# Probability of Greater Boston's Biotech Startups Reaching Financial Success

Zhirui Xiong ‘23, Data Science Major Capstone

## Background and Research Questions

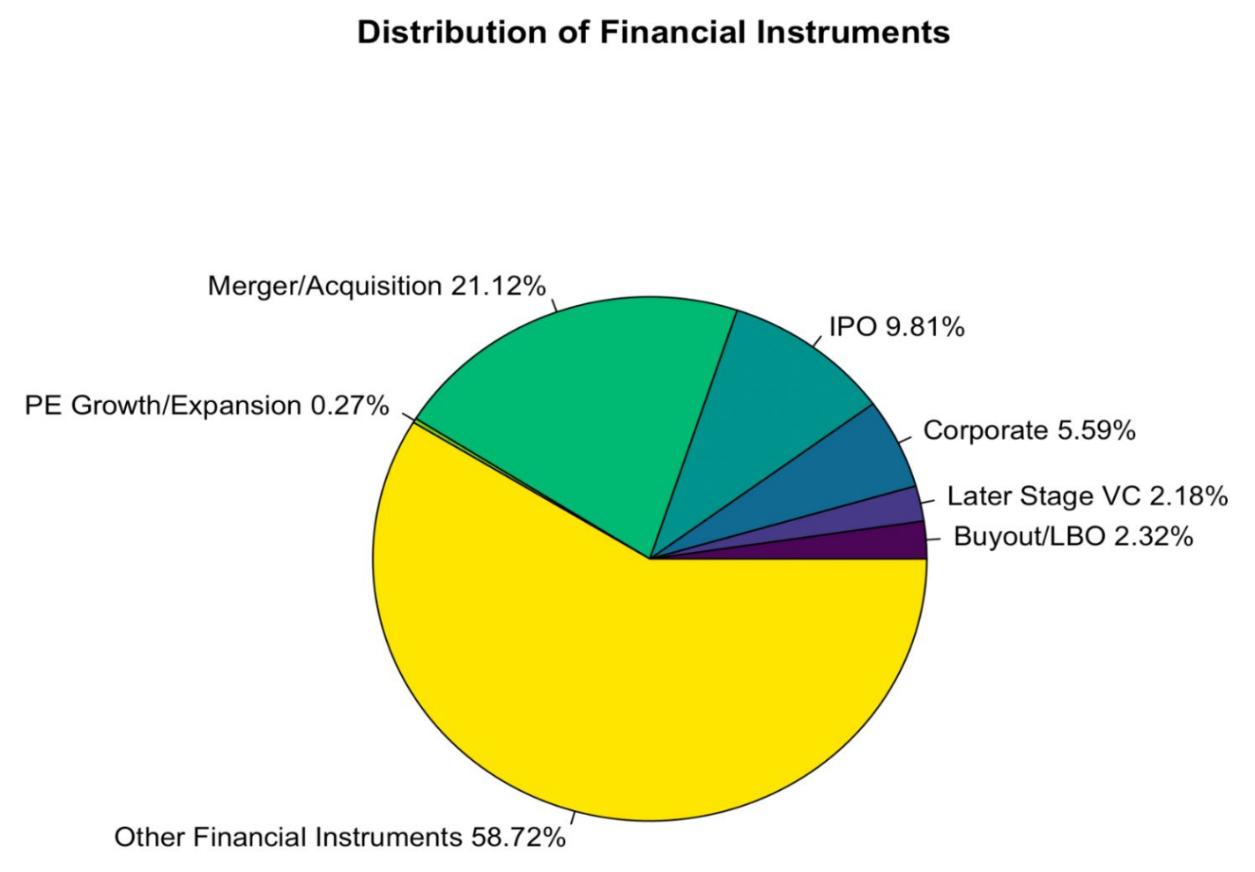
- **Research Background**: this research is inspired by Startup Cartography Project
  - Measure entrepreneurial ecosystem statistics for the United States from 1988 to 2016
  - Build a predictive analytics approach to estimate the probability of “extreme” growth close to the time of founding
- **Question Addressed**: 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 have the positive implications on the long-term enterprise value evaluation of the publicly listed biotechnology firms.

