Capstone Project 1: Milestone

Proposal

For my First Capstone Project I decided to use Crunchbase data to build a model to see if it was possible to predict whether or not a company would be funded in the following year based on data at a given time.

Problem Statement

It is difficult to truly understand what it takes to start a company and successfully raise capital as there are an endless amount of variables many of which cannot accurately be quantified. Using data provided by Crunchbase I hope to gain some level of understanding of the startup funding.

Why This Topic?

I chose this topic as I have a strong interest in startups. I found the crunchbase data very interesting and wanted to build some sort of model based on what I was learning in the course. In the end I decided to see if it was possible to predict whether a start up would get funding in a specific time frame based on round, company, investor, and other information.

Startups, funding, and venture capital are growing topics in the US and around the world with more and more startups getting funded everyday. I have always loved business and am constantly trying new things and starting new projects but the more I thought about starting or joining a startup the more I realized that I had no clue about the market, risks, and overall viability of getting an idea, raising a capital and being successful. For this reason, I chose to explore this as the topic for my first capstone to try to gain some insight into the startup & funding world.

Procedure

1. Data Cleaning & Wrangling
2. Exploratory Data Analysis
3. Inferential Statistics
4. Machine Learning
5. Model Evaluation
6. Feature Analysis

Data Collection Summary

The data I will be using for this project was taken from a GitHub repo. We did have access to the Crunchbase Open Dataset, however, the data only went up to 2013. We were able to find a GitHub repo of someone with access to a more complete version of the data (up to 2015) which we compared to the open data to test for accuracy and decided that the more complete dataset would be better to work with.

Although the data is not up-to-date, given access to the full dataset we would be able to run it through the model in the same way to get an updated version of the analysis as the variables are exactly the same.

As we did pull the data from a GitHub repo, no complicated scraping or data extraction was required for the sake of this project so no interesting display of abilities here. However, cleaning and wrangling of the data was extremely time consuming due to the amount of data used and the variability in the data.

Data Wrangling Summary

The data wrangling process was quite extensive as we needed to accumulate all of the data into features which could be used to train my Random Forest and Gradient Boosting model. This involved transforming all non-numeric data to numeric features as well as filling in missing values and creating new features to help summarize the data more effectively.

Raw Data

The data was broken up into the following dataset:

	Shape	Numeric Data	Text Data	Datetimes
Investment/Rounds	(212810, 18)	1	16	1
Companies	(51146, 14)	2	9	3
Organizations	(606064, 16)	0	16	0
People	(605630, 15)	0	15	0
Acquisitions	(18968, 18)	1	15	2
IPOs	(1259, 13)	2	8	3

Basic Cleaning Steps:

1. Rename Column names
2. Indicate Null Values
3. Deal (Impute/Drop) with Null Values
4. Strings to Numbers/Datetimes
5. Export Clean Data

Basic Wrangling Steps

1. Create format of dataframe with IDs, Dates, etc
2. Add or summarize numeric values and add to dataframe
3. Text variables as dummies, summarized, or vectorized
4. Export Processed Data

Processed Dataset

Using the data we want to create a single final dataframe with features that will summarize relevant information. As we want to predict whether a startup will be funded in a given time each row will represent a quarter of a specific company after the first round of funding. The target variable will note whether that company raised a round in the preceding year.

As we are planning to use a Random Forest Classifier for this model we will need to represent all non-numeric variables as numeric variables. The random forest is able to ignore noise from less important features but we will do our best to minimize the number of features as we want to be as computationally efficient as possible.

The wrangling section is broken up in the same way as the dataset themselves and pull key features from each dataset summarizing them in the format we decided on for the final table. An additional section was added as well which was Macro variables as we want to see how the market as a whole is performing as we predict this has an impact on funding.

Exploratory Data Analysis Summary

The EDA for the project was broken up into three main sections.

1. Investor Analysis
2. Investment Analysis
3. Category Analysis

As there is a wide range of data it took a large amount of exploratory data analysis (EDA) to really get a feel for the entirety of the data and the important features. We split the EDA into 3 separate notebooks to help with organization. The notebook provided includes the analysis of the features finally selected. This being my first big project I did a lot of extra exploration both to be sure that I got everything from the data and to just practice different visualization and analysis tools learned in the course.

