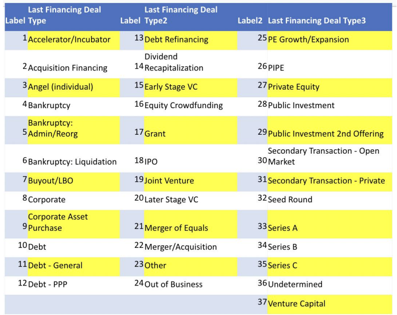
Capstone

Zhirui Xiong
Graduating in May 2023
Capstone Advisor: Professor Pattanayak
Github link: https://github.com/zx106/Capstone-Project

Purpose of Capstone Project

My research is inspired by Startup Cartography Project (SCP), which measure entrepreneurial ecosystem statistics for the United States from 1988 to 2016 by building a predictive analytics approach to estimate the probability of "extreme" growth (IPO or high-value acquisition) at or near the time of founding.

This capstone project addresses the question: what is the probability of biotech startups of the Greater Boston area reaching success determined by their qualitative features such as industry sector, name-based observables, and legal information? The success of a biotechnology startup is defined as the company reaching the IPO, later stage venture capital funds, and other funds received at the late stage startups. These financial instruments will have the positive implications on the long-term enterprise value evaluation of the public listed biotech firms.



Label	Last Financing Deal Type	Label	Last Financing Deal Type2	Label2	Last Financing Deal Type3
1	Accelerator/Incubator	13	Debt Refinancing	25	PE Growth/Expansion
2	Acquisition Financing	14	Dividend Recapitalization	26	PIPE
3	Angel (individual)	15	Early Stage VC	27	Private Equity
4	Bankruptcy	16	Equity Crowdfunding	28	Public Investment
5	Bankruptcy: Admin/Reorg	17	Grant	29	Public Investment 2nd Offering
6	Bankruptcy: Liquidation	18	IPO	30	Secondary Transaction - Open Market
7	Buyout/LBO	19	Joint Venture	31	Secondary Transaction - Private
8	Corporate	20	Later Stage VC	32	Seed Round
9	Corporate Asset Purchase	21	Merger of Equals	33	Series A
10	Debt	22	Merger/Acquisition	34	Series B
11	Debt - General	23	Other	35	Series C
12	Debt - PPP	24	Out of Business	36	Undetermined
				37	Venture Capital

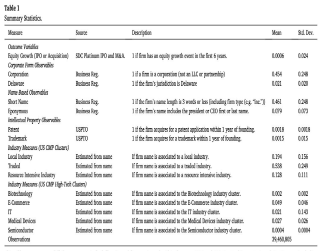
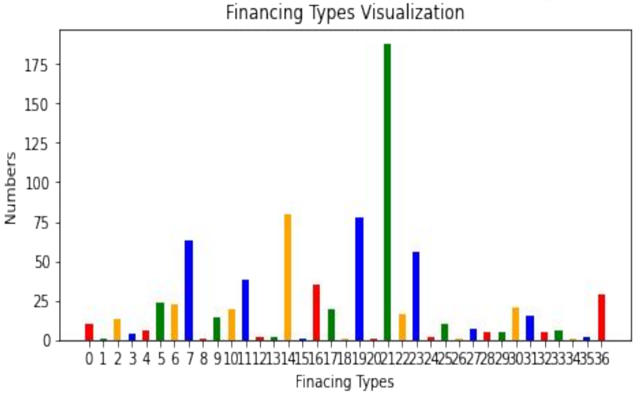


Table 1
Summary Statistics.

Measure	Source	Description	Mean	Std. Dev.
Outcome Variables				
Equity Growth (IPO or Acquisition)	SDC Platinum IPO and M&A.	1 if firm has an equity growth event in the first 6 years.	0.0006	0.024
Corporate Form Observables				
Corporation	Business Reg.	1 if a firm is a corporation (not an LLC or partnership)	0.454	0.248
Delaware	Business Reg.	1 if the firm's jurisdiction is Delaware	0.021	0.020
Name-Based Observables				
Short Name	Business Reg.	1 if the firm's name length is 3 words or less (including firm type (e.g. "inc."))	0.461	0.248
Eponymous	Business Reg.	1 if the firm's name includes the president or CEO first or last name.	0.079	0.073
Intellectual Property Observables				
Patent	USPTO	1 if the firm acquires for a patent application within 1 year of founding.	0.0018	0.0018
Trademark	USPTO	1 if the firm acquires for a trademark within 1 year of founding.	0.0015	0.015
Industry Measures (US CMP Clusters)				
Local Industry	Estimated from name	If firm name is associated to a local industry.	0.194	0.156
Traded	Estimated from name	If firm name is associated to a traded industry.	0.538	0.249
Resource Intensive Industry	Estimated from name	If firm name is associated to a resource intensive industry.	0.128	0.111
Industry Measures (US CMP High-Tech Clusters)				
Biotechnology	Estimated from name	If firm name is associated to the Biotechnology industry cluster.	0.002	0.002
E-Commerce	Estimated from name	If firm name is associated to the E-Commerce industry cluster.	0.049	0.046
IT	Estimated from name	If firm name is associated to the IT industry cluster.	0.021	0.143
Medical Devices	Estimated from name	If firm name is associated to the Medical Devices industry cluster.	0.027	0.026
Semiconductor	Estimated from name	If firm name is associated to the Semiconductor industry cluster.	0.0004	0.0004
Observations			39,460,805	

