A Machine Learning Approach to Detect Early Signs of Startup Success

Abhinav Nadh Thirupathi, Tuka Alhanai*, Mohammad M. Ghassemi*

thirupa3@msu.edu tuka.alhanai@nyu.edu ghassem3@msu.edu

*Both authors contributed equally

The Authors

Team of students, professors, and entrepreneurs

Abhinav Nadh Thirupathi

Is a senior at MSU pursuing degrees in computer science and economics. He is also a data science intern and a teaching and research assistant. His research interests lie at the intersection of AI and finance.

Tuka Alhanai, Ph.D.

Is an Assistant Professor of Computer Science at New York University. She received a Ph.D. from MIT where she was recognized as one of the world's top innovators under 35. Her research was highlighted by Bill Gates as a frontier area.

Mohammad M. Ghassemi, Ph.D.

Is an Assistant Professor of Computer Science and Engineering at MSU and National Service Scholar at the National Institutes of Health. Before joining MSU, he was a director of data science at S&P Global, and a strategic consultant with BCG.

Motivation

Differentiating between successful and unsuccessful startups is a hard problem

Predicting startup success is a difficult challenge

SpaceX, founded in 2002, crossed $100 billion[1] in valuation while many other rocket startups such Rotary Rocket are now defunct. This leads to the question: how do we detect startup success?

Many factors further complicate a difficult problem

Most investors, including the top venture capitalists, have a 25%[2] success rate. More importantly, many factors such as work ethic, economic conditions, geography, individuals' background, etc. exacerbate an already challenging problem.

[1] https://www.cnbc.com/2021/10/08/elon-musks-spacex-valuation-100-billion.html
[2] https://www.wsj.com/articles/SB10000872396390443720204578004980476429190

Research Objectives and Contributions

Can we predict the success of startups using initial conditions?

1. Collect and curate a novel large scale dataset from SBIR/STTR and Crunchbase
2. Predict future success (IPO and/or M&A) of a startup using initial conditions
3. Identify time-independent factors associated with future startup success
4. Release code and data to facilitate reproducibility and extensions

Data

Used 3,160 companies and people data to predict startup success

Startups collected from SBIR/STTR recipients

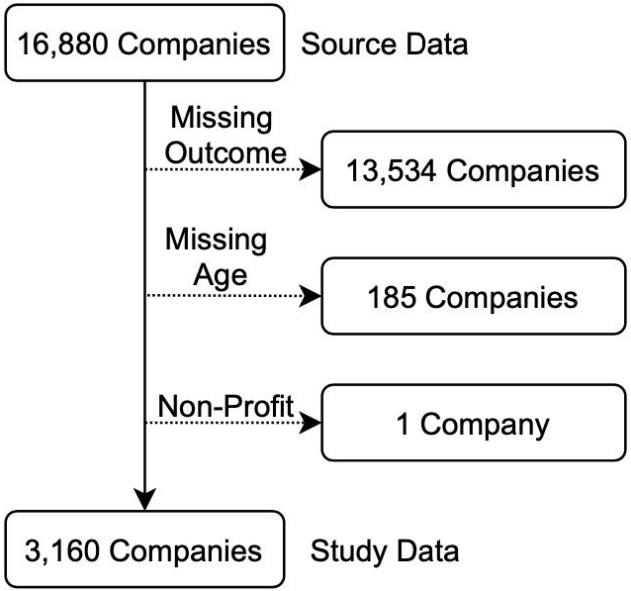
Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) awards are granted through individual government agencies to startups (less than 500 employees) to pursue innovations for consumer potential. We collected 16,880 companies who were recipients of the SBIR/STTR awards.

Utilized data from 3160 Companies and their People

After following our exclusion criteria (described in a later slide), we had a novel dataset on 3160 companies and their people. The dataset contains characteristics of the companies *and* individuals from the SBIR/STTR databases and Crunchbase.

Outcome: Initial Public Offering (IPO) and Acquired (M&A)

A startup was considered successful if it underwent an initial public offering (IPO), and/or merged with, and/or was acquired by another entity (M&A). Out of the 3160 companies, 1657 companies were successful and 1503 companies were considered unsuccessful.