## Data

- **Dataset Sources**: Pitchbook and Orbis (Subscription-Based Data Platform)
- **Selection Criteria**: Biotechnology NAICS industry code, only include sectors in diagnostic analytics, biotechnology, biomnaufature equipment, biotech laboratory
- **Data Processing**: Algorithm created to match the same company from Pitchbook and Orbis based on the names, address, 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<br>(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<br>(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<br>(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 |

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



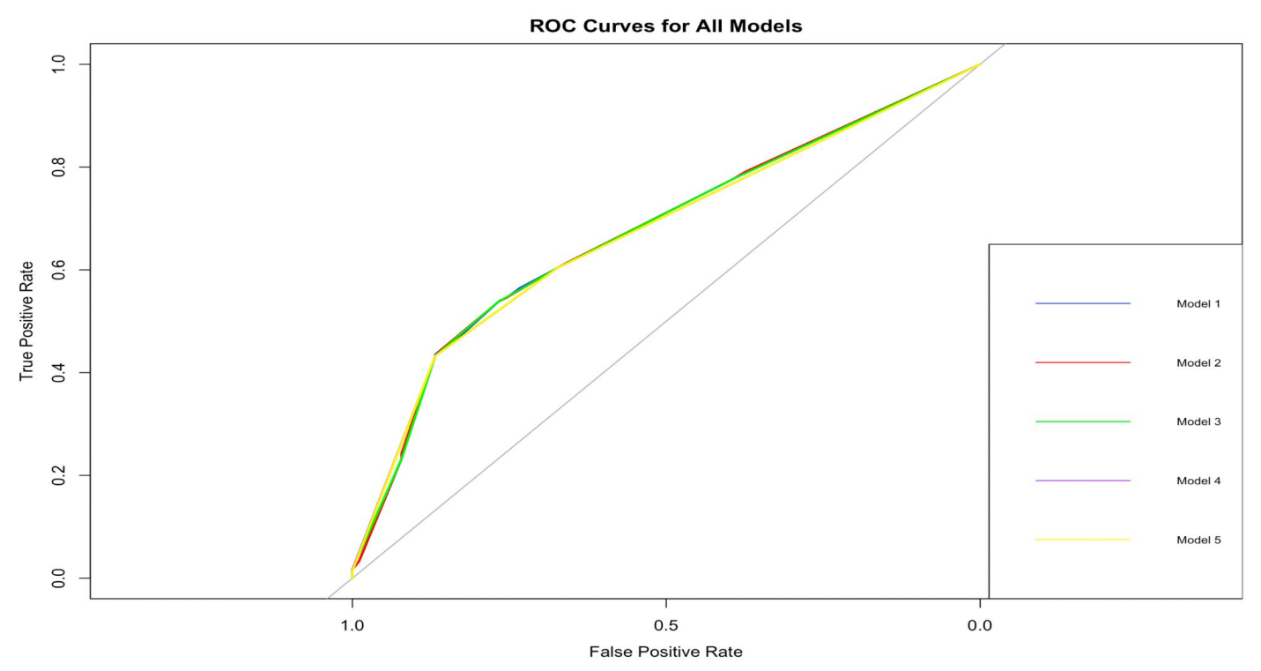
## Data Modeling

- **Model Types**: Binary Logistic Regression
- **Reasons**: Binary logistic regression is the best model used for estimating probability of an event occurring (startup being successful or not) based on multiple predictors. The model identifies predictors that are mostly influential to the outcome. In addition, the model estimate direction and magnitude of the effect from each predictors on the outcome.
- **Model Assumptions**:The model satisfies the following assumption: linear relationship between predictors and log odds of independent variables, eros being independent from 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 goodness of fit of the model with different predictors. The lower AIC and deviance indicates the model is better fitted. Furthermore, the lower AIC indicates a better trade-off between goodness of fit and complexity of the model. All 5 models have similar Deviance, while their AIC decreases as less predictors are involved in the regression.
- **Hosmer and Lemeshow Test**: measures how well the observed value of 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 has the perfect result, as their p-value are all 1, demonstrating the good fit to the data.
- **Area Under Receiver Operating Characteristic Curve (AUC)**: measure the performance of binary classifier and how well the model can differentiate two possible outcomes of dependent variable. The higher AUC indicates that the model can better distinguish two possible outcomes (postive and negative cases), while 0.5 AUC indicates the model to be a random classifier. In the graphs below, AUC value are all similar across 5 models, approximately at 0.67.