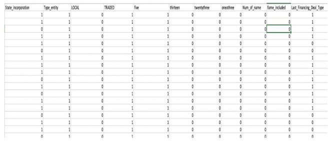


Data Collection and Processing

I have compiled a dataset by processing and merging data from two sources: Pitchbook and Orbis. The dataset is downloaded separately from Orbis and Pitchbook by selecting specific definitions of sectors and NAICS industry code related to biotech. The NAICS industry codes were then organized into subcategories defined by US Cluster Project: Biopharmaceuticals, Distribution and Electronic Commerce, Downstream Chemical Products, Downstream Metal Products, Education and Knowledge Creation, Information Technology and Analytical Instruments, Insurance Services, Medical Devices, Upstream Chemical Products, and Local Health Services.

Although the entire dataset includes companies from 5 major sectors of healthcare (biotech, research and lab, insurance, healthcare institutions, and pharmaceuticals), this capstone project uses only a subset of data (only companies from the biotechnology sector). The biotechnology sectors include diagnostic analytics, biotechnology, biomnaufature equipment, and biotech laboratory. Replicated and missing data have been removed manually – for the next step, a python algorithm should be created to double-check the anomalies. I have also collected data for patent from PatSnap, trademark registration from Massachusetts Business Registration and IPO information from SDC, however the data then was observed to be irrelevant to the purpose of this research.

734 Rows (Data) X 11 Columns (Variables)



State_incorporation	Type_entity	LOCAL	TRADED	five	thirteen	twentythree	oneothree	Num_of_name	fame_included	Last_Financing_Deal_Type
1	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
0	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	0
0	1	0	1	1	0	0	0	0	0	0
1	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
0	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
0	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	1
0	1	0	1	1	0	0	0	0	0	0
1	1	0	1	1	0	0	0	0	0	0
1	1	0	1	1	0	0	0	0	0	0
0	1	0	1	1	0	0	0	0	0	0

Statistical Model

This capstone project uses a binary logistic regression and binomial generalized linear models to investigate the relationship between biotech startups reaching financial success and its qualitative features. These features include name-based observables (such as the number of words in the name of the company and many more), industry sector defined by NAIC codes, and legal information (such as states of incorporation and type of entity).

In the binary logistic regression model, variables included are incorporated state, firms with national networks, five as biopharmaceutical cluster, thirteen as education and knowledge creation, twentythree as information technology and analytical instruments. The variables indicating local firms are omitted for comparison to the national firms, while industry cluster local health services as noted as oneothree is omitted for comparison with other non-omitted industry clusters. Type of entity is omitted from the model due to a few firms not being corporate types. The dependent variable are based on the last financing deal type, as 1 indicating financing types being later stage venture capital fund, IPO, and other funds for later stage growth firms and 0 indicated other financial types.

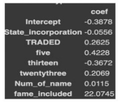
Explanations on Binary Logistic Regression Model

The regression models using a complete dataset show a similar pattern as the analysis in the published research paper Start-up Cartography Project, Measuring and mapping entrepreneurial ecosystems, where having a short company name, having a national business network as a firm, and being in the biotechnology and information analytical technology sector can increases the odds ratio of reaching financial success. Opposite to the previous research, incorporation in Delaware and being in sector of knowledge creation (research lab) can decrease the odds ratio of reaching the success. Surprisingly, if the name of firm includes the name of research institutions from Greater Boston area (such as MIT, Tufts, Harvard), then the likelihood of reaching success jumps enormously.

