The Automated Venture Capitalist: Data and Methods to Predict the Fate of Startup Ventures

Mohammad M. Ghassemi,[1,2*] **Christopher Song,**[1,3] **Tuka Alhanai**[1,4]
[1] Ghamut Corporation, Okemos MI, USA [2] Michigan State University, East Lansing MI, USA
[3] Johns Hopkins University, Baltimore MD, USA [4] New York University, Abu Dhabi, UAE

Supplementary Materials available at: https://github.com/ghamut/automated-venture-capitalist

Abstract

We investigate how the composition of early-stage start-up teams, and the properties of their ventures, predict their nomination to a premier entrepreneurship competition, and their continued operation two years following. We collected a novel dataset of 177 ventures, comprising 374 individuals. The dataset contained the characteristics of the entrants, free-text descriptions of the ventures, and crowd assessments of venture ideas. Using sixteen descriptors of each venture, we trained several models to predict both the nomination of the teams by the competition judges, and the survival of the ventures two years later. The best performing model exceeded the performance of the competition judges in predicting venture survival (AUC 0.72). We found that teams with diverse professional and academic backgrounds were more likely to survive ($p < 0.05$), while ventures with highly-optimistic business abstracts ($p < 0.03$), or ideas that targeted established markets ($p < 0.01$) were less likely to survive. Furthermore, the judgment of crowd workers were strongly associated with survival ($p < 0.02$). We conclude that while immense personal commitment, professional aptitude, and market volatility have major roles in the destiny of ventures, the quantifiable initial conditions of teams also carry predictive weight.

Introduction

Human cooperation played a significant causal role in the historical development of human language, and advanced intelligence (McNally, Brown, and Jackson 2012). Cooperation allowed for organized society, which in turn facilitated culture, science, and ever more cooperation. Cooperation, however, does not always guarantee positive outcomes. Poorly constructed teams are less effective than the sum of their parts, and even ideal teams can fail in the pursuit of challenging objectives (e.g. obtaining a grant), or when the metrics of success are beyond their control (e.g. consumer products) (McNally, Brown, and Jackson 2012). In this study we investigate the attributes of teams, and the individuals that comprise them, that contribute to their short term success, and longer term survival. Specifically, we study early stage start-up ventures, and follow their survival for two years.

*Authors contributed equally. For questions about this work, please contact research@ghamut.com

The literature outlines three factors that determine the survival or failure of teams: (1) their composition (Amason, Shrader, and Tompson 2006) (2) their objectives (Thomas and McDaniel 1990), and (3) their evaluators (Hackman 2002). Each of these factors, the interactions between them, and their association with 'survival' are subjects of extensive and ongoing investigation. For our purposes, the literature is of two varieties: investigative and practical. Investigative studies isolate previously unknown aspects of teams that are most predictive of performance (Eppler and Hoffmann 2012; Lingard et al. 2004). As their purpose is the advancement of knowledge, investigative studies are understandably less concerned with the practical deployability of findings. Indeed, features such as multi-tasking, attitude, and risk-aversion are strongly associated with team outcomes, but scalably collecting these latent factors outside of a structured setting (e.g. laboratory environment, job interview) is difficult. Conversely, practical studies identify aspects of teams that are easily collected, and could be deployed for immediate practical ends (e.g. employee screening, investment targets) (Bercovitz and Feldman 2011; Guzman and Stern 2016). Practical studies tend to investigate how success is predicted by: qualities of teams, qualities of venture ideas, and human assessment of ideas.

Prior work on the qualities of *teams* that predict success have studied the effects of demographics (age, academic status, academic specialization, and prior work experiences) (Visintin and Pittino 2014; Delmar and Shane 2006; Beckman, Burton, and O'Reilly 2007). Other investigators have studied demographics, while also accounting for technical innovation, venture strategy, and competition in the market (Eisenhardt and Schoonhoven 1990). In the context of entrepreneurship competitions specifically, the relationship between team demographics and competition outcomes has also been studied (Der Foo, Wong, and Ong 2005). Prior work on the qualities of *ventures* that predict success utilized data about when companies started, their funding levels, media coverage (e.g. Crunchbase, Techcrunch, etc.)(Krishna, Agrawal, and Choudhary 2016), and social networks (Golshan, Lappas, and Terzi 2014; Xu et al. 2016). The Entrepreneurial Quality Index, for instance, used passively collected features including geographic location of a venture, patents filed and governing state to assess the future performance of the venture (Guzman and Stern 2016).

