A Machine Learning Approach to Detect Early Signs of Startup Success

Abhinav Nadh Thirupathi
Michigan State University
East Lansing, Michigan, USA
thirupa3@msu.edu

Tuka Alhanai*
New York University
Abu Dhabi, UAE
tuka.alhanai@nyu.edu

Mohammad M. Ghassemi*
Michigan State University
East Lansing, Michigan, USA
ghassem3@msu.edu

ABSTRACT

In this study, we investigate a heterogeneous set of startup ventures (different ages, products, teams, levels of maturity, etc.) to identify the time-independent factors associated with their future success. More specifically, we investigated 3,160 unique companies, all of which were recipients of Small Business Innovation Research (SBIR) or Small Business Technology Transfer (STTR) awards. For each company, we collected any publicly available information: the SBIR/STTR award (amount, agency, principal investigator, etc.), and Crunchbase business profile. The collected data were used to train a XGBoost model that predicts whether a company had an initial public offering (IPO), and/or merged with, and/or was acquired by another entity (M&A). The performance of the model assessed using leave one-out-cross validation (LOOCV) was strong: 84% accuracy and 0.91 AUC. We found that employees with entrepreneurial experience, arts, and/or STEM educational backgrounds, among other characteristics played a significant role in predicting the success of small businesses. Our results indicate that machine learning models may be used to assess the viability of small ventures.

CCS CONCEPTS

• **Computing methodologies** → **Ensemble methods**; • **Applied computing**;

KEYWORDS

Ensemble methods, Business Success, Data Mining, Venture Capital, Startups, Factors Extraction

ACM Reference Format:
Abhinav Nadh Thirupathi, Tuka Alhanai, and Mohammad M. Ghassemi. 2021. A Machine Learning Approach to Detect Early Signs of Startup Success. In *2nd ACM International Conference on AI in Finance (ICAIF'21), November 3–5, 2021, Virtual Event, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3490354.3494374

1 INTRODUCTION

There are a myriad of complex factors that influence the success or failure of business ventures: innovation, technology, market conditions, and team to name a few. Young ventures (e.g. startups) may be more sensitive to these factors because they are less likely to have reliable revenue streams or capital buffers needed to optimize product strategies, employee configurations and sales approaches in volatile and competitive markets. As ventures mature, the importance of some factors may change while others remain constant. In this paper, we study the evolution of these factors' importance on success as ventures mature; that is, we identify those factors that consistently influence a venture's ability to succeed, regardless of its level of maturity. In determining the answer to this question, we also provide a quantitative tool to augment the qualitative assessments of investment entities.

To survive in the long-term, for-profit ventures must produce goods and services that maximize revenues while minimizing costs. For many young ventures however, long-term profitability is not feasible without up-front investment. Small ventures often require initial capital investments to support research & development (R&D) efforts, or to scale the size of their operations following successful product development efforts. This required capital may be provided by either private entities (e.g. venture capitalists) and/or public entities (e.g. governments) [11, 20]. In either case, funding entities are typically seeking ventures with the greatest likelihood of returns on investment (ROI). To accomplish their goals, funding entities must discriminate between those ventures that will provide an acceptable ROI from those that will not, *given an investment*. While the precise threshold for what constitutes an acceptable ROI will vary from one investor to another, investments that lead to an acquisition or Initial Public Offering (IPO) of a company are generally understood as successful [4].

Discriminating between successful and unsuccessful ventures is challenging due to the many factors influencing business success, the interactions between those factors, and the potential evolution in the importance of those factors as ventures mature. Furthermore, many factors are often immeasurable or difficult to precisely quantify. For these reasons, it is common for funding entities to use a qualitative review by experts in the funding process [25]. We hypothesize that a correctly designed quantitative methodology may be able to overcome the sparsity challenges inherent to venture data and augment the capability of these investment bodies.

2 BACKGROUND

In this section, we report the survey of the academic literature. Specifically, we describe studies examining team-independent and team-dependent factors of venture success. Furthermore, we categorized the studies into two types: (1) investigative, and (2) practical.

*Both authors contributed equally to this research.

