Investigating Startup Success Factors

Ben Piekarz, Jose Urruticoechea

Team Name: Innovation Oracles

Introduction

The goal of every investor and entrepreneur is to maximize the success of their businesses. In this project, we aimed to investigate how various factors influence the exit success of startups. To do this, we tested several hypotheses and constructed two different ML models intended to predict startup "success".

Hypotheses

1. Software startups typically exhibit higher multiples on invested capital (MOIC), a measure of ROI, than other industries.
2. California is disproportionately represented in the software industry compared to the rest of the world.
3. Startups in California are more likely to exit with higher MOICs.
4. Cybersecurity startups are more likely to achieve successful exits than other industries.

Models

1. Predicting startup exit based on factors at time of seed raise with logistic regression
2. Predicting total amount of fundraising based on factors at time of seed raise

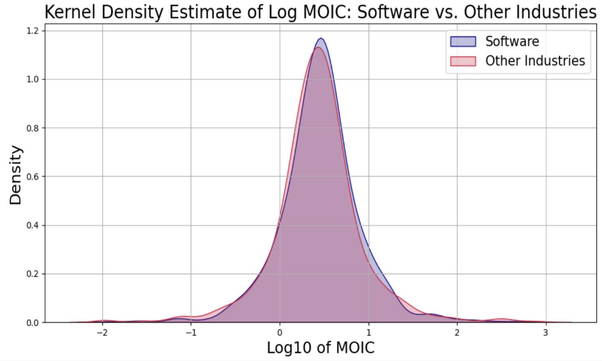
Methods

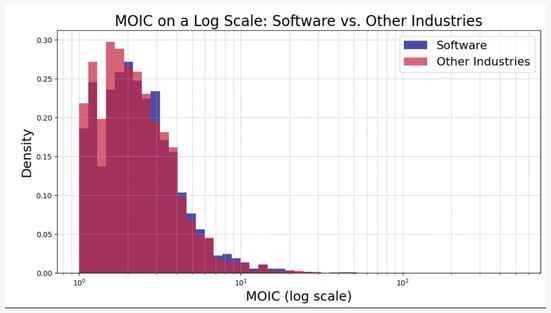
- Scraped and parsed Pitchbook dataset with startup characteristics such as industry, location, and exit valuations
- Note: data contains over 4392 records, each representing a startup that raised Seed stage funding in the year 2015.
- Calculated a multiple on invested capital (MOIC) figure for every entry.
- Used Mann-Whitney U and Chi-Squared tests to evaluate hypotheses
- Used logistic regression to create predictive ML models

Results

Hypothesis 1

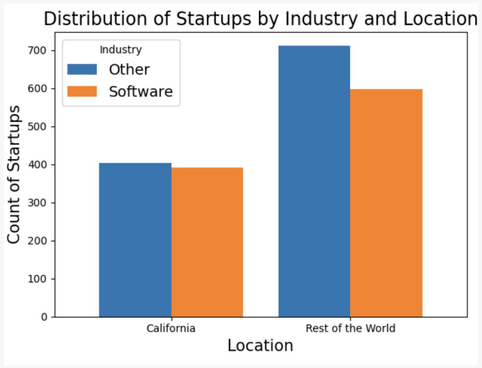
P-value: 0.051 -> While we're on the edge of statistical significance, we fail to reject the null hypothesis, suggesting there is not enough evidence to conclude that software startups exit with higher MOICs than startups in other industries.





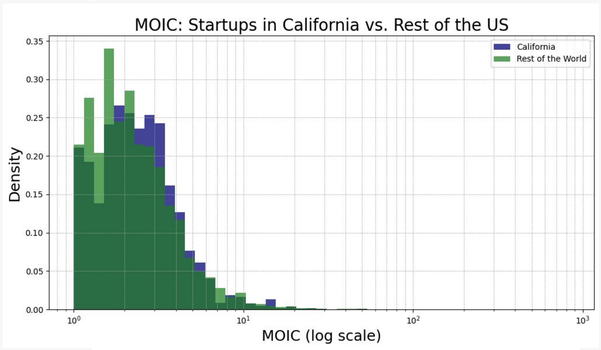
Hypothesis 2

P-value: 0.12-> There is not enough evidence to suggest a significant association between a startup's HQ location and its primary industry group being Software.



Hypothesis 3

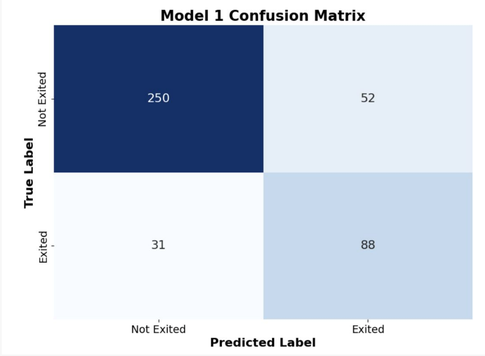
P-value: 0.048 -> We reject the null hypothesis, suggesting that startups in California are likely to exit with higher MOICs than those based elsewhere.