Prior work on how effectively *human assessment* predicts success has looked at the perceptions of entrepreneurs themselves (Keh, Foo, and Lim 2002) and the metrics investors report when evaluating ventures (Sudek 2006). Other studies have attempted to define metrics (novelty, workability, relevance, and specificity) for how humans *should* evaluate ideas (Soukhoroukova, Spann, and Skiera 2012). Previous investigations have also looked at the ability of "the crowd" to evaluate ventures using information about team demographics and venture ideas (Mollick 2013).

More broadly, teams "in the wild" tend not to formally leverage the half a century of research in this area when creating companies, academic collaborations, or other ventures. One ambitious and recent attempt to formally leverage the existing knowledge was made by Google's Project Aristotle (PA). The project investigated 180 teams within Google over several years in an effort to improve team formation and cohesion within the organization. Even with unprecedented levels of information at their disposal (from team gender balance and educational background to lunch habits), the project's results were mostly inconclusive (Duhigg 2016). Such results highlight that the prediction of team performance, even for the most technically capable, is a problem without an easy solution.

The difficulty of predicting team performance may explain why even the most quantitatively inclined investment firms still use in-person meetings and "gut-feel" to evaluate investment candidates, with perceptions of team passion and trustworthiness often overriding quantitative metrics of performance (Sudek 2006); even then, rates of identification do not exceed 25% (Wiltbank et al. 2009; Gage 2012).

Our Objective

Building on the literature, we investigate how the composition of a team, and their objectives (as reflected by free-text abstracts) predict their nomination to a premier entrepreneurship competition, and their continued operation two years following their entry. Our study is novel for the following reasons: (1) most of the surveyed studies investigate how venture and team qualities predict their *current* success; less research has evaluated how the initial properties of teams predict their *future* success, (2) most studies investigate venture qualities in isolation; less research has used a combination of qualities (team demographics, business ideas, and crowd perception) to predict success and (3) our data is at the same scale (albeit of lower resolution) as PA and has been made publicly available as part of this study.

Before we proceed, it is important to highlight that creating a successful venture requires personal commitment, professional aptitude, market receptivity, and a host of other factors that are not easy to ascertain from a resume, business plan, or team presentation alone. Hence, the best that a venture capital (VC) firm, an accelerator, or a predictive algorithm can hope to identify are the teams with the right *initial conditions* of survival, and *not* the complex trajectory traversed by the teams to attain it. The prediction of such initial conditions is our goal in this paper.

Methods

Data

In May of 2015, we collected the names and members of all ventures competing in the *2015 Massachusetts Institute of Technology $100K Launch competition* (the *competition*, hereafter). Established in 1990, the competition claims credit for the creation of 160 successful companies, 4600 jobs, and $16 Billion in market capitalization (Matheson 2017). A total of 613 individuals comprising 192 teams were collected from the competition's publicly facing website. The website provided the name of each team, an abstract describing the team's idea, as well as the names, academic affiliations, and class standing of all team members. Using the name and academic affiliation provided by the competition's website, we performed an Internet search to manually curate additional information on the ventures and entrants. Of the 613 listed entrants, 374 individuals had additional publicly available photographs, information on skills, background, and experiences. Our working data comprised these 374 individuals, and the 177 teams they belonged to.

Outcomes

There were two outcomes we aimed to predict using the collected data: (1) the decisions of the competition's nomination committee, which we refer to as *nomination* and (2) the continued operation of the venture two years following their entry, which we refer to as *survival*.

Nomination Outcome The nominations were generated by the competition judging committee. In 2015, the competition committee was comprised of approximately 100 individuals with backgrounds in venture capital, technology, entrepreneurship, and industrial leadership. The committee nominated 56 of the 192 teams[1].

Survival Outcome We defined a venture as a survivor if it remained in operation for at least two years following entry into the competition. We manually determined the operational status of each organization through an Internet search involving the name of the company, and its founding members. Two years was selected as the threshold of survival because the average length of time until a seed-funded company fails is 20 months (cbi 2014).