ICAIF'21, November 3–5, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-9148-1/21/11…$15.00
https://doi.org/10.1145/3490354.3494374

2.1 Team Factors of Success

Previous studies have identified *team-independent* properties of small companies that improve their likelihood of success; these factors include: (1) innovation [19], (2) economics [14], (3) environment surroundings [27], (4) geography [1] and even the year of founding [29]. Other studies have identified *team-dependent* properties of small companies that improve their likelihood of success; these factors include: team skills and prior experiences [2], team objectives [26] and demographic diversity [22]. Previous research has often studied team-independent and team-dependent factors independently, limiting the ability of investigators to identify *interactions between factors* that play as critical a role in success as the individual factors alone.

2.2 Investigative Studies

Investigative studies focus on identifying (and discerning the effects of) factors that are statistically relevant to venture success but may not necessarily be practical to collect; for instance, the interpersonal dynamics of a founding team have been associated with venture outcomes [8], but it is practically infeasible to reliably characterize team dynamics outside a formal laboratory setting. A survey of 100 CEOs of small ventures found that positive entrepreneurial motivation [14] was associated with venture performance [10]. Another study found that education and training lead to the success of small businesses [24]. Finally, a study developed a theoretical framework to model success of small and medium-sized enterprises [23]; this approach cannot be scaled and quantitatively applied to companies due to its academic nature.

A few studies have performed investigative analysis on SBIR/STTR data. One study conducted econometric analysis on estimating the probability of commercialization of National Institutes of Health (NIH) SBIR awardees; it found that conditioning Phase II award on external funding/awards increased the probability of commercialization [17]. Likewise, Department of Defense (DoD) SBIR/STTR award winners also increased odds of company success by leveraging external business support [18].

2.3 Practical Studies

Prior works have applied machine learning algorithms to identify team-independent and/or team-dependent factors associated with company success and subsequently predict the success of companies. In Table 1, we summarize the results of earlier studies deploying machine learning methods. The methods deployed include: Logistic Regression [9], Random Forest [12, 13], Gradient Tree Boosting [3], SVM [16], XGBoost [28], among others. Moreover, machine learning models have previously been applied to data collected from Crunchbase [3, 13, 16, 28], LinkedIn [21] and/or startup pitch competitions [9]. All these studies signal the potential to automatically predict the success of companies for investment purposes as concluded by the studies' results.

One study used machine learning to model SBIR/STTR data [12] to predict the success factors of technology-based companies in Phase III of the program. The study had an accuracy of about 93%, and AUC of 0.95 by using Random Forest which highlights the potential to model the SBIR/STTR data using boosting algorithms, but the study remains restricted to predicting success in Phase III of the program. Leveraging the predictive signals within the SBIR/STTR data enriched with data from Crunchbase [13, 21, 28] enables us to build a superior machine learning model to effectively discriminate between successful and unsuccessful ventures.

3 PROBLEM STATEMENT

Building upon past research, we investigate the factors of early-stage R&D ventures that predict success in the marketplace. Our investigation is novel for the following reasons: (1) most studies have focused on ventures of a certain category, product offering, market, and/or region, while less research has evaluated the success of ventures with a unifying source of seed-funding, (2) most studies investigate venture qualities in isolation; less research has used a combination of qualities (team demographics, and business profile) to predict success, and (3) our data is at a relatively large scale with granular information and has been made publicly available[1] as part of this study.

4 OUTCOME SPECIFICATION

The binary target variable was generated to capture the success of a company; more specifically, a company was considered a success if it underwent an initial public offering (IPO), and/or merged with, and/or was acquired by another entity (M&A) [4]. Following this definition, 52% of companies in the data were successful. Our definition of success is reasonable assuming that acquisition and/or IPO is a goal of small ventures and their investors [6].

5 METHODOLOGY

In this section we describe the data preparation pipeline, outcome specification, model specification, performance metrics and validation approach used in this study. All data and software needed to reproduce and extend the methodology and results described herein are publicly available online[1].

5.1 Data Preparation Pipeline

The data preparation pipeline for this work is presented in Figure 1; it is comprised of five steps each of which we describe in the following subsections (see Supplementary Materials[1] for detailed description of the first four steps):

5.1.1 ***Source and Extracted Data***.

U.S. federal small business R&D seed-fund

Small Business Innovation Research (SBIR) and Small Business Technology Transfer Awards are granted through individual government agencies. SBIR and STTR awards provide zero-equity seed-funding to small businesses (less than 500 employees) to pursue R&D activities for technological innovations that have commercial potential and are relevant to each agency's mission. In May of 2020, we collected any publicly available information on SBIR/STTR awards (amount, investigators, etc.) granted between 1983 to 2020 to 27,260 unique companies.