Pt 1. Initial Visualizations

Basic Visualizations were made during the data wrangling of the project to try to find interesting features to add to the model. During this initial exploration I found myself asking different questions about the data which were then grouped into the EDA notebooks that are found in the project folder. These EDA notebooks contain the more in-depth EDA and Inferential Statistics work.

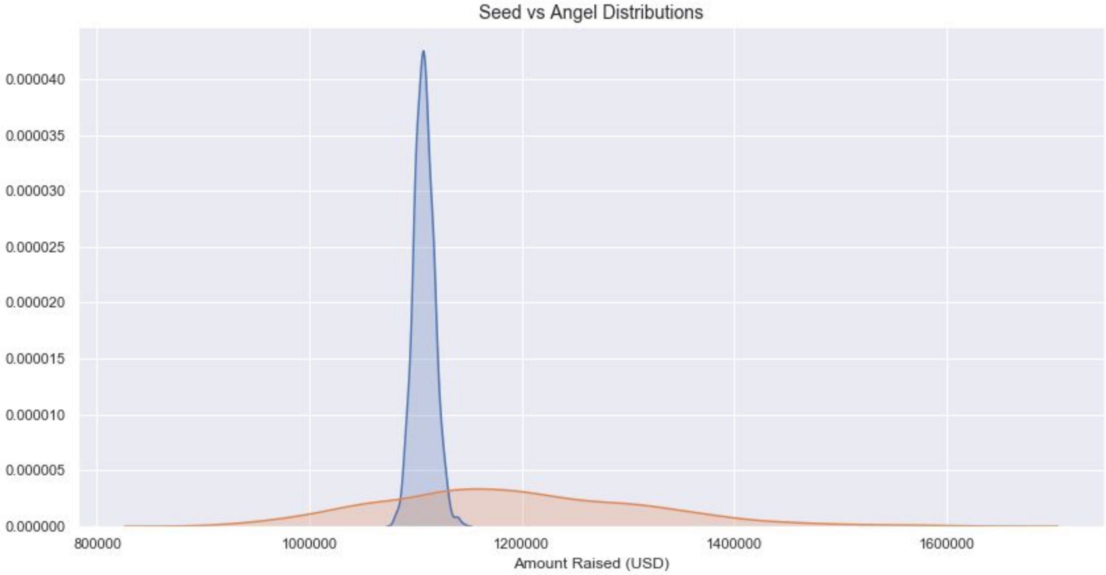
Pt 2. Inferential Statistics

The EDA Notebooks consist of more detailed visualizations and the inferential statistics. For more clear organization this section was broken up into the three notebooks: Investor, Investment and Category Analysis. The EDA was based off questions that came up during the initial visualization and exploration of the data when choosing features and building the dataset for the machine learning model. The Analysis and Questions are organized as follows:

