Springboard Foundations of Data Science

Capstone Project “Investigating Returns on Seed-Stage Capital in the US”

Milestone Report

*Background*

Angel Investing is the business of providing capital to early-stage growth companies. The capital is typically the first outside money the company takes; they offer in exchange some portion of their company’s equity. The company uses investor’s money to fund the growth of their business. As the company grows and increases proportionately in value, additional rounds of funding may be raised, fueling additional growth. Ultimately, investors (whether Seed- or Later-stage) will crystallize their investment and generate returns via an exit. An exit results from the sale of the company to a strategic (larger, corporate) buyer or via IPO. Companies which do not proceed to exit will typically either cease operations/close, in which case investor’s money is lost and return is zero, or continue to operate, where a non-zero value is achieved but the investment remains illiquid and real value unrealized.

Unlike other asset classes, there is a dearth of information around outcomes (% success/failure) and returns ($ invested, $ returned) in this space. Information that is available is largely self-reported creating a potential source of bias/positive skew. Two companies (Pitchbook and Crunchbase) collect and make available (fee basis) proprietary subsets of data. Crunchbase in 2014 published for the public a sample set of their then-current (2014) data.

*Introduction*

My study seeks to use this Crunchbase data set to investigate returns in this space. The data set includes four separate tables which separately describe funding rounds, companies, acquisitions and people (investors). For my purposes, acquisitions (specifically the “acquisition price” field which describes the value of the exit, which becomes the numerator in the returns equation) and funding rounds (specifically total funding raised, returns denominator) will be the most useful.

Specific lines of inquiry include:

1. Describing the data scarcity: of the companies listed in this set, what proportion are missing data critical to understanding outcomes (status field from companies database) and among those how many are missing data critical to understanding returns (acq’n price/total funding).
2. What do returns look like for those companies that are properly described (outcome, funding, price)? Is this a sufficiently large sample from which to draw conclusions about the population?

describe (using statistical analysis where appropriate) what returns in the Angel space My analysis is based on the publicly-available Crunchbase data

* An introduction to the problem (based on your earlier Capstone submissions).
* A deeper dive into the data set:
  + What important fields and information does the data set have?
  + What are its limitations i.e. what are some questions that you cannot answer with this data set?
  + What kind of cleaning and wrangling did you need to do?
* Any preliminary exploration you’ve performed and your initial findings.
* Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?