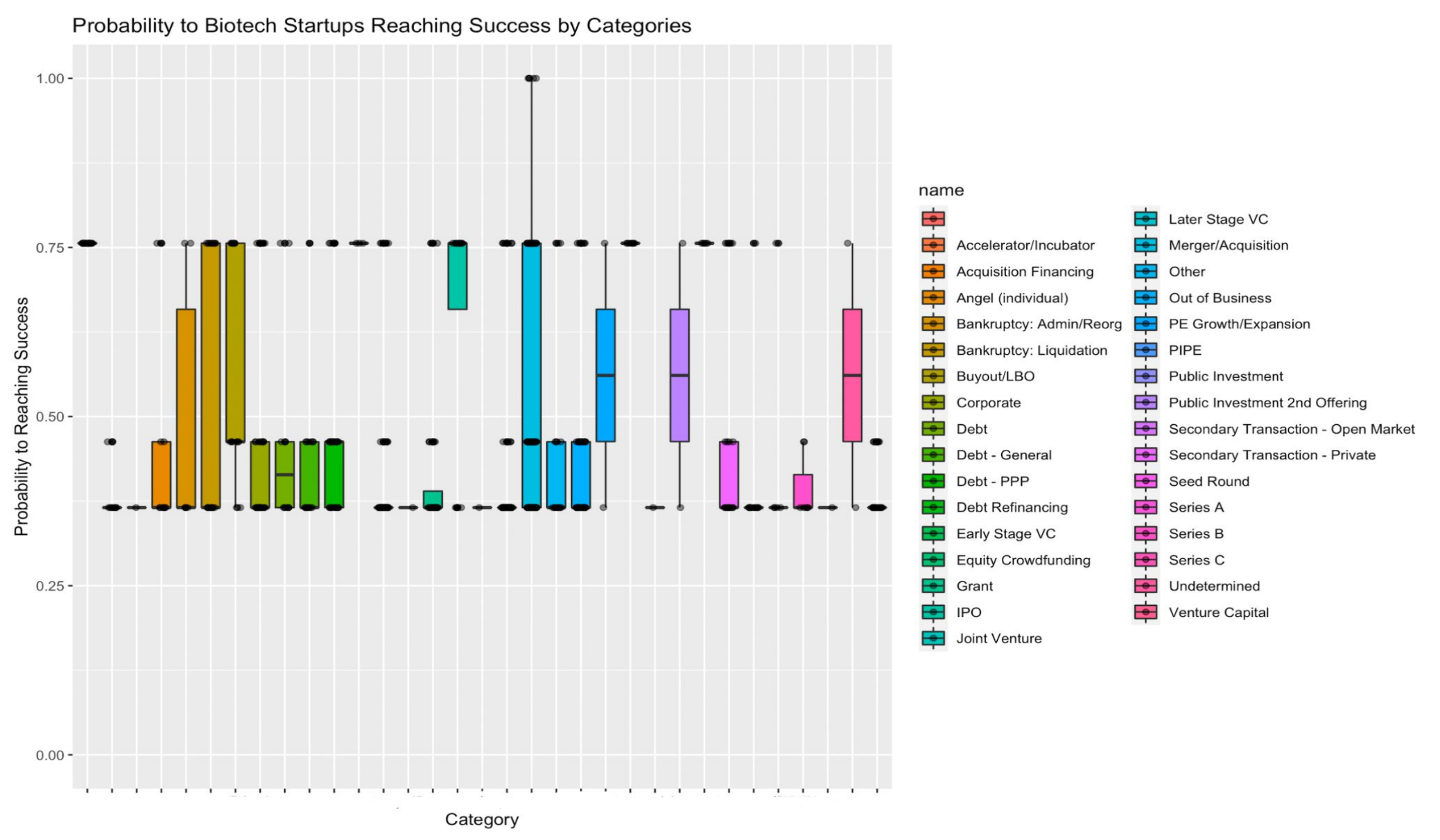
- **Model 2, Best Model Based on Goodness of Fit**: Model 2 has the lowest Deviance and AIC values, as well as the highest values in area under ROC Curve (AUC).