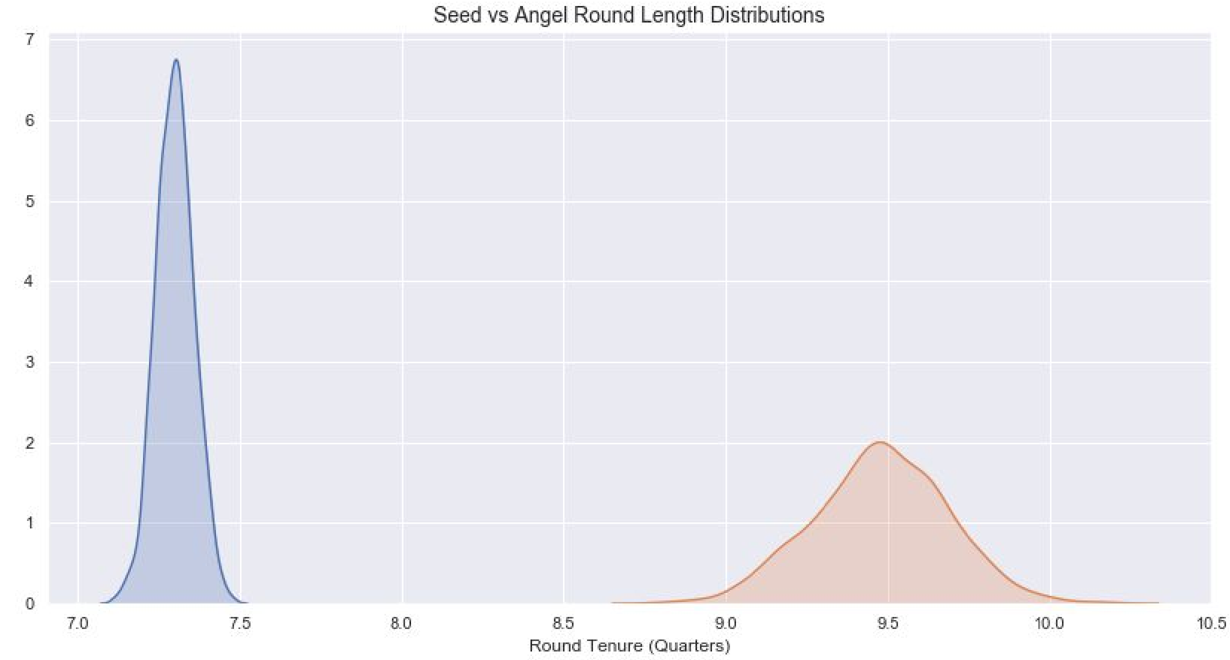
1. Investment Analysis

a. **Do Different Rounds Raise Different Amounts?**
b. **Do Different Round Types Have Different Tenures?**

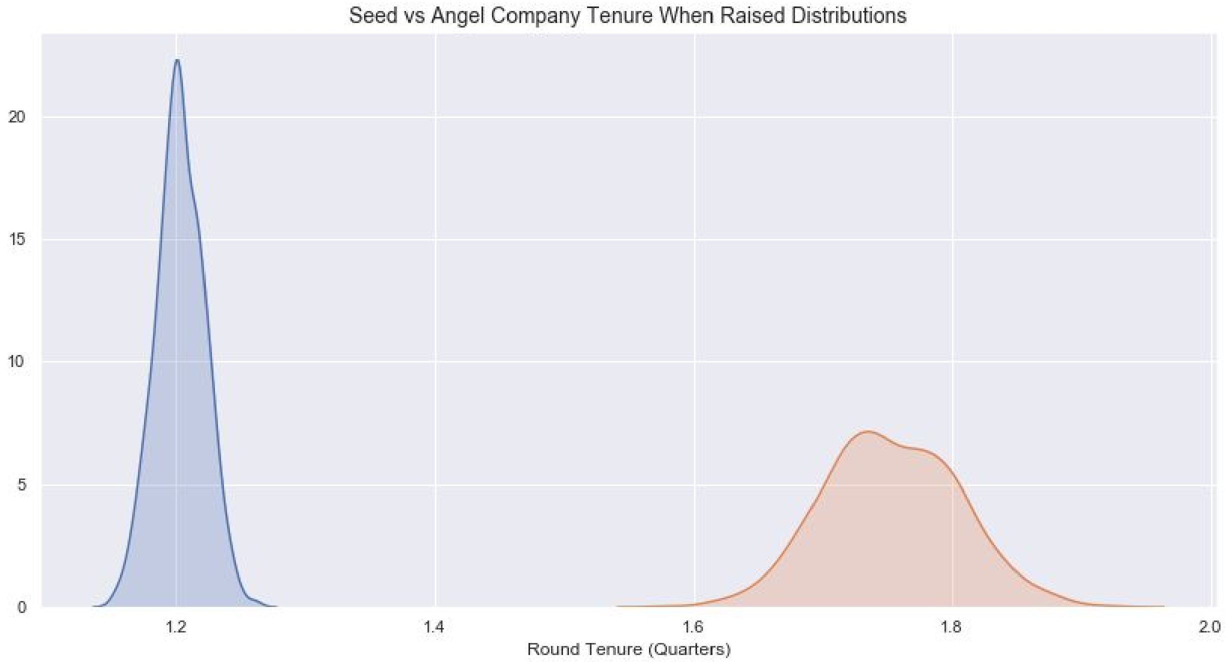
Seed Vs Angel Funding Distribution (Other Rounds In Notebook)



Seed vs Angel Round Lengths (Other Rounds In Notebook)



Seed vs Angel Company Tenure at Funding (Other Rounds In Notebook)