Features

Utilized 16 company-level features describing the headquarters region and sector

Company Features (16 features)

Headquarters Regions	(1) New England (2) Midwestern US ...
Sectors	(6) Finance (7) Energy (8) Industrial ...

People Features (32 features)

Number of Founded Orgs.	(1) Average number founded
Gender	(2) Male (3) Female ...
Regions	(8) East Coast (9) Midwestern US ...
University Attended	(18) Prestigious ...
Academic Degree	(20) Bachelors (21) Masters ...
Academic Major	(25) Engineering (26) Art …

Example Company Features

3 Sigma

Headquarters Region: Midwestern US
Sector: Industrials

Features

32 people-level features aggregated at the company-level

Company Features (16 features)

Headquarters Regions	(1) New England (2) Midwestern US ...
Sectors	(6) Finance (7) Energy (8) Industrial ...

People Features (32 features)

Number of Founded Orgs.	(1) Average number founded
Gender	(2) Male (3) Female ...
Regions	(8) East Coast (9) Midwestern US ...
University Attended	(18) Prestigious ...
Academic Degree	(20) Bachelors (21) Masters ...
Academic Major	(25) Engineering (26) Art …

People Features Aggregation

A company can have many people associated with it, so we used different statistical summarization and aggregation methods. Specifically, peoples' text features and continuous features were averaged by company, and nominal features were summed at the company-level.

Method*

Utilized an extensive data preparation pipeline

Extracted Data Preprocessing

Columns missing >90% of the data were removed, nominal columns were one-hot coded, free text descriptions were vectorized using Doc2Vec, and continuous columns remained as numerical features.

Study Data Curation

Study data was processed to exclude: (1) features correlated with the age of the company, (2) companies with an unknown age, (3) companies missing their outcome status, and (4) companies younger than 2 years because it took at least 2 years for a companies to IPO or M&A based on our data statistics.

* See the Paper and Supplementary Materials for more information

Method*

Extra measure to remove correlation between company's age and features

Data Normalization

The probability of a company achieving an IPO or M&A is likely to increase as a function of the age of the company. We removed the correlation between the age of the company and the features through a z-score (i.e. zero-mean unit-variance) normalization of all continuous features for all companies that shared a founding year.

* See the Paper and Supplementary Materials for more information

Results

XGBoost model exhibited excellent performance

Applied the XGBoost algorithm

We utilized XGBoost due to its ability to handle missing data.

Assessed performance using LOOCV

Leave-one-out cross validation was used to evaluate the performance of the model.

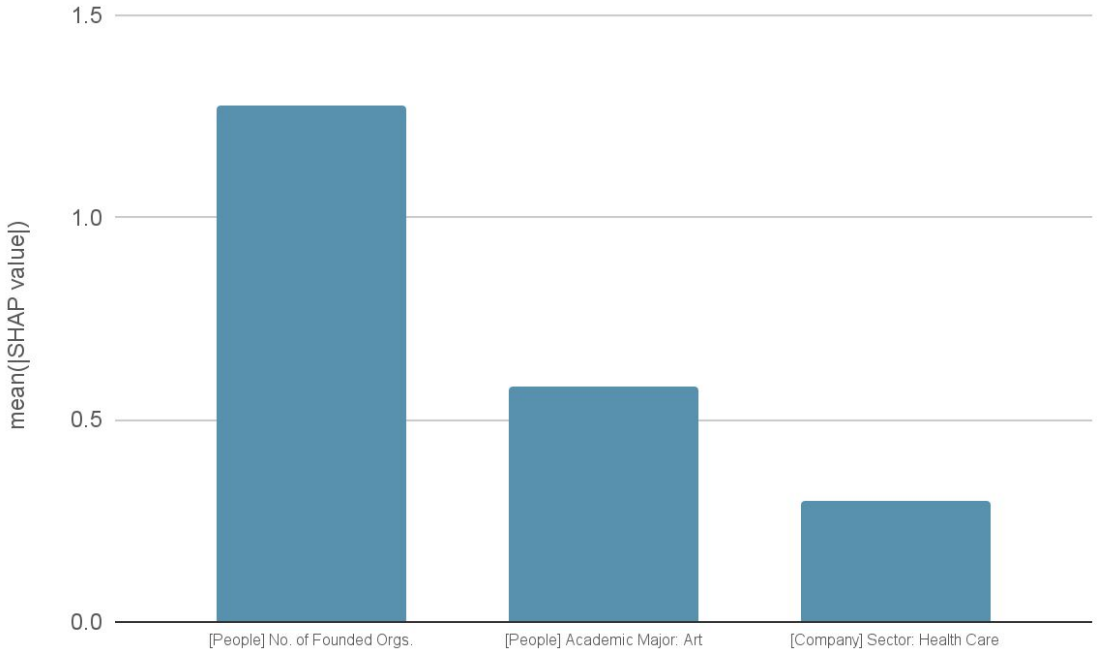
SBIR/STTR Program has a high barrier to entry