## Results

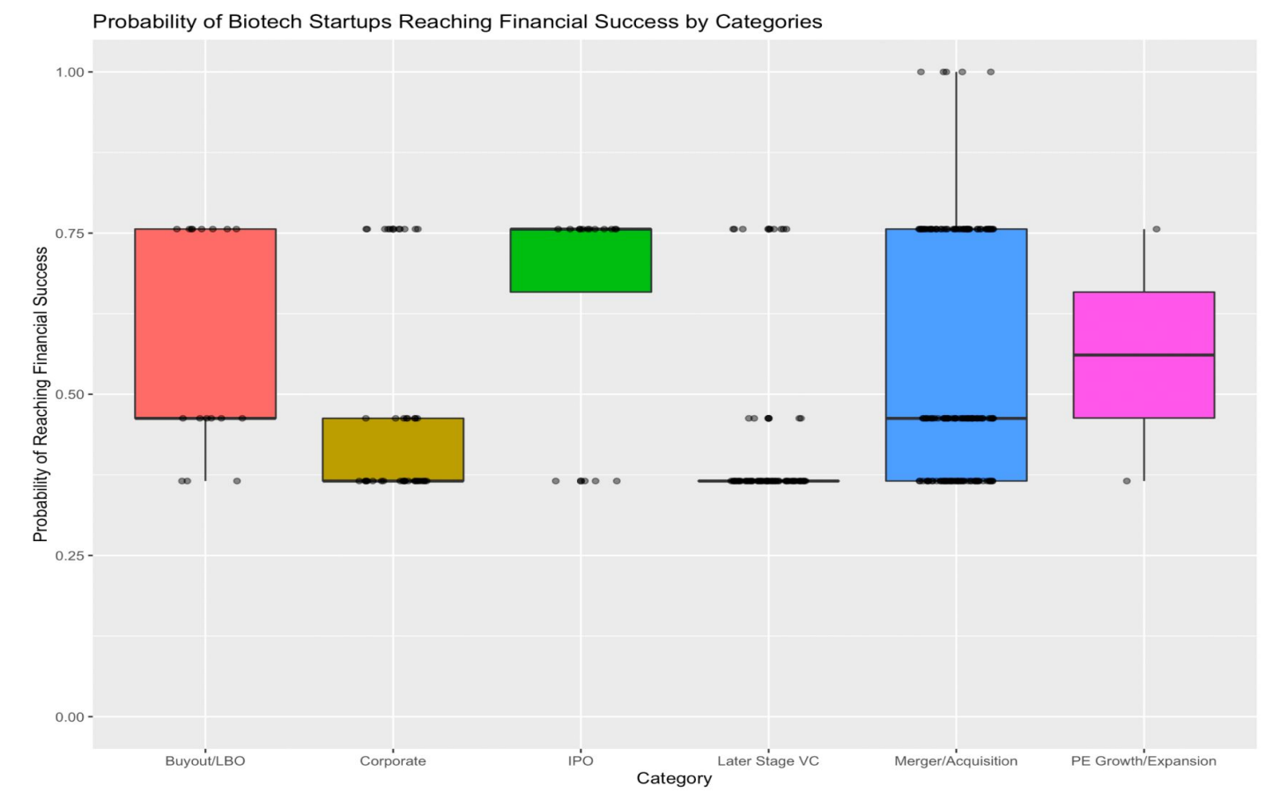
- **Regression Model**: log(odds of reaching financial success) = 14.28664 + 0.06177(Incorporated In Delaware) - 14.47653(Corporation) + 1.27943(Biopharmaceutical) - 0.39061(Education and Knowledge Creation) + 0.04522(Short Company's Name) + 15.25418(Affiliated to Higher Education Institution)
- **Interpret 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 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 utmost numerical effect on success.