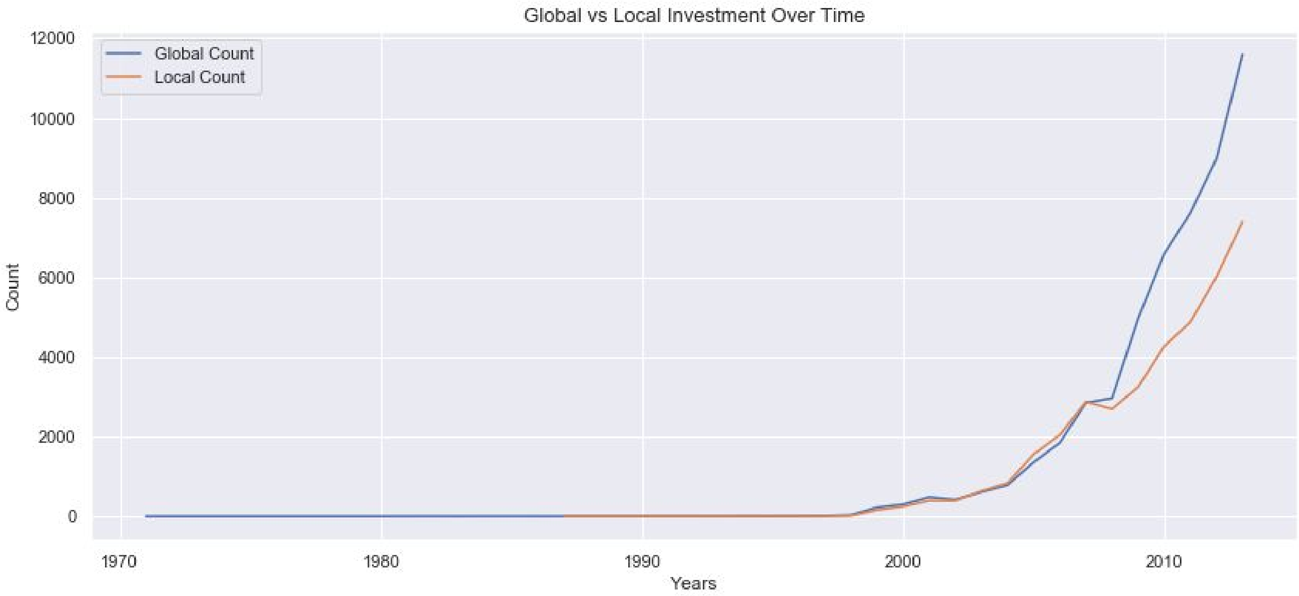
Investment Analysis: Key Observations

- **Amount Funded by Angels Varies More Than Seeds**
- **Later Round Typically Have Higher Variance in Amount & Tenure. (Not Shown)**
- **Venture Round Lengths Tends to be between 10 -15 Quarters. (Not Shown)**
- **Equity Funding is Greater Than Debt Before IPOs but Less After. (Not Shown)**