Company Profiles

Using the unique company names extracted from the SBIR/STTR

[1] https://github.com/abhit20/ML-Startup-Success

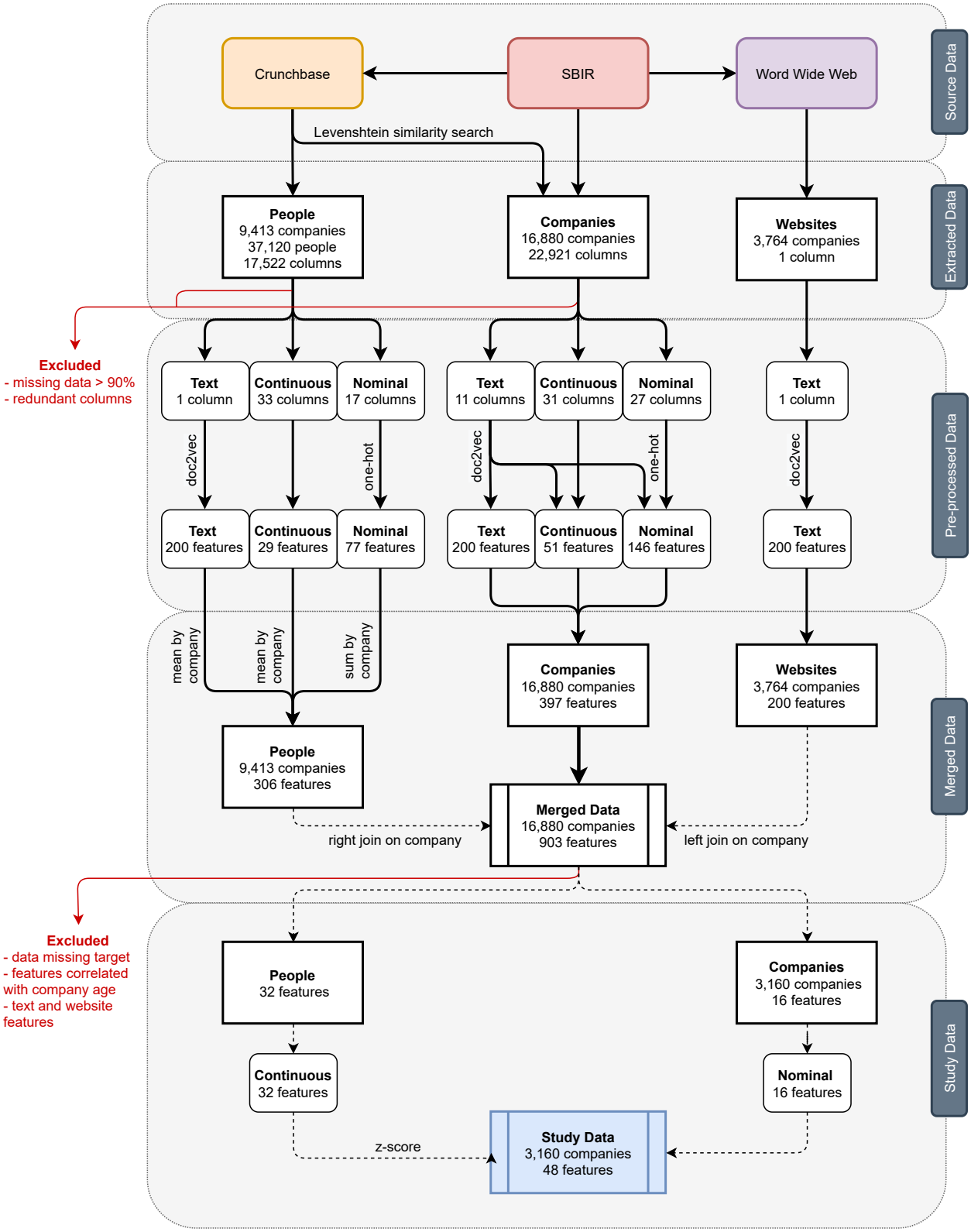

Figure 1: *Data Preparation Pipeline*: Pictorial representation of the five steps of the data preparation pipeline for this work: source data, extracted data, pre-processed data, merged data, and study data.