The superior performance of the model may be partially attributed to the high quality of companies accepted into the SBIR/STTR program.

Metric	Score
AUC	0.91
Accuracy	0.84
Precision	0.84
Recall	0.85
F-score	0.84
Brier Score	0.16

Results
SHAP magnitudes indicate *absolute* feature importance

Feature Importance
We identified the important features using model-agnostic SHAP technique. Number of founded organizations and academic background in arts played a major predictive in the outcome.

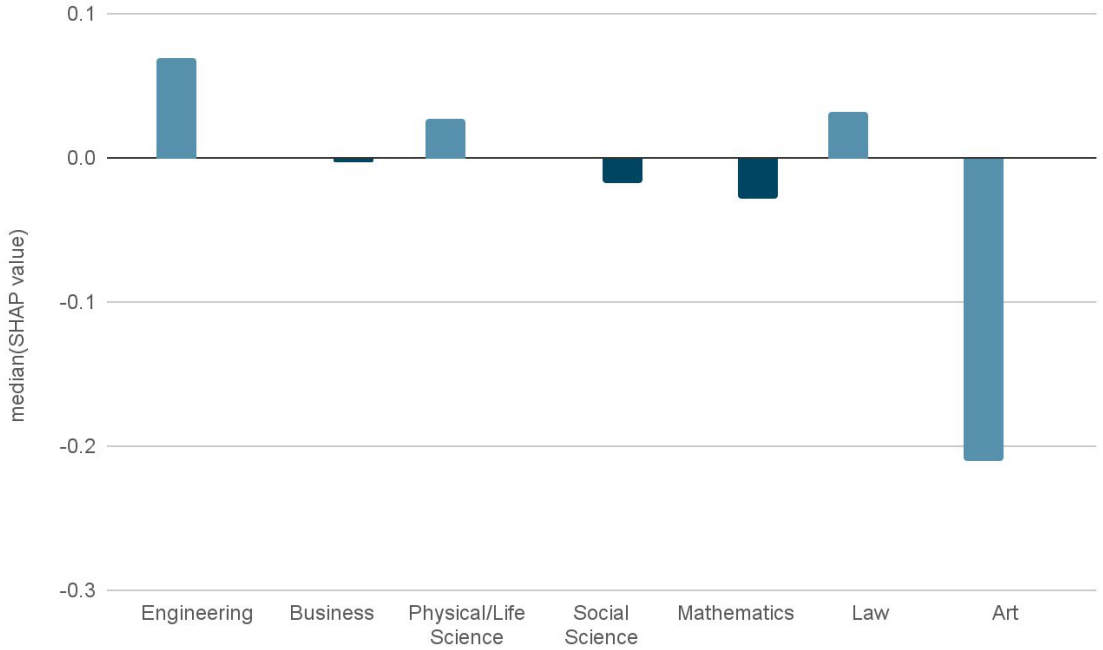


Results

Proportion of STEM majors is positively associated with startup success

Academic majors impact on outcome

While STEM majors had a positive impact on predicting startup success, an academic major in art had a significant negative impact on the outcome.