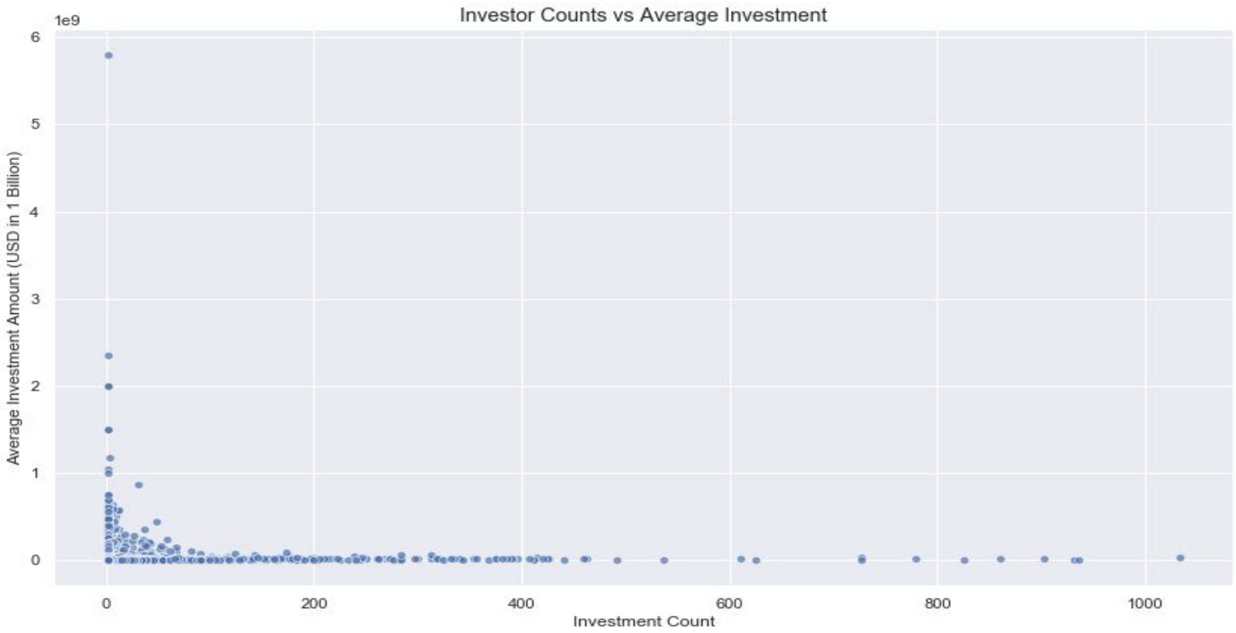
2. Investor Analysis

a. **Do Global Investors Fund Differently Than Local Investors?**
b. **Do Investors Fund Differently?**
   i. **Quantity vs Quality?**
   ii. **Focus on Specific Rounds?**

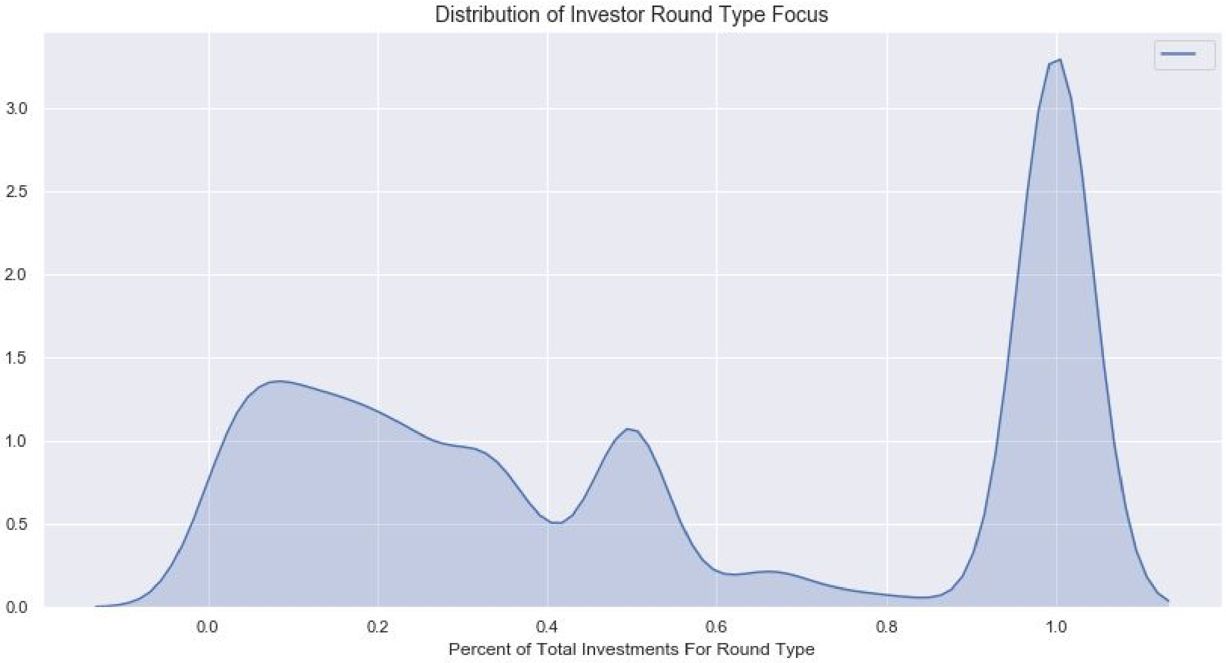
Analyzing Globalization of Investing (Full Analysis In Notebook)



Analyzing Difference in Investment Strategies (Full Analysis in Notebook)