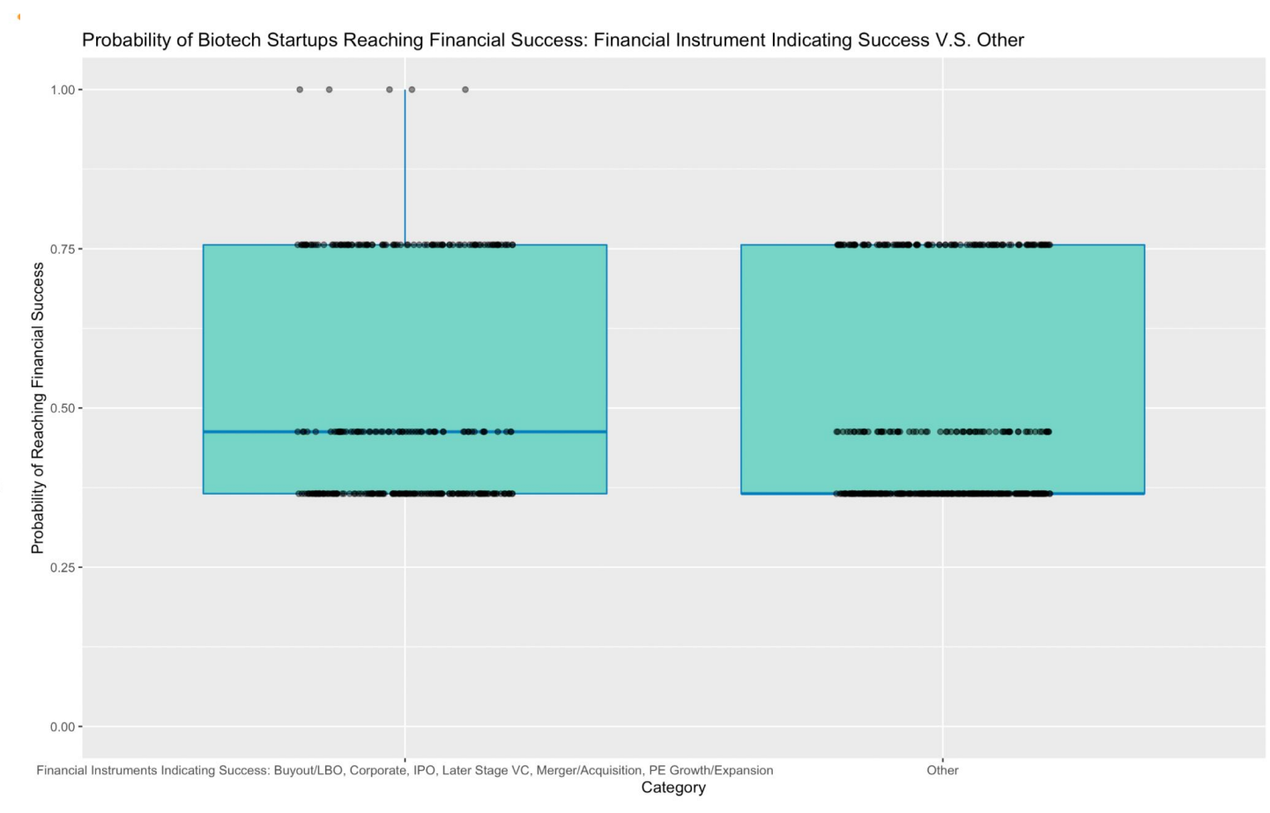
## Probability of Success by Categories

- **By All Financial Instruments Categories:** Companies receiving the later stage venture capital funds and IPO has the most potential to reach success, while companies with public investment and merger and acquisition have moderate probability. Interestingly, bankruptcy aimed at reorganization has wide range of probability, an indication that bankruptcy is not the end of the world for the startups and firms in biotechnology sector of 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, compared to other instruments. Within all financial instruments indicating startups’ success, startups with IPO tends to have higher probability to become successful, while corporate acquisition tends to fail the development of the startup. Companies underwent mergers and acquisitions has wide variety of possibility ranging from failing or succeeding the business.





## Limitation

- 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.
- Data is harnessed through the subscription-based website, so the original data shouldn’t be published in GitHub for ethical practice in data science.