Descriptors

Team-level Descriptors For each entrant, the following descriptors were collected: sex, academic institution (MIT, Harvard, etc.), academic status (undergraduate, graduate, postdoctoral fellow), current degree pursued (Bachelors, Masters, PhD, MBA, Other), past degrees pursued (Bachelors, Masters, MBA, PhD, MD, JD, Associates), academic major (social science, physical/life science, engineering, business, law, health, mathematics, art, and other), previous job functions (creative, assistant, engineer, entrepreneur, manager, scientist, student, marketer, military) and professional skills (engineering, hardware, software, algorithms, research, hard science, life science, health, legal, finance,

[1]Nomination criteria are available in Supplementary Materials

Table 1: *Venture Categories.* Representative examples of venture categories.

Venture Category	n	Representative Example
health	31	Clothing that detects impact of gunshot wounds
goods	26	Peer-to-peer market for personal storage
information	25	Platform presenting statistics on US legislation
transportation	19	Parking spot rental and management service
finance	14	Pre-paid debit card for international travel
social	12	Social networking for the elderly
energy	11	Gasoline delivery service
entertainment	11	Crowd-funding for independent films
impact	10	Search engine constrained by social values
education	10	Mobile fabrication lab for high school students
children	9	Toys to inspire interest in science
employment	7	On-demand employment platform
food	7	Drones to monitor crop health

education, military, management, relationships, communication, creative, and design). More information on specific majors, degrees, job titles, and skills used for the coding of the entrants may be found in the Supplementary Materials[2].

For each entrant we also collected the total number of self-indicated: honors, awards, publications, prestigious publications[3], professional experiences, professional skills, months of experience in last job, total months of experience, and years since (or until) graduation. In total, we collected 64 individual-level descriptors.

Venture-level Descriptors We also extracted information at the venture level. For each venture we collected the total number of members in the team, and aggregated the entrant level descriptors of all members within the team (mean academic status, mean academic degree, mean skills etc.). We also coded the category that best described the venture (see Table 1 for a list of categories, and representative examples).

The team abstracts were also processed to create a bag-of-words representation of the text. The abstracts contained a total of 15,811 words of which 3,655 words were unique. After removing all numbers, punctuation, special characters, stop words and words that occurred with a frequency of less than ten, our corpus contained a total of 140 unique words.

The abstracts were further processed using the Stanford coreNLP toolkit to generate Part-of-Speech tags (42 in total), named entities labels (date, duration, location, misc, none, ordinal, organization, percent, set, and time), and sentence-level sentiment scores (very positive, positive, neutral, negative, very negative) (Manning et al. 2014).

Crowd-level Descriptors Finally, we evaluated the team abstracts using three crowd-sourcing tasks. Each task required workers to select the better of two presented ideas (where 'ideas' were representative sentences from the abstract of each venture). Workers were presented with the following instructions: *"You are a judge in a prestigious entrepreneurship competition. Many startups apply for the chance to win $100K in seed funding. You are presented with two team names, and a short abstract of their idea. From the provided information, select the team you would nominate. Choose between Option A and B."* The first task asked workers to choose between ideas of the nominated and unnominated teams, the second task asked workers to choose between the ideas of surviving and failing teams, the third and final task asked workers to choose between two random teams (regardless of their outcome). For every experiment, each of the teams appeared at least ten times in the shuffling of pairwise team comparisons. The raw worker vote for each team in each experiment was retained as a descriptor. These experiments were conducted on Amazon Mechanical Turk (AMT). AMT workers received an average hourly wage of $7.20.

Lastly, we collected any publicly available profile photos of each entrant. Three research assistants independently evaluated the profile photos of each entrant according to the following instructions *"Does the following individual look competent? Based on their look alone, would you hire this person?"*. The majority of the votes across the three workers (yes/no) were used as a descriptor.

Descriptor Selection

To support model generalizability, we constrained the observation-to-descriptor ratio to be greater than ten (Peduzzi et al. 1996). This criteria required us to eliminate all but sixteen of the collected descriptors, which were selected on the basis of literature guidance, and investigator intuition. Ultimately, we selected seven venture-level descriptors, seven team-level descriptors, and two crowd-level descriptors which we describe in greater detail below.

Selected Team-Level Descriptors: For each entrant, the following seven team-level descriptors were utilized:

Academic Institution (2 descriptor): the primary academic affiliation of all entrants was encoded as a vector describing the proportion of the team affiliated with the Massachusetts Institute of Technology or Harvard University with "Other" Universities set as the reference group. Academic affiliation categories were determined on the basis of data densities: 75.1% of entrants were affiliated with MIT, the academic institution with the greatest number of entrants following MIT was Harvard (6.7%), while the remaining entrants (18.2%) held primary academic affiliations across other institutions.