Do Investors Focus On Specific Round Types (Full Analysis In Notebook)

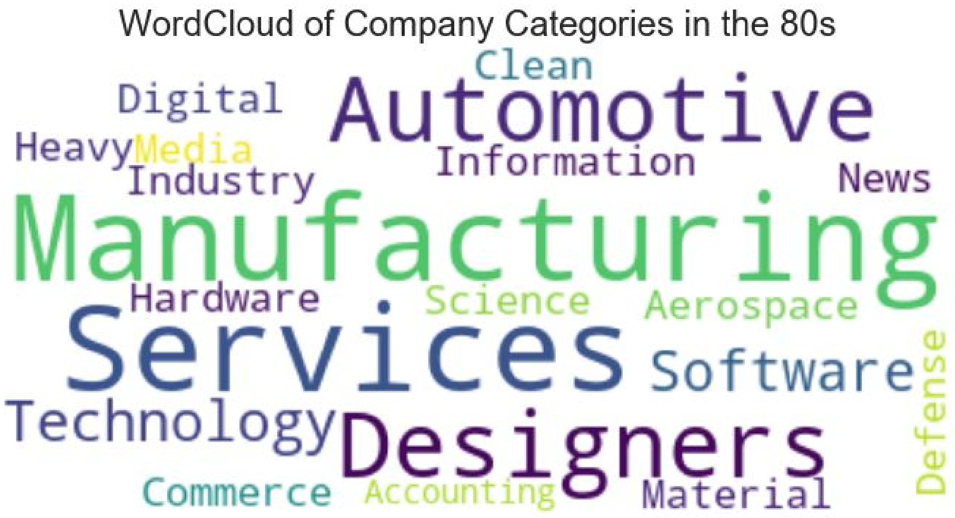


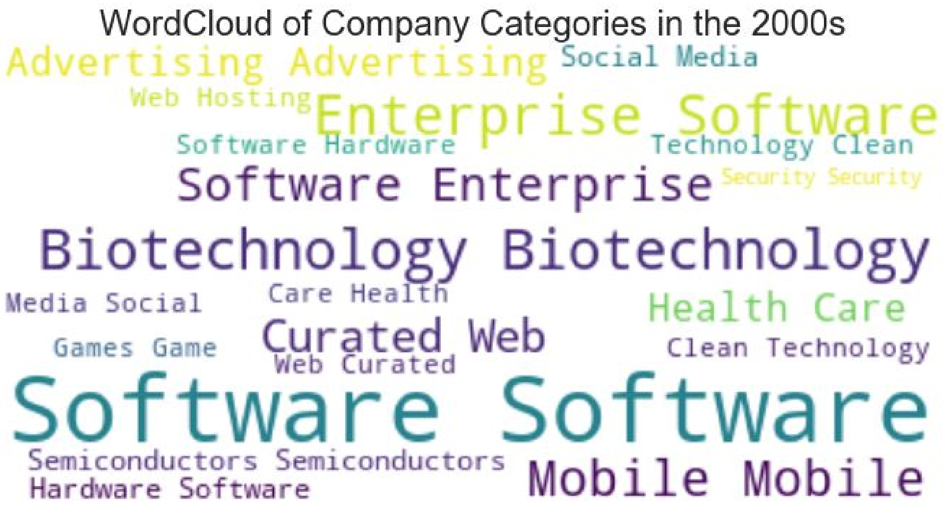
Investor Analysis: Key Observations

- **There are more gaps in data of global investors**
- Difference in invested amount of global and local investors is **NOT** statistically significant
- Investors do NOT focus on just **many small** or **few large** investments.
- **Most investors focus on more than one round type.**
- **Seed investors** tend to **focus more** on seed funding.

3. Category Analysis

a. How has funded categories changed in the past 30 years?





Category Analysis: Key Observations

- **1980s:** Manufacturing, Services, Designers, Automotive, Technology
- **1990s:** Software, Technology, Curated Web, Service, Internet
- **2000s:** Software, BioTech, Enterprise Software, Mobile, Curated Web
- **2010s:** Software, Enterprise Software, Mobile, Curated Web, Commerce

References and Resources

GitHub Repo
https://github.com/daikiminaki/Springboard/tree/master/Capstone_Project_Crunchbase_Funding