ML Model 1: Predicting Startup Exit

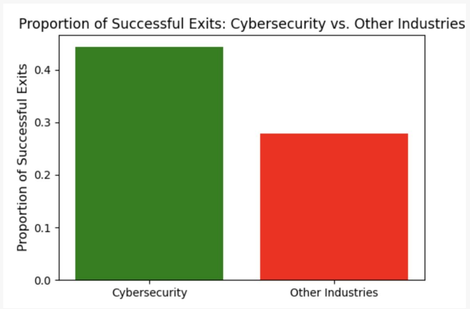
Accuracy: 0.80

Based on our results, it appears that whether or not a private seed-stage startup exits can be predicted fairly well using variables such as location, industry, and information about its first funding round.



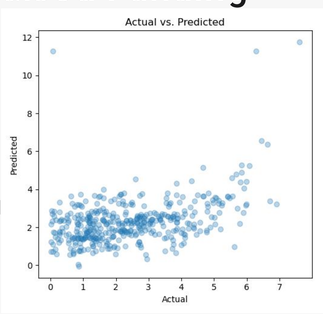
Hypothesis 4

P-value: 0.004 -> We find evidence to suggest that cybersecurity startups are more likely to exit compared to startups in other industries.



ML Model 2: Predicting Total Raised Funding

Our linear regression model predicts companies' funding (in millions) post-logarithm with a MSE of 2.33, indicating it struggles with complex financial predictions.



Discussion

- We find that several circumstantial factors do correlate with startup success: For example, HQ location and industry seem to materially impact a startup's return on investment
- Linear regression on our data was not powerful enough to effectively predict something as complex and variable as money raised.
- **Limitations**: data sourced only from startups that raised seed funding in 2015 which limits relevance of extrapolating
- Moving forward, we'll source more data from different years and stages of a startup's fundraising journey to make more advanced predictions

Acknowledgements

Thank you to Professor De Stefani and for our mentor TA, Alex Ding, for guiding us through the course and project.

Contact: ben_piekarz@brown.edu (+1 408 505 3534) | jose_urruticoechea@brown.edu (+1 401 241 6917)

Startup Success Factors

bpiekarz, jurruti

Hypothesis

We aimed to understand the dynamics affecting startup success by testing four specific hypotheses related to industry characteristics and geographical location. These included the performance and exit outcomes of software and cybersecurity startups and the prevalence and success metrics of startups based in California.

Data

We used a comprehensive Pitchbook dataset (which we have access to through the university) detailing startup characteristics such as industry, location, and exit valuations, from which we also calculated a multiple on invested capital (MOIC) figure for every entry. The dataset contains over 4392 records, each representing a startup that raised Seed stage funding in the year 2015. Pitchbook sources its data by employing over 650,000 web crawlers that scan the web for information on public and private companies. Attributes of the data include descriptors such as the primary industry group, the total amount raised, exit and ownership statuses, and more regarding the startups. The dataset's complexity required rigorous data cleaning to handle missing values and ensure accurate categorization for analysis.

Findings

Note that the claims below represent our alternative hypotheses. In order to provide evidence-based backing for them, we must reject the inverse of them (the null hypotheses).

Claim #1: Software startups typically exhibit higher MOICs than other industries.
Support for Claim #1: Using the Mann-Whitney U test, we analyzed MOIC distributions, finding a p-value of 0.05085, so we failed to reject the null hypothesis at our desired 95% confidence interval. These results do suggest a trend towards higher MOICs for software startups but are not statistically significant at the conventional 0.05 level.

Claim #2: California is disproportionately represented in the software industry compared to the rest of the world.
Support for Claim #2: We ran a Chi-square test, which showed no significant association between being located in California and belonging to the software industry. Hence, we failed to reject the null hypothesis.

Claim #3: Startups in California are more likely to exit with higher MOICs.
Support for Claim #3: Using the Mann-Whitney U test, we found that California-based startups tend to have higher MOICs, with a statistically significant p-value of 0.04745, rejecting the null hypothesis and suggesting that geographic location might influence financial success.

Claim #4: Cybersecurity startups are more likely to achieve successful exits than other industries.
Support for Claim #4: Our analysis revealed that cybersecurity startups have a higher likelihood of successful exit (p-value = 0.00388), rejecting the null hypothesis and supporting the hypothesis that industry type significantly impacts exit outcomes.

Model #1: We used a logistic regression model to predict startup exits—defined as being acquired, going public, or entering IPO registration—using features such as the HQ location, industry, and information about the initial fundraising. It achieves an accuracy of 80.3% and processes these features through a pipeline that includes imputation, scaling for numerical data, and one-hot encoding for categorical data. During testing, the model predicted 250 true negatives (correctly predicted non-exits) and 88 true positives (correctly predicted exits), with 52 false positives and 31 false negatives, indicating it is more adept at identifying non-exits than exits.

Model #2: We used a linear regression model to predict the total amount of funding a startup raised based on its seed-stage funding and other factors determined at the time of seed-stage funding. The model predicts companies' funding (in millions) post-logarithm, but testing results in a Mean Squared Error of 2.33, indicating that this model struggles with prediction.