VCPredict: A Predictive Model of Venture Capital Investments

Akilesh Potti AVP39@CORNELL.EDU
Siddharth Reddy SGR45@CORNELL.EDU

Abstract

Venture capital firms are a major source of funding to start-up companies that would otherwise be unattractive investments to traditional investors and financial institutions. A wealth of public data is available about start-ups and VC firms through platforms like Crunchbase and AngelList. These data are useful for understanding the underlying structure of the venture capital space. We develop a predictive model of venture capital investments using various characteristics of start-up companies and VC firms, motivated by the following possible applications: (1) optimizing the fundraising process for start-up companies, and (2) optimizing the investment process for VC firms. As a result of our analysis, we find that

1. Introduction

1.1. Motivation

The goal of our analysis is to create a predictive model of venture capital investments, specifically our model will predict whether or not a venture capital firm will invest in a start-up company, and if so, the specific dollar amount. Start-up companies could use this model to optimize their efforts in search of VC funding; rather than seek investment from VC firms unlikely to invest in companies with their characteristics, they could focus their time on VC firms that are active in their industry, geographic region, or problem space. VC firms could use this model to optimize their search for new start-up companies to add to their portfolios, and to develop data-driven characterizations of their investment philosophies to use as marketing tools.

2. Problem Setup

We use unsupervised learning methods to discover structure in the Crunchbase data set, then use those results to construct a feature set for the following supervised learning tasks:

- Prediction of whether or not a venture capital firm will make future investments in a start-up company
- Conditional on the outcome of the previous task, prediction of the dollar amount invested

3. Data Sources

- Crunchbase: `http://www.crunchbase.com/`[1]
- AngelList: `https://angel.co`

We use a combination of datasets as no one dataset truly captures all of the pertinent information for making the analysis. Crunchbase is our primary source, with over 80,000 data points and 13 features. AngelList offers start-up data regarding founders, board members, employees, etc. that the Crunchbase data set lacks.

4. General Approach

In our unsupervised learning phase, we cluster VCs by average investment size, targeted industries, and investment stage using heuristic-based clustering methods.

For our prediction of binary VC investments (i.e. will a VC invest in this start-up?), we use a logistic regression model.

For our prediction of specific dollar amounts invested by VCs, we use multiple linear regression and a regression tree model.

5. Methods

6. Unsupervised Learning Tasks

We expect inherent structure in the diversity of VC firms. Some firms are early-stage, seed funds that invest primarily in young start-ups. Others are mid-stage, late-stage, tend to make large investments (¿40 million USD), tend to make small investments (¡10 million USD), etc. In order to

ORIE 4740 Data Mining, Final Project

[1] http://info.crunchbase.com/about/crunchbase-data-exports/

characterize this structure according to available data fields such as number of Series A rounds made, number of private equity investments made, and average investment size, we apply the k-means clustering algorithm to the VC firms represented in the Crunchbase data set. Prior knowledge suggests the existence of three major clusters of firms, so we set K=3 and produce the following clusters.

Characteristics of VC firms such as distribution of investments across various funding rounds, geographic regions, and start-up industries are typically correlated. As such, we expect variation in VC firms to be characterized by relatively few factors. Principal component analysis of the VC firms represented in Crunchbase reveals two main principal components that account for much of the variation in VCs.

In practice, certain VC firms tend to invest in the same start-up companies as a result of similar interests, investment strategies, domain expertise, etc. The graph is defined as follows: VC firms are represented as nodes, and edges are drawn between VC firms with weights proportional to the number of companies they have invested in together. The following ranking of VC firms is constructed using various graph centrality measures. A simple web page has also been setup to allow users to explore the co-investment graph.

7. Supervised Learning Tasks

In our prediction of binary VC investments, we use the following feature set:

- Age (days)
- Total funds raised to date (USD)
- Binarized company industry (e.g. biotech, ecommerce)
- Binarized company location (country)
- Binarized types of past investors (e.g. has received investments from early-stage, mid-stage, and/or late-stage funds)
- Average centrality of founders, advisors, board members, current and past investors in the co-investment graph
- Number of employees

Table 1. Centers produced by k-means clustering of VC firms.