Years Since Graduation (1 descriptor): To capture the professional and life experiences of teams, we used the average time since graduation from their latest degree program. A negative number represented years until graduation while a positive number represented years since graduation.

Academic Degree (2 descriptors): Each team's most recent academic degree were encoded as a vector reflecting the proportion of the team having (or working towards) an MBA degree or a PhD degree with other academic degrees set as the reference category.

Academic Major (1 descriptor): We captured the proportion of team members pursuing (or having attained) degrees in science, technology, engineering, or math (STEM).

[2]Tables S1, S2, and S3 in Supplementary Materials.

[3]Nature,Science, New England Journal of Medicine, Proceedings of the National Academy of Science, Cell, The Lancet, Journal of the American Medical Association, Chemical Reviews, Circulation, Physical Review Letters, Nature Genetics, Journal of the American Chemical Society, Nature Medicine, Journal of Clinical Oncology, Journal of Biological Chemistry

Non-Technical Skills (1 descriptor): In an entrepreneurship context some skills beyond technical training may be useful, to capture this dimension of experiences we counted the total number of non-technical skills held by team members. These skills included: creative, design, communication, legal, management, relationships, and education.

Selected Venture-Level Descriptors: For each venture, the following seven venture-level descriptors were utilized:

Target Market (3 descriptors): Ideas were categorized in accordance with their target market: fundamental (e.g. food, energy, finance), periodic (e.g. employment platform, delivery service, tax service), and inessential (e.g. social platform, high-end foods).

Linguistic Formality (2 descriptors): Ideas were processed to capture the level of linguistic formality in the written abstract including: the use of possessive pronouns (e.g. 'you', 'your') and the number of rhetorical questions asked. Both descriptors were normalized by the number of sentences in the given abstract.

Descriptive (1 descriptors): We measured how descriptive the written abstract were using the number of adjectives, normalized by the number of sentences in the abstract.

Sentiment (1 descriptors): To capture the emotive content of the abstract, we measured the number of sentences with a positive or very positive sentiment, normalized by the number of sentences in the abstract.

Selected Crowd-level Descriptors: For each venture, the following two crowd-level descriptors were utilized:

Perceived Team Competence (1 descriptor): Three research assistants independently evaluated the profile photos of each entrant. A consensus judgment of positive competence across the three judges was coded as a descriptor.

Idea Rating (1 descriptor): A continuous descriptor was generated to measure the popularity of team ideas. An idea was considered popular if at least 40% (20 out of 50) of the AMT crowd workers voted for the idea relative to other randomly selected ideas.

Models and Analysis

Using the selected descriptors, we compared the classification performance of the following modeling approaches: Decision Trees, Discriminant Analysis, Logistic Regression, Support Vector Machines, k-Nearest Neighbors (k-NN), Ensemble Learning, and Neural Networks. All Neural Networks were feed-forward, and topology optimized using grid search. The search was constrained to networks with one or two hidden layers, 0-5 nodes per layer, $tanh$ activation functions, and random initialization of weights (Gaussian with mean=0, and variance=1). Network parameters were optimized using stochastic gradient descend (with momentum) on 1000 epochs of data and an early termination condition if validation set performance diminished for six epochs. We compared the best performing 1 hidden-layer, and 2 hidden-layer network to the other modeling approaches. Then, using the best performing modeling framework, we compared approaches that used team-level descriptors, venture-level descriptors, both team- and venture-level descriptors combined, and crowd-only descriptors to predict the two outcome classes (nomination and survival).

Performance Metrics and Validation

All models in this study were assessed using leave one-out-cross validation (LOOCV). The classification performances of all models were 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. We also evaluated the False Positive Rate (FPR) and True Positive Rate (TPR) of the models at various points on the Receiver Operator Curve. For the final survival and nomination models (trained on all data) we performed the Hosmer-Lemeshow Test (HL-test) to evaluate statistical calibration and report the values of the model coefficients, odds ratio, and the statistical significance.

