Probability of Greater Boston's Biotech Startups Reaching Financial Success

Zhirui Xiong '23, Data Science Major Capstone

Background and Research Questions

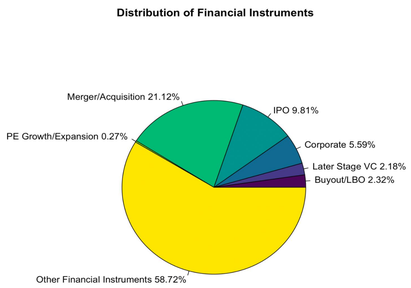
- **Research Background**: this research is inspired by Startup Cartography Project
  - Measure entrepreneurial ecosystem statistics for the United States from 1988 to 2016
  - Build a predictive analytics approach to estimate the probability of "extreme" growth close to the time of the founding
- **Question**: 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?
- **Outcome Measurement**: The success of a biotechnology startup is defined as the company reaching the IPO and capital received, usually by late-stage startups. These financial instruments will positively affect publicly listed biotechnology firms' long-term enterprise value evaluation.

Data

- **Dataset Sources**: Pitchbook and Orbis (Subscription-Based Data Platform)
- **Selection Criteria**: Biotechnology NAICS industry code only include sectors in diagnostic analytics, biotechnology, bio-manufacture equipment, biotech laboratory
- **Data Processing**: Algorithm created to match the same company from Pitchbook and Orbis based on the names, addresses, and industry sector
- **Data Cleaning**: Replicated data have been removed manually
- **Data Summary**: 734 companies in total, 11 Variables for model building

X Variables (Predictors)	Binary Value Given to X Variables
Company incorporated in Delaware	1 if the company is incorporated in Delaware; 0 if otherwise (Delaware has business-friendly laws and low taxes)
Type of entity as corporation	1 if the type of entity is a corporation; 0 if otherwise (Corporation has separate legal identities from its owners to raise capital and limit liability)
Local business	1 if the company is a local business; 0 if otherwise
Nationwide business	1 if the company is a traded business nationwide; 0 if otherwise
Biopharmaceutical sector	1 if the company is in the sector; 0 if otherwise
Education and knowledge creation sector	1 if the company is in the sector; 0 if otherwise
Information and analytical Instruments sector	1 if the company is in the sector; 0 if otherwise
Medical facility sector	1 if the company is in the sector; 0 if otherwise
Short firms' name	1 if the company's name is less than four words; 0 if otherwise
Affiliated with higher education Institutions	1 if yes; 0 if no (Affiliation to institutions provides more access to funding, research resources, and many more)
Y Variable (Response Variable)	**Binary Value Given to X Variables**
Most recent financing instruments used	1 if Merger and Acquisition, IPO, Corporate Merging, Later Stage Venture Capital Funds, Buyout/LBO, Private Equity Growth/Expansion are used as of September 2022; 0 if otherwise

- **Distribution of Outcome Vraible**: financial instruments used recently shows
  - Approximately 60% of the Y variable is 0; the other 40% is 1.
  - Merger and acquisition and IPO, an indicator of successful startups, consists majority of outcome variables noted as 1.



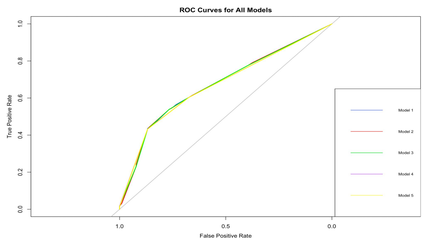
Statistical Model

- **Model Type and Reasons**: Binary logistic regression is the best model for estimating the probability of an event occurring (startup being successful or not) based on multiple predictors. The model identifies predictors that are primarily influential to the outcome. In addition, the model estimates the direction and magnitude of each predictor's effect on the outcome.
- **Model Assumptions**: The model satisfies the following assumption: the linear relationship between predictors and log odds of independent variables, eros being independent of each other, and no multicollinearity between predictors.
- **Summary of 5 Models**:

Model	Formula	Deviance	AIC	Hosmer and Lemeshow GOF test	Area Under ROC Curve (AUC)
1	Incorporated In Delaware + Type of Entity + Local Business + Pharmaceutical + Education and Knowledge Creation + Short Company's Name + Affiliated to Higher Education Institution	924.88	938.88	X-squared = 7.2397e-15, df = 8, p-value = 1	0.6733
2	Incorporated In Delaware + Type of Entity + Pharmaceutical + Education and Knowledge Creation + Short Company's Name + Affiliated to Higher Education Institution	924.90	936.90	X-squared = 7.3807e-15, df = 8, p-value = 1	0.6734
3	Incorporated In Delaware + Type of Entity + Pharmaceutical + Education and Knowledge Creation + in Company's Name + Affiliated to Higher Education Institution	925.05	935.05	X-squared = 7.4697e-15, df = 8, p-value = 1	0.6728
4	Type of Entity + Pharmaceutical + Education and Knowledge Creation + Affiliated to Higher Education Institution	925.61	933.61	X-squared = 7.3987e-15, df = 8, p-value = 1	0.6704
5	Pharmaceutical + Education and Knowledge Creation + Affiliated to Higher Education Institution	925.61	933.61	X-squared = 5.3681e-15, df = 8, p-value = 1	0.6702