Table 1: Prior Practical Studies: Summary of previous studies that applied machine learning algorithms to predict the success of companies

First Author, Year	Features Used	Data Used, Sample Size	Model Used	Model Performance
Ghassemi, 2020 [9]	Peoples' academic institution, degree, major, and years since graduation; companies' target market and abstract's linguistic formality, descriptiveness, and sentiment; crowd's perceived team competence and rating of idea	2015 Massachusetts Institute of Technology $100K Launch competition, 177 Samples	Logistic Regression	AUC = 0.72
Krishna, 2016 [13]	Founded date and seed fund raised (Y/N)	Crunchbase, 11000 Samples	Random Forest	AUC = 0.6
Johnson, 2018 [12]	Number of employees, annual revenue, funding requesting and major agency, region, minority and/or women owned	SBIR, 6295 Samples	Random Forest	Accuracy = 92.2%, AUC = 0.96
Arroyo, 2019 [3]	Companies' location, age, social media presence, funding information (total funding raised, number of funding rounds, etc.), founders information (number of male and female founders, etc.)	Crunchbase, 120000 Samples	Gradient Tree Boosting	Accuracy = 82%
Li, 2020 [16]	Operation status, region, industry, funding information (number of funding rounds, first and last funding dates, etc.), funding rounds (seed, round A, etc.)	Crunchbase, 22000 Samples	SVM	Accuracy = 88%, AUC = 0.51
Zbikowski, 2021 [28]	List of organizations' categories and subcategories, founder's gender, founder's completed one degree and more than one degree, rank of region and city in number of startups, and duration of college and duration between graduation and company founding	Crunchbase, 200000 Samples	XGBoost	Accuracy = 85%, F-score = 0.43

data, we used Google Custom Search JSON API[2] and standard Google Search homepage to search and collect the links to the profile pages of the companies on Crunchbase[3]. We were able to retrieve 23,691 Crunchbase links out of the 27,230 companies in the SBIR/STTR data. Using the Crunchbase links, the data from the Crunchbase profiles of the companies was collected using a custom-built scraping and parsing tool[4]. The company name from the SBIR/STTR data was compared with the name on Crunchbase using the Levenshtein distance algorithm [15]; all companies with distances above 15 were discarded. As a result, we retained Crunchbase data for 16,880 companies.

The company-level data contained information on individuals affiliated with the company under the categories of 'Current Team' and 'Board Members and Advisors'. We collected additional information on these individuals from Crunchbase using our custom-built scraping and parsing tool. This resulted in data on 37,120 individuals from 9,413 companies.

[2]https://developers.google.com/custom-search/v1/overview
[3]https://www.crunchbase.com/
[4]https://github.com/abhit20/Crunchbase-Scraper-Parser

Word Wide Web

Utilizing the website links available in the SBIR/STTR data, we extracted the names of the companies and links to the company websites. 5,468 companies had links to their website, out of which 3,764 (69%) links were found to be active. We collected the raw HTML pages from these active websites.

5.1.2 ***Preprocessed Data.***

After removing sparse columns (greater than 90% of values missing) and columns with redundant or unnecessary information such as links to social media accounts, the data was reduced to 51 columns comprising people-level information, 69 columns representing company-level data, and 1 column of website content.

The extracted data was pre-processed into three different types of features for company-level and people-level information. The three types are the following: text features, continuous features, and nominal features. Some of the text columns were transformed into text features using Doc2Vec, continuous features, and nominal features using one-hot encoding. Following are the number of features after pre-processing: 600 text features, 80 continuous

features, and 223 nominal features. See Figure 1 and Supplementary Materials[5] for more information.

5.1.3 ***Merged Data***.

The pre-processed data for the companies, people, and websites was merged by using a few different aggregation, statistical summarization, and join methods. Specifically, peoples' text features and continuous features were averaged by company, and nominal features were summed by company (see Supplementary Materials for more details).

5.1.4 ***Study Data***.

Data was further processed with the following exclusion criteria: (1) removal of any features correlated with the age of the company (to maintain time-independence of the features), (2) removal of companies of unknown age, (3) removal of companies younger than two years old, since it would be too young to evaluate its market success; this threshold was informed by statistics from the data, a company took a minimum of two years to have an initial public offering (IPO), and/or merge with, and/or be acquired by another entity (M&A), and (4) removal of companies that had missing information on their outcome status (i.e. IPOs, M&A, etc.).