	Late-stage	Mid-stage	Early-stage
angel	0.017	0.023	0.324
crowdfunding	0	0.0001	0.0001
other	0	0.012	0.005
post.ipo	0.008	-0	0
private.equity	0.842	0.012	0.004
series.a	0.067	0.194	0.449
series.b	0.042	0.214	0.075
series.c.	0.017	0.253	0.031
venture	0.008	0.293	0.112
total	4.170	59.500	26.800

Table 2. Importance of components

	Proportion of Variance (> 0.01)
Comp.1	0.861
Comp.2	0.075
Comp.3	0.042

Table 3. Importance of components

	Proportion of Variance (> 0.1)
Comp.1	0.362
Comp.2	0.276
Comp.3	0.184
Comp.4	0.117

Table 4. Top 10 VC firms by Pagerank (eigenvector centrality) in the co-investment graph

	Pagerank	In.degree.Centrality	C
SV Angel	0.025	0.385	
Intel Capital	0.014	0.395	
Accel Partners	0.013	0.359	
First Round Capital	0.013	0.306	
Kleiner Perkins Caufield & Byers	0.012	0.332	
New Enterprise Associates	0.012	0.346	
Andreessen Horowitz	0.011	0.271	
Draper Fisher Jurvetson (DFJ)	0.011	0.346	
Sequoia Capital	0.010	0.281	
Greylock Partners	0.010	0.294	

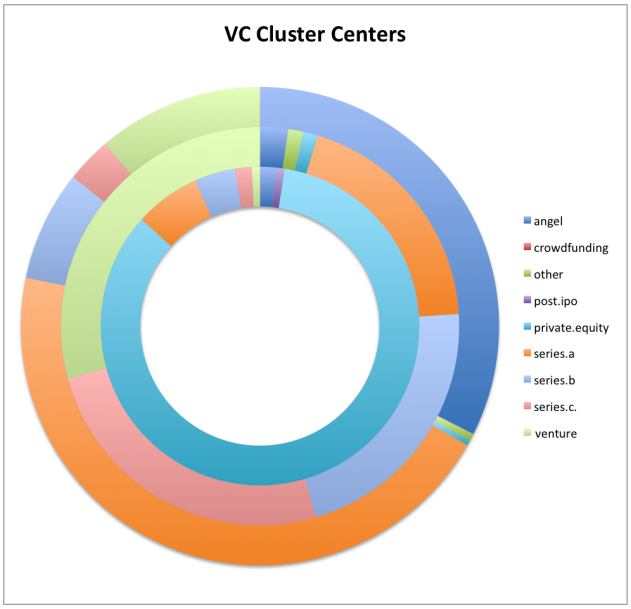

Figure 1. Centers produced by k-means clustering of VC firms. Outer ring, middle ring, and inner ring correspond to early-stage, mid-stage, and late-stage funds respectively.

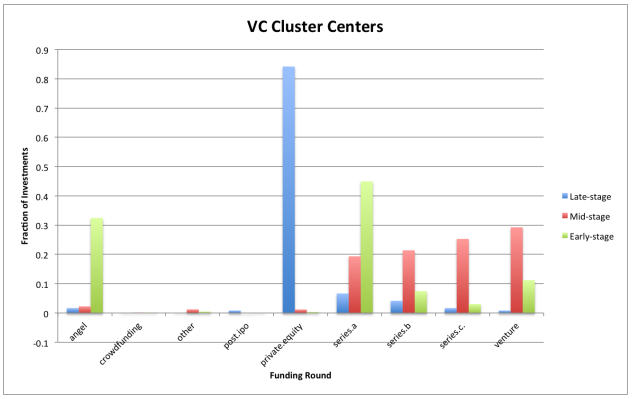

Figure 2. Centers produced by k-means clustering of VC firms.