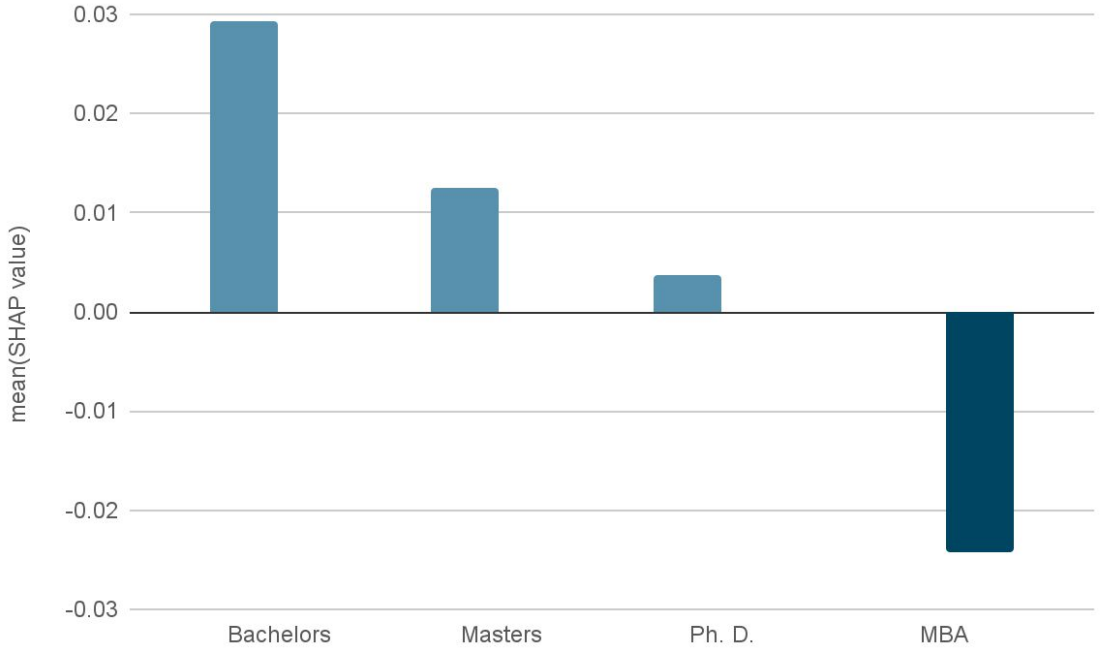


Results

High proportion of problem solvers to managers associated with positive success

Level of education impact on the outcome

People with an undergraduate degrees had a positive impact on predicting startup success.



Results

Model possibly could be applied to provide investment recommendations

Prediction Thresholds

The prediction thresholds may be used to inform potential investment decisions. More specifically, if an investor's threshold for a false positive investment recommendation was 10%, the investor could estimate the true positivity of the recommendation at about 72.5%.

FPR at TPR			TPR at FPR		
90%	**95%**	**100%**	**10%**	**5%**	**0%**
26.3%	46.2%	98.5%	72.5%	58.1%	3.8%

Conclusions

Machine learning model can predict startup success using initial conditions

1. **Collect and curate a novel large scale dataset from SBIR/STTR and Crunchbase**

   A large scale dataset comprising both company-level and people-level features was collected from the U.S. federal government's SBIR/STTR programs and Crunchbase. A total of 3160 companies and their people characteristics were used in this study.

2. **Predict future success (IPO and/or M&A) of a startup using initial conditions**

   We extracted 16 company-level features and 32 people-level features for 3160 companies. Using the data, we trained and evaluated a XGBoost classification model to detect startup success.

Conclusions

Machine learning model can predict startup success using initial conditions

3. **Identify time-independent factors associated with future success**

   We found that employees with entrepreneurial experience, arts, and/or STEM educational backgrounds, among other characteristics played a significant role in predicting the outcome of small businesses.

4. **Release code and data to facilitate reproducibility and extensions**

   All the data and code has been made available online so anyone can reproduce our model, figures, and performance metrics. This will make it easier for others to expand and build on our work.

   https://github.com/abhit20/ML-Startup-Success

Any Questions?

Paper Code Contact