The study data was consisted of 48 features: 16 features represented company-level information, and 32 features represented people-level information. A description of the final representation of features utilized in the study are described below.

People

As mentioned in Section 5.1.3, to generate people-level features, the features of individuals affiliated with a given company were grouped together. 32 people-level features were generated and are described below:

Number of Founded Organizations (1 feature): An individual's entrepreneurial experience was represented by counting the number of organizations founded by each individual in the data.

Gender (6 features): Gender was encoded with 6 binary features for the categories of agender, male, female, non-binary, other, and prefer not to identify.

Regions (10 features): The regions listed for each individual were one-hot encoded. 9 features captured the nine unique regions inside the US while the remaining regions were captured by the final feature. Examples of regions encoded included Northeastern US, East Coast, and Midwestern US.

University Attended (2 features): An individual's academic background was one-hot encoded according to the prestige of the university they attended. A university was considered prestigious if it is an Ivy League school and/or Top 10 in WSJ/THE 2021 College Ranking List[6] and resulted in two features: prestigious, and non-prestigious academic institution.

Academic Degree (5 features): Academic degrees attained by each individual was one-hot encoded with the following categories: Bachelors, Masters, MBA, PhD, and Other.

Academic Major (8 features): Academic majors pursued by each individual was one-hot encoded with the following categories: Engineering, Business, Physical/Life Science, Social Science, Mathematics, Law, Art, and Other.

Companies

To represent a company's location, and sector, 16 features were utilized for each company, as follows:

Headquarters Regions (5 features): 4 features captured four unique regions inside the US while international regions were captured by the final feature. Examples of regions encoded included New England, and Midwestern US.

Sectors (11 features): 91 industries that companies operated in and occurred more than 100 times in the data were classified into 10 sectors (health care, finance, information technology, real-estate, communication services, industrial, energy, consumer discretionary, consumer staples, materials), while the remaining industries were classified as the 'Other' sector. These classifications were based on the Global Industry Classification Standard (GICS).

5.2 Data normalization

The probability of a company achieving an IPO or M&A (our outcome which we describe in Section 4) is likely to increase as a function of the age of the company (e.g. a company that is a year old is significantly less likely to be acquired than a ten year old firm). Hence, we are interested in building a model that can learn the attributes of a successfully company which are *independent* of the company's age. That is, we would like to remove correlations (if any) between the age of the organization, and any of our selected features.

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 and text features for all companies that shared a founding year. For instance, the one hot encoding of people's university attended for a company founded in 2001, would be z-scored using the statistical distribution of all companies founded in 2001.

To ensure fairness in modeling companies of different ages we removed potential biases in the data by (1) grouping companies that were the same age according to their years since founding (e.g. all companies that were 10 years old were grouped together), and (2) converted all continuous features into a z-score within a company's respective age group. Standardized features were the peoples' number of organizations founded, peoples' gender, peoples' location, peoples' university attended, peoples' academic degree, and peoples' academic major.

5.3 Model Specification

The data retained some missing values after the pre-processing steps described above. To account for missing data, an XGBoost model was used due to its superior performance on sparse data. More specifically, we trained an XGBoost model with the following hyperparameters: boosting rounds='100', L2 regularization='1', maximum tree depth for base learners='6', decision tree model='gbtree', learning rate=0.3 [5].

[5]https://github.com/abhit20/ML-Startup-Success

[6]https://www.wsj.com/articles/best-colleges-2021-explore-the-full-wsj-the-college-ranking-list-11600383830?st=o78haqf5zxhl4is

Table 2: *Performance Metrics*: AUC, Accuracy, Precision, Recall, F-score, and Brier Score for XGBoost model predicting whether a company had an initial public offering (IPO), and/or merged with, and/or was acquired by another entity (M&A).

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

5.4 Performance Metrics and Validation

The model in this study was assessed using leave one-out-cross validation (LOOCV). The classification performances of the model was measured using the Area Under the Receiver Operator Characteristic Curve (AUC). The AUC is a useful performance metric for problems where the cost of misclassification is not necessarily balanced, and where we wish to understand the performance of our models for various levels of misclassification tolerance. Thus, we evaluated the model True Positive Rate (TPR) and False Positive Rate (FPR) at numerous points on the Receiver Operator Curve. We also evaluated the model by using the following metrics: Accuracy, Precision, Recall, F-score (the weighted average between precision and recall). We calculated the Brier Score and plotted a Reliability Diagram [7] to evaluate the statistical calibration of the model.

