distribution quite well. The model I obtained is:

$$ROI^{0.4} = 0.85 + 1.214x_1 - 0.106x_2 + 0.24x_3 - 0.008x_4 - 0.047x_5 - 0.267x_6$$

I run out of space to show the model for a later stage investor but the process is similar (just additional features such as series B amount, time between A and B rounds). In any case, the series A decision is the most difficult (less features available).

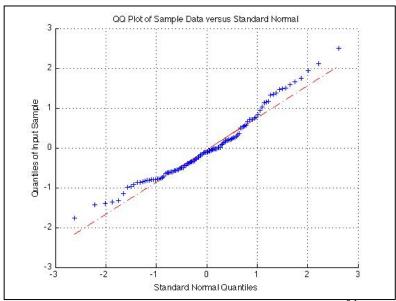


Figure 3. QQ-plot for residuals of linear regression of ROI^{0.4}

5. Conclusions and Future Work

It is surprising that such a simple model can guess failures correctly in 88% of the cases and successes in 74% of the cases. This is without taking into consideration any market factors such as whether the company has a key technology, differentiated products, revenue, number of employees, competitors, good macro-economic conditions etc. I suspect that taking this later features into account would greatly improve the model error. The success rate of a venture capitalist is very low, only 23% (27/115). I believe VC robot could help assist bring that rate up.

The model still has high bias and more features are needed to achieve better modeling results. I plan to work on this. Additionally I plan to estimate revenue stage at the time of financing. This will help with the model level accuracy for a late stage investor.

One of the conclusions from simply collecting the data is that if the initial founder CEO is still the CEO at the time of a liquidity event, the chances of success are much higher. As a result the most important question an investor must ask himself is whether the founder can remain CEO. To assist with this, I plan to continue to gather additional features that would allow to build a simple model to predict whether a founding CEO can remain CEO. I plan to gather features that are relate to the characteristics of the CEO (education, work history, relation to founding team, etc).

I have just scratched the surface of this exciting project. However it is exciting and I have found many people interested in continuing the work.

6. References

- [1] VentureSource by Dow Jones VentureOne, http://www.venturesource.com/
- [2] Crunchbase by Techcrunch (now AOL), http://www.crunchbase.com/
- [3] Venturebeat, http://venturebeat.com/
- [4] SiliconTaps, by SocalTech LLC, http://www.silicontap.com/
- [5] "Silicon Times" by Pinestream Communications http://www.pinestream.com/
- [6] "Wayback Machine", http://www.archive.org/web/web.php/