If a firm is not incorporated in Delaware, having only local business network, being in local healthcare service, having words of the firm's name greater than three but not including name of research institutions from Greater Boston area (such as MIT, Tufts, Harvard) can lead to decrease in the likelihood of reaching success at 1-1/(1+exp(-0.3878))=60%.

Unfortunately for binary logistic regression model, all coefficients are statistically insignificant at the significant level of 0.05. Because of statistical insignificance from binary logistic regression model, this capstone project continues to explore an another regression model: Binomial Generalized Linear Model.

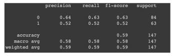
Coefficients and Binary Logistic Regression Model



	coef
Intercept	-0.3878
State_incorporation	-0.0556
TRADED	0.2625
five	0.4228
thirteen	-0.3672
twentythree	0.2069
Num_of_name	0.0115
fame_included	22.0745

Odds Ratio of Reaching Financial Success =
-0.3878
-0.0556*State Incorporation Being Delaware
+0.2625*National Firms
+0.4228*Biopharmaceuticals
-0.3672*education and knowledge creation
+0.2069*information technology and analytical instruments
+0.0115*number of name below three words
+22.0745*name of research institution included in the firms' name

Accuracy of the Regression Model Prediction:
60%



```
              precision    recall  f1-score   support

           0       0.64      0.63      0.63        84
           1       0.52      0.52      0.52        63

    accuracy                           0.59       147
   macro avg       0.58      0.58      0.58       147
weighted avg       0.59      0.59      0.59       147
```

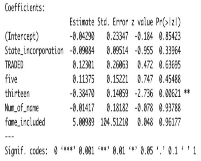
Explanations on Binomial Generalized Linear Regression Model

The regression models using a complete dataset show a similar pattern as the analysis in the published research paper Start-up Cartography Project, Measuring and mapping entrepreneurial ecosystems, where having a short company name, having a national business network as a firm, and being in the biotechnology sector can increases the odds ratio of reaching financial success. Opposite to the previous research, incorporation in Delaware and being in sector of knowledge creation (research lab) can decrease the odds ratio of reaching the success. Surprisingly, if the name of firm includes the name of research institutions from Greater Boston area (such as MIT, Tufts, Harvard), then the likelihood of reaching success jumps enormously.

If a firm is not incorporated in Delaware, having only local business network, being in local healthcare service or information analytical industry sector, having words of the firm's name greater than three but not including name of research institutions from Greater Boston area (such as MIT, Tufts, Harvard) can lead to decrease in the likelihood of reaching success at 1- 1/(1+exp(-0.0429))=45%.

Happily for binomial generalized linear regression model, the decreased likelihood of financial success from being in sector of knowledge creation (research lab) is statistically significant at the significance level at 0.01. This result shows that research labs are less likely to be involved in capital game to reach its operational success. Many literatures suggest that research labs has their funding resources from the government or corporations, so the lab might less likely to seek fund in the form of public stocks.

Coefficients and Binary Generalized Linear Regression Model



```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.04290    0.23347  -0.184  0.85423
State_incorporation -0.09084    0.09514  -0.955  0.33964
TRADED               0.12301    0.26063   0.472  0.63695
five                 0.11375    0.15221   0.747  0.45488
thirteen            -0.38470    0.14059  -2.736  0.00621 **
Num_of_name         -0.01417    0.18182  -0.078  0.93788
fame_included        5.00989  104.51210   0.048  0.96177
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Odds Ratio of Financial Success =
-0.0429
-0.09084*State Incorporation Being Delaware
+0.12301*National Firms
+0.11375*Biopharmaceuticals
-0.38470*education and knowledge creation
-0.01417*number of name below three words
+5.00989*name of research institution included in the
firms' name

Modification of Binomial Generalized Linear Regression Model

According to the Step function and checking up with AIC level, the better fitted model only includes predictors the knowledge creation sector and firms' name having the Greater Boston area research institutions' names. In the aforementioned linear regression model, the knowledge creation sector is statistically significant at the significance level 0.05, further confirming the observations made in the original model.

Limitations

The limitation of my initial work is that the model is built based on companies in the healthcare industry instead of narrowing down to the biotechnology sector. In this capstone project, the model entails patterns in biotechnology sectors.

More quantitative variables (such as financial measurement from Pitchbook) should included in the model for having a better fitted models. More data collection should be conducted for the better result and understanding of the research problem.

This capstone project is the extension of an ongoing research project, and there should be no limitation on using any material from this research project as all the research materials are created and compiled by myself. Data is harnessed through the private (subscription-based) website, so data shouldn't be published in GitHub for ethical practice in data science.