5.5 Data Sharing

To facilitate reproducibility and extensions of this work, we have publicly released a de-identified version of the collected data and code in an online repository[7].

6 RESULTS

6.1 Model Performance

In Table 2 we provide the performance of the XGBoost model (assessed using LOOCV) when predicting whether a company had an initial public offering (IPO), and/or merged with, and/or was acquired by another entity (M&A). The XGBoost classifier model was found to have an accuracy of 0.84 and AUC of 0.91. In addition, the model was measured to have a precision of 0.84, recall of 0.85, and F-score of 0.84. Finally, in Figure 2, we illustrate the statistical calibration of the XGBoost model using a Reliability Diagram [7] and a Brier Score of 0.16 show a well calibrated model.

6.2 Feature Importance

In Figure 3, we have plotted the feature importance based on the mean absolute SHAP values in the model. Number of founded organizations of the individuals had the highest mean absolute SHAP value in the XGBoost classifier model. In addition, we used permutation importance with LOOCV, a model agnostic metric, to identify important features for the model. In Table 3 we show the top

[7]https://github.com/abhit20/ML-Startup-Success

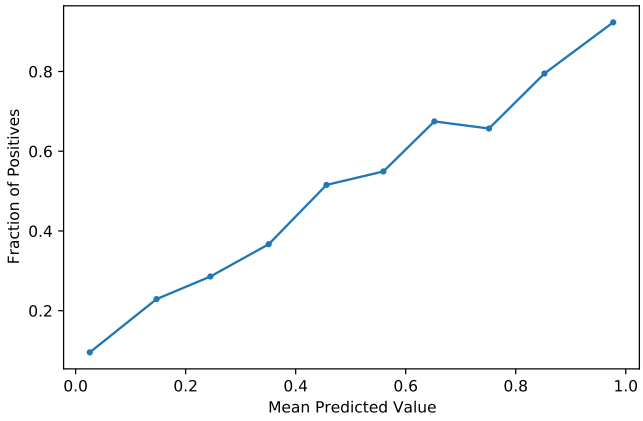

Figure 2: *Reliability Diagram*: Statistical calibration of the XGBoost model with the predictions discretized into 10 bins

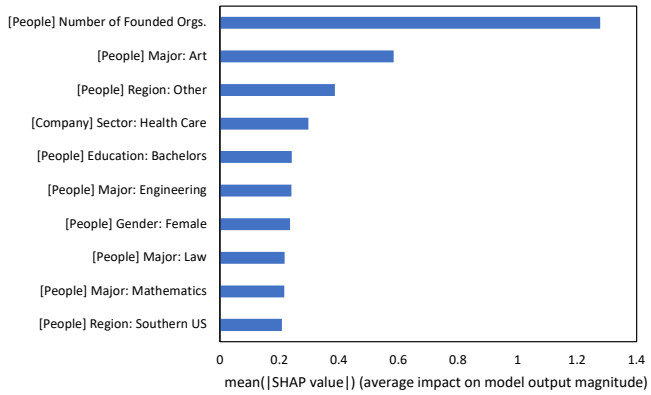

Figure 3: *SHAP Feature Importance*: Top 10 features based on the mean absolute SHAP values from the XGBoost classifier model.

Table 3: *Permutation Importance*: Top 5 features calculated using 3160 shuffles for each feature.

Feature	Weight
[People] Number of Founded Organizations	0.114 +/- 0.005
[People] Academic Major: Art	0.026 +/- 0.002
[People] Region: Other	0.018 +/- 0.002
[People] Gender: Male	0.016 +/- 0.002
[Company] Sector: Health Care	0.012 +/- 0.002

5 features calculated using 3160 shuffles for each feature. Moreover, the average number of organizations founded by individuals within a company was the top feature in predicting whether a company had an initial public offering (IPO), and/or merged with, and/or was acquired by another entity (M&A). This result implies that the number of organizations founded by individuals of a given company

Table 4: *Prediction Thresholds*: FPR at various TPR thresholds and TPR at various FPR thresholds for 3160 samples

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

captures the entrepreneurial experience of the individuals, hence the importance of the feature in predicting the success of a company.