Proposed Cost Matrix

Investment entities are incentivized to select companies that will maximize their future market capitalization. By virtue of their early investment, such entities may assume credit for growth following their investment, but must also accept responsibility for failure. There is nothing to be gained by successfully identifying a failing venture and there is much to be lost by mistaking a failing venture as a survivor. Given that firms have limited resources, the cost of false positives (mistaking a failing venture as a survivor) may be higher than the reward of true positives (correctly identifying ventures that will survive) because false positives incur both an investment cost in addition to an opportunity cost. With this in mind, we identified a positive prediction rate that minimized the overall model cost across several different penalties, where the cost of a false positives was -0.1x, -0.5x, -1x, and -10x the cost of a true positive. We assumed the cost of a true negative to be 0, and the cost of a false negative to be 0. This comparison was performed across a range of ratios for true and false positive costs which allows for the substitution of a dollar cost.

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[4]. For privacy reasons, we maintain a higher-resolution version of the data, available upon request. Access to the restricted data requires investigators to sign a data-use agreement promising not to intentionally identify the entrants, teams, or judges of the competition.

Results

Comparison of Models

In Table S4 (see Supplemental Materials), we compare the the results of seven modeling frameworks for the prediction

[4]https://github.com/ghamut/automated-venture-capitalist

Table 2: *Judge Comparison.* Predictive performance of judges, crowd popularity and model (algorithm) both including, and excluding competition nominees that were finalists in the competition. Baseline incidence of survival was 28%.

Prediction — Truth	# Teams	Crowd % (teams)	Judges % (teams)	Algorithm % (teams)
Survival — Nomination	54	33% (18)	39% (21)	41% (22)
Nomination (excl. 7 finalists)	47	34% (16)	36% (17)	40% (19)
Survival — Survival	49	37% (18)	43% (21)	45% (22)
Survival (excl. 7 finalists)	42	36% (16)	38% (17)	42% (19)

of venture survival using the 16 selected descriptors. The logistic regression model was found to have the best LOOCV AUC of the tested approaches (AUC = 0.72), and the best TPR at various FPR thresholds (12% at 0%, 20% at 5%, and 31% at 10%). Compared to the next best performing model (Linear SVM), the logistic regression exhibited minor improvements in AUC (0.01 absolute improvement), but significant improvements in TPR at an FPR of 0% (3x relative improvement, 8% absolute improvement). Because the logistic regression approach had the best overall performance, the rest of the investigation was performed using this model.

Descriptor Impact

In Table S5 (See Online Supplemenary Materials) we show the performance of the logistic regression model using different subsets of the selected descriptors (team, idea, team+idea, and crowd only) for the prediction of survival, nomination, and survival+nomination. For the prediction of nomination, the best performing model used all descriptors (AUC of 0.63). For the prediction of survival, the best performing model also used all descriptors (AUC 0.72). For the prediction of survival, models using all descriptors tended to have higher TPR for FPR thresholds but such was not the case for the prediction of nomination.

Crowd vs. Judges vs. Algorithm

In Table 2 we compare the performance of our model (i.e. logistic regression) to the competition judges, and crowd workers for the prediction of nomination and survival. The probability of venture survival, given nomination by the judges, was 43% (21/49). In contrast, the probability of survival given selection by the crowd workers was 37% (18/49), and 45% (22/49) given selection by our model (a 2% absolve improvement over the judges). Given nomination by the judges, ventures has a 39% probability of two-year survival. In contrast, ventures selected by the crowd had a 33% probability of survival, while ventures selected by our algorithm had a 41% chance of survival. The relative performance of the crowd, competition judges, and model remained the same after excluding competition finalists from the analysis.

Table 3: *Model Coefficients.* Coefficients, odds ratios and p-values for logistic regression models predicting competition nomination and two-year survival. Nomination Model HL-test p-value: 0.0418. Survival Model HL-test p-value: 0.11.