8. Results

8.1. Clustering

8.2. Principal Component Analysis

8.3. Co-investment Graph

9. Conclusions

10. Bibliography

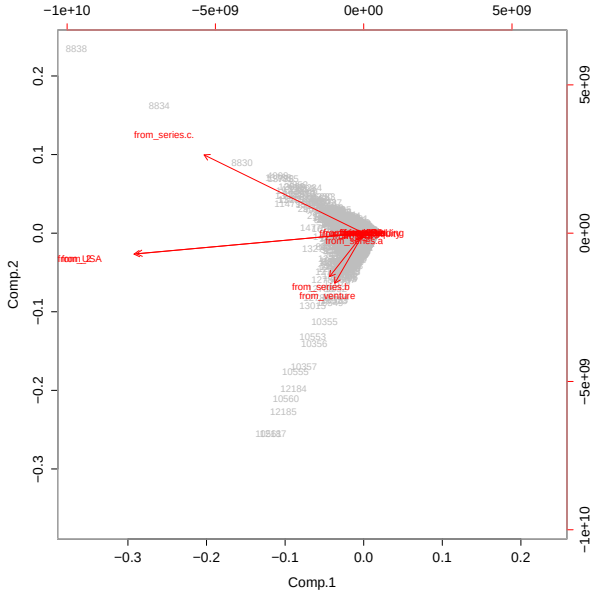

Figure 3. PCA of VC-Startup investments based on how much the start-up has raised from different types of VCs in the past, age, and total funds raised to date

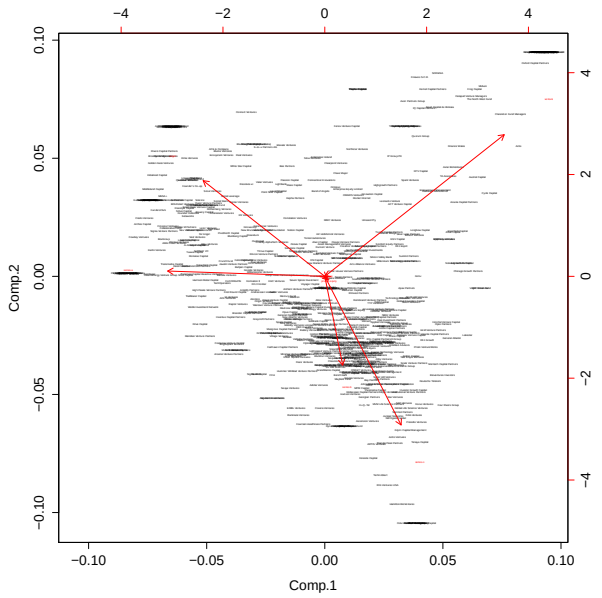

Figure 4. PCA of VC firms based on distribution of investments across different funding rounds (e.g. Series A/B/C, Angel)

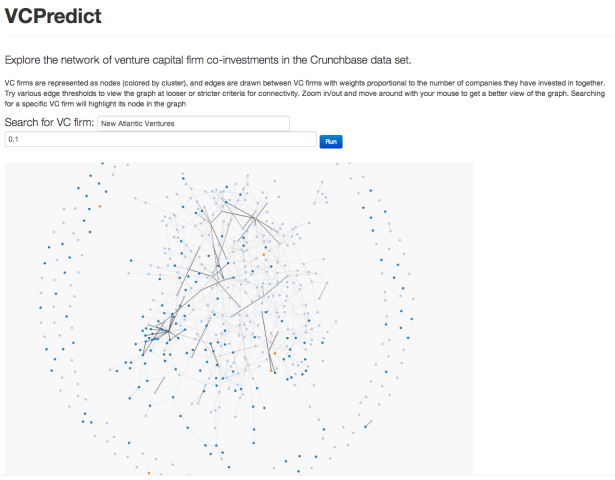

Figure 5. Screenshot of interactive web page that allows users to explore the VC co-investment graph