6.3 Prediction Thresholds

In Table 4 we show the FPR at various TPR thresholds and TPR at various FPR thresholds. A TPR of 90% and 95%, the FPR is 26.3% and 46.2%, respectively. Likewise, at the given FPR of 10% and 5%, the TPR is 72.5% and 58.1%, respectively.

7 DISCUSSION

In this paper, we studied a novel dataset comprising 3,160 unique with 32 people-level features and 16 company-level features. Using the collected data, we trained an XGBoost classifier to predict a binary outcome: whether a company had a liquidation event in the form of an IPO and/or M&A. The model was validated using leave-one-out cross validation (LOOCV), and performance was evaluated using the Area Under the Receiver Operator Characteristic Curve (AUC), accuracy, precision, recall, TPR for various FPR thresholds, and a reliability diagram for statistical calibration. Our model performed strongly (0.91 AUC) based on these metrics.

7.1 Time-independent Indicators of Success

The core insight of this study is that time-independent factors of companies and their teams can be extracted to predict the *future* success of a company. More precisely, the time-independent factors of companies were able to predict if a company had an IPO and/or M&A. Our work expands the scope of previous research to include companies of different ages, product maturities, industries, and location, while accounting for the teams' academic backgrounds and work experiences.

The success factors of companies have not significantly changed across time due to the inherent stability of the fundamentals of business. The legal requirement that establishes fiduciary duty of officers and directors to the shareholders to maintain the long-term solvency of an enterprise reinforces the primary objective of business: maximize profit and returns for the equity holders. Thus, our findings indicate that factors of companies and their teams can predict the long term success within a heterogeneous investment pool.

7.2 Model Application

The model may be deployed to provide recommendations. The prediction thresholds in Table 4 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 is about 72.5%. Hence, the model can be calibrated to fit the investor's risk threshold to inform the investor's decision. Practically, the investor should choose a threshold and accept the recommendations with probabilities above that threshold to enhance the judgement of the investment.

7.3 Lack of Comparable Baseline

As reviewed earlier, many prior research efforts have focused on identifying success factors of companies belonging to specific industries or geographic regions. However, no studies focused on applying machine learning to classify a combination of SBIR/STTR and Crunchbase data. Nevertheless, there are a few studies that have applied machine learning algorithms to Crunchbase [13, 28] and/or LinkedIn data [21] but the differing data blend of these studies yields a lack of the characteristic baseline for this study to be a direct reference to. Therefore, we were unable to find and provide a comparable baseline for benchmarking the results of our study.

7.4 Future Work

Legal founding documents such as articles of incorporation and operating agreement are required when a business is incorporated. These documents outline the objectives, structure, and other factors of the venture that potentially signal the viability of a company. Future studies should augment factors extracted from these legal documents with the factors used in this study to make inferences about the long-term success of companies. Founding documents are important to investigate because the information these documents provide may prove to be a significant early indicator of corporate success

Furthermore, future research should focus on predicting the valuation of the company's IPO / M&A and the number of years it might take to accomplish that goal. Investors prefer new ventures that produce the largest returns in the shortest time while minimizing the risk of loss. Hence, predicting the valuation of the IPO / M&A and the number of years will help quantify the risk and returns of an investment while enabling the investor to account for various external investment factors such as inflation, and other macroeconomic metrics.

REFERENCES

[1] Thomas J Allen, Peter Gloor, Andrea Fronzetti Colladon, Stephanie L Woerner, and Ornit Raz. 2016. The power of reciprocal knowledge sharing relationships for startup success. *Journal of Small Business and Enterprise Development* (2016).

[2] Allen C Amason, Rodney C Shrader, and George H Tompson. 2006. Newness and novelty: Relating top management team composition to new venture performance. *Journal of Business Venturing* 21, 1 (2006), 125–148.

[3] Javier Arroyo, Francesco Corea, Guillermo Jimenez-Diaz, and Juan A Recio-Garcia. 2019. Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. *Ieee Access* 7 (2019), 124233–124243.