Descriptors	Nomination Coeff.	Nomination Odds Ratio	Nomination $p <$	Survival Coeff.	Survival Odds Ratio	Survival $p <$
Intercept	-5.59	0.00	0.01	-5.67	0.00	0.01
Team Descriptors						
Educational Institution						
MIT	1.17	3.23	0.10	1.36	3.90	0.09
Harvard	2.65	14.11	0.02	2.58	13.16	0.04
Years since Graduation	0.04	1.04	0.56	-0.31	0.74	0.01
Degree Type						
MBA	0.15	1.16	0.85	2.17	8.75	0.05
PhD	0.90	2.47	0.14	-0.16	0.85	0.82
Major STEM†	0.12	1.13	0.88	2.30	9.97	0.02
Skills Non-Technical‡	-4.06	0.02	0.04	-4.21	0.01	0.06
Crowd Descriptors						
Perceived Team Competence	0.70	2.02	0.21	-1.20	0.30	0.07
Idea Rating	3.39	29.68	0.03	4.44	84.86	0.02
Venture Descriptors						
Target Market						
Inessential	0.29	1.33	0.81	-0.09	0.92	0.92
Periodic	0.54	1.71	0.64	-0.54	0.58	0.53
Fundamental	0.78	2.18	0.52	-3.58	0.03	0.01
Venture Abstract						
Possessive Pronouns	-0.31	0.73	0.85	-5.11	0.01	0.01
Questions	-2.72	0.07	0.03	1.77	5.86	0.11
Adjectives	1.22	3.39	0.01	-0.27	0.76	0.60
Positive Sentiment	0.15	1.16	0.84	-1.85	0.16	0.03

†STEM: Science, Technology, Engineering and Math.
‡Non-Technical Skills: Design, Creative, Communication, Relationships, Legal, and Education.

Model Coefficients

In Table 3 we show the coefficients, odds ratios and p-values of the logistic regression model using all sixteen of our selected descriptors for the prediction of both nomination and survival. Harvard University affiliation(s) and the opinion of the crowd were positively associated with both nomination and survival ($p < 0.03$). The opinion of the crowd was more strongly associated with survival than nomination. Ideas that the crowd consistently preferred were 30 times as likely to receive nomination, and over 80 times as likely to survive than ideas the crowd did not consistently prefer. The usage of adjectives in the venture abstracts had a positive association with nomination ($p < 0.01$); for every one additional adjective per-sentence, the odds of nomination increased by 3.4 times. Teams with exclusively non-technical skills (e.g. management and communication) were nearly 50 times less likely to be nominated ($p < 0.04$) than teams with exclusively technical skills. The use of rhetorical questions in venture free-text abstracts had a negative association with nomination ($p < 0.01$); abstracts that were entirely rhetorical were 14 times less likely to receive nomination.

Teams with STEM major(s) had a positive association with survival ($p < 0.02$); teams composed entirely of STEM major(s) were nearly 10 times as likely to survive as teams composed entirely of non-STEM majors. The average time since graduation was negatively associated with venture survival ($p < 0.05$); for every one additional year since graduation (on average), the odds of survival diminished by 1.35 times. Ventures that targeted fundamental daily needs (e.g. energy, water) were 30 times less likely to survive ($p < 0.01$) than ventures which targeted other sectors. Venture free-

text abstracts that used possessive pronouns ('you', 'your', etc.) were negatively associated with survival ($p < 0.01$); for every one additional possessive pronoun per-sentence, the odds of survival decreased by 100 times. Positive sentiment was negatively associated with venture survival ($p < 0.03$); venture abstracts with consistently positive sentiment were approximately 6 times less likely to survive than venture abstracts without any positive sentiment. The model was found to be well calibrated according to the HL-test.

Discussion

In this study we collected a novel dataset characterizing 177 ventures (and 374 individuals) that competed in a premier entrepreneurship competition. The collected information was condensed into: seven venture-level descriptors reflecting target-market and abstract style, seven team-level descriptors of educational background and skills, and two crowd-level descriptors of idea quality and team competence (based on visual appearance). Using the sixteen descriptors, we trained a variety of models to predict two binary outcomes: (1) the decisions of the competition's nomination committee, and (2) the continued operation of the ventures two years following their entry. We tested all models using leave-one-out cross validation, and evaluated their performance using the Area Under the Receiver Operator Curve, the Hosmer-Lemeshow test for statistical calibration (HL-test), and TPR for various FPR thresholds (0%, 5% and 10%). Our algorithm out-performed the competition judges in the prediction of team survival (2% absolute improvement).

Model Interpretation

Only two descriptors were positively associated ($p < 0.05$) with both nomination and survival (Table 3): Harvard affiliation, and crowd preference. One possible reason why Harvard affiliation improved outcomes are the constraints of the competition itself. Team affiliation with MIT is a prerequisite of entry, hence, Harvard students must overcome a higher barrier to participate, inadvertently selecting for the more resourceful among them (Costello and Keane 2000). For both nomination and survival, venture assessment by the crowd was the strongest positive indicator. The significance of crowd-preference is sensible, and supported by recent studies (Costello and Keane 2000; Mollick 2013; Soukhoroukova, Spann, and Skiera 2012). If the market is ultimately driven by the end consumer, then a random sampling of judgment from such consumers should correlate with venture survival, as well as the judgment of investors (which are, at least in-part, attempting to predict the receptivity of the end consumers to the ideas).