Goodness of Fit

- **AIC and Deviance**: measure the model's goodness of fit with different predictors. The lower AIC and Deviance indicate the model is better fitted. Furthermore, the lower AIC indicates a better trade-off between the goodness of fit and the complexity of the model. All five models have similar Deviance, while their AIC decreases as fewer predictors are involved in the regression.
- **Hosmer and Lemeshow Test**: measures how well the observed value of the dependent variable in logistic regression matches the predicted value by the model. The test shows the level of accuracy of logistic regression predicting the probability of the outcome. All five models have the perfect result, as their p-value are all 1, demonstrating the excellent fit to the data.
- **Area Under Receiver Operating Characteristic Curve (AUC)**: measures the performance of the binary classifier and how well the model can differentiate two possible outcomes of the dependent variable. The higher AUC indicates that the model can better distinguish two possible outcomes (positive and negative cases), while 0.5 AUC suggests the model is a random classifier. In the graphs below, the AUC value is similar across five models, approximately at 0.67.



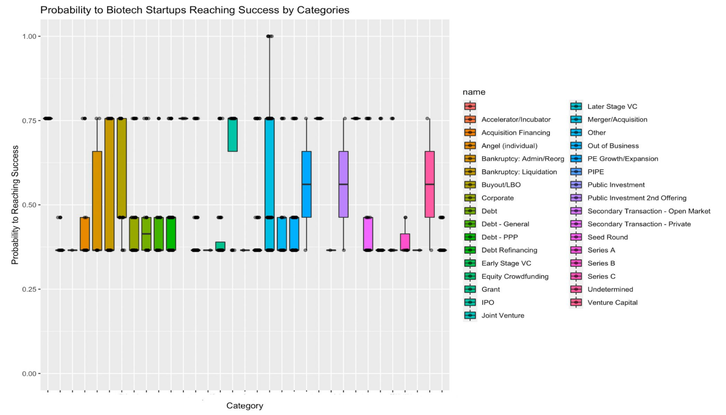
- **Model 2, Best Model Based on Goodness of Fit**: Model 2 has the lowest Deviance and AIC values, as well as the highest values in area under ROC Curve (AUC).

Results

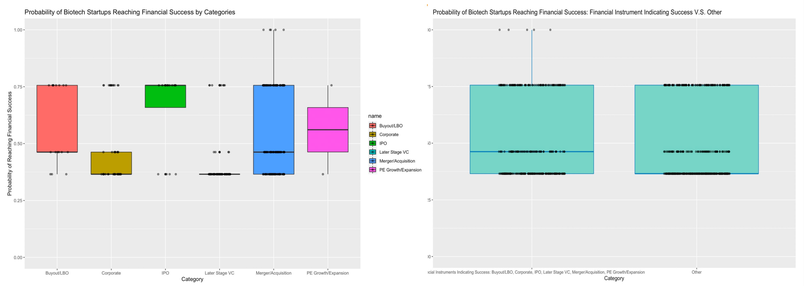
- **Regression Model**: log(odds of reaching financial success) = 14.28664 + 0.06177(Incorporated In Delaware) - 14.47653(Corporation) + 1.27943(Biopharmaceutical) - 0.39061(Education and Knowledge Creation) + 0.04522(Short Company's Name) + 15.25418(Affiliated to Higher Education Institution)
- **Interpret Coefficients**: Holding all other variables constant for each case, the probability of success is negatively impacted when the firm becomes a corporation or is in the education and knowledge creation sector. The incorporation in Delaware, pharmaceutical sectors, having a short name, or being affiliated with higher education institutions positively influence the probability of success, with affiliation with institutions having the maximum numerical effect on success.

Probability of Success by Categories

- **By All Financial Instruments Categories:** Companies receiving the later stage venture capital funds and IPO have the most potential to reach success. In contrast, companies with public investment and mergers and acquisitions have a moderate probability of becoming successful. Interestingly, bankruptcy aimed at reorganization has a wide range of possibilities, indicating that bankruptcy is not the end of the world for the startups and firms in the biotechnology sector of the Greater Boston Area.



- **Financial Instruments Indicating Startups' Success V.S. Other Instruments:** When using financial Instruments Indicating Startups' success, the startups have more possibility to reach success than other instruments. Within all financial instruments indicating startups' success, startups with IPO tend to have a higher probability of becoming successful, while corporate acquisition tends to fail the development of the startup. Companies undergoing mergers and acquisitions have various possibilities ranging from failing to succeeding the business.



Limitation

- More quantitative variables (such as financial measurements from Pitchbook) should be included in the model to have a better-fitted model. More data collection should be conducted for a better result and understanding of the research problem.
- Data is harnessed through the subscription-based website, so the original data shouldn't be published in GitHub for ethical practice in data science.