[4] Onur Bayar and Thomas J Chemmanur. 2011. IPOs versus acquisitions and the valuation premium puzzle: A theory of exit choice by entrepreneurs and venture capitalists. *Journal of Financial and Quantitative Analysis* (2011), 1755–1793.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Veroniek Collewaert. 2012. Angel investors' and entrepreneurs' intentions to exit their ventures: A conflict perspective. *Entrepreneurship Theory and Practice* 36, 4 (2012), 753–779.

[7] Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)* 32, 1-2 (1983), 12–22.

[8] Elli Diakanastasi, Angeliki Karagiannaki, and Katerina Pramatari. 2018. Entrepreneurial team dynamics and new venture creation process: an exploratory study within a start-up incubator. *SAGE Open* 8, 2 (2018), 2158244018781446.

[9] Mohammad M. Ghassemi, Christopher Song, and Tuka Alhanai. 2020. The Automated Venture Capitalist: Data and Methods to Predict the Fate of Startup Ventures. In *KDF at the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

[10] Paul Hribar and Holly Yang. 2016. CEO overconfidence and management forecasting. *Contemporary accounting research* 33, 1 (2016), 204–227.

[11] Mazhar Islam, Adam Fremeth, and Alfred Marcus. 2018. Signaling by early stage startups: US government research grants and venture capital funding. *Journal of Business Venturing* 33, 1 (2018), 35–51.

[12] Michele Jeffrey Johnson. 2018. *Using Machine Learning Algorithms to Predict Technology-Based Small Business Commercialization.* Ph.D. Dissertation. The George Washington University.

[13] Amar Krishna, Ankit Agrawal, and Alok Choudhary. 2016. Predicting the outcome of startups: Less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 798–805.

[14] Seol-Bin Lee. 2017. An analysis on the critical startup success factors in small-sized venture businesses. *Asia-Pacific Journal of Business Venturing and Entrepreneurship* 12, 3 (2017), 53–63.

[15] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.

[16] Jinze Li. 2020. Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forset. In *2020 2nd International Workshop on Artificial Intelligence and Education*. 5–11.

[17] Albert N Link and Christopher J Ruhm. 2009. Bringing science to market: Commercializing from NIH SBIR awards. *Economics of Innovation and New Technology* 18, 4 (2009), 381–402.

[18] Erick W Littleford. 2014. Probability of small business innovation research (SBIR) commercialization as a result of participating in the Navy's transition assistance program. (2014).

[19] Satish Nambisan and Robert A Baron. 2013. Entrepreneurship in innovation ecosystems: Entrepreneurs' self–regulatory processes and their implications for new venture success. *Entrepreneurship theory and practice* 37, 5 (2013), 1071–1097.

[20] Charles Ou and George W Haynes. 2006. Acquisition of additional equity capital by small firms–findings from the national survey of small business finances. *Small Business Economics* 27, 2-3 (2006), 157–168.

[21] Boris Sharchilev, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten de Rijke. 2018. Web-based startup success prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2283–2291.

[22] Sanjay Sharma. 2000. Managerial interpretations and organizational context as predictors of corporate choice of environmental strategy. *Academy of Management journal* 43, 4 (2000), 681–697.

[23] Mike Simpson, Joanne Padmore, and Nicki Newman. 2012. Towards a new model of success and performance in SMEs. *International journal of entrepreneurial Behavior & Research* (2012).

[24] Mike Simpson, Nicki Tuck, and Sarah Bellamy. 2004. Small business success factors: the role of education and training. *Education+ Training* (2004).

[25] Richard Sudek. 2006. Angel investment criteria. *Journal of Small Business Strategy* 17, 2 (2006), 89–104.

[26] Elisabeth J Teal and Charles W Hofer. 2003. The determinants of new venture success: strategy, industry structure, and the founding entrepreneurial team. *The journal of private equity* (2003), 38–51.

[27] Marco Van Gelderen, Roy Thurik, and Niels Bosma. 2005. Success and risk factors in the pre-startup phase. *Small business economics* 24, 4 (2005), 365–380.

[28] Kamil Żbikowski and Piotr Antosiuk. 2021. A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management* 58, 4 (2021), 102555.

[29] Y Lisa Zhao, Dirk Libaers, and Michael Song. 2015. First Product Success: A Mediated Moderating Model of Resources, Founding Team Startup Experience, and Product-Positioning Strategy. *Journal of product innovation management* 32, 3 (2015), 441–458.