Unsurprisingly, Judges based at an "Institute of Technology" showed strong preference against teams with non-technical skills when selecting nominees. Furthermore, while such judges preferred well-described ideas (adjective use), they also penalized teams that were gratuitous in their language use (rhetorical questions) (Dean et al. 2006; Hindle and Mainprize 2006).

We found that technical teams (STEM majors) composed of recent graduates were more likely to survive while teams that used less professional communication techniques (possessive pronouns), targeted fundamental markets (energy, finance, food), or were overly optimistic in their written abstracts (positive sentiment) were more likely to fail. The importance of STEM majors may be related to the increasingly technical demands of modern employment (Machin and Van Reenen 1998; Mollick 2013). The effects of graduation may result from younger teams having less existing professional or social commitments. Fundamental markets are often saturated, highly regulated, and resistant to rapid change (Machin and Van Reenen 1998; Mollick 2013), increasing the difficulty of penetration. Finally, teams that were overly optimistic in product descriptions may be seen as "too good to be true" by customers, while teams that utilized possessive pronouns ("you", "your", etc.) may be perceived as unprofessional.

Comparing Judges to Models

The limited overlap in descriptors that predicted both nomination and survival indicates that the requirements for nomination differ from the requirements for survival alone. However, these outcomes are not completely independent. It is reasonable to assume that judges anticipate team survival as a necessary condition of nomination. We find evidence for this assumption within the collected data itself: 43% of teams that survived (21 out of 54) were nominated while only 23% (28 out of 123) of un-nominated teams went on to survive. If judges only nominated ventures they believed would survive, then we may compare the predictive power of judge nomination against our survival model. This comparison revealed a 2% absolute improvement in the prediction of survival using our model (22 out of 54), relative to the judges. Such results provide evidence that automated approaches using publicly available data can assess the viability of teams as well as experts, using only a fraction of the available information.

Model Application

Our survival model may be deployed in two ways. First, investments may be scaled according to the probability of survival predicted by the model, which was well-calibrated according to the Hosmer-Lemeshow goodness of fit test ($p > 0.05$). In Figure 1, we illustrate the calibration of the model for various predicted probabilities of survival. Second, the model may also be used in a discriminative fashion. In Figure 2, we show the expected gains/costs to an organization deploying our survival model. Expected costs are shown as a function of various model classification thresholds and cost trade-offs. For each cost trade-off in the figure, we display the optimal classification threshold that maximizes projected return on investment. For practical model use, investors may select a preferred cost structure, and disregard all ventures with probabilities of survival below the corresponding classification threshold, after which evaluators may apply qualitative metrics (e.g. interviews) to further screen teams.

Nomination vs. Survival

We found it was easier to predict the survival of a team over its nomination to the competition (AUC 0.72 vs. 0.63). One

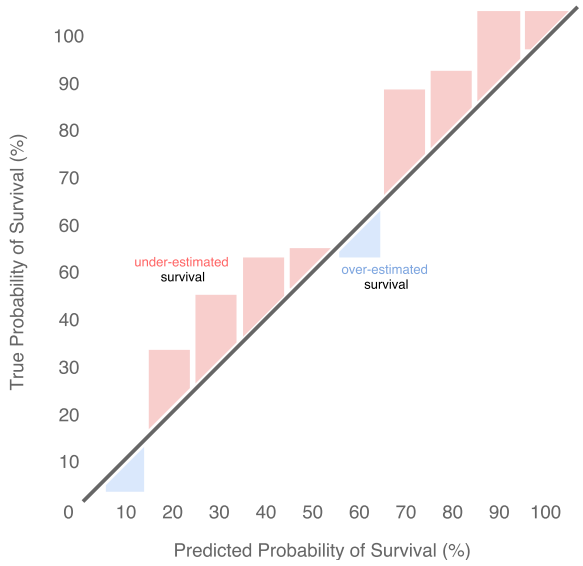

Figure 1: Calibration plot of the best performing survival model. Red bars represent underestimation of survival probability while blue bars represent overestimation of survival probability. Difference in predicted and actual probabilities was statistically insignificant (HL-test = $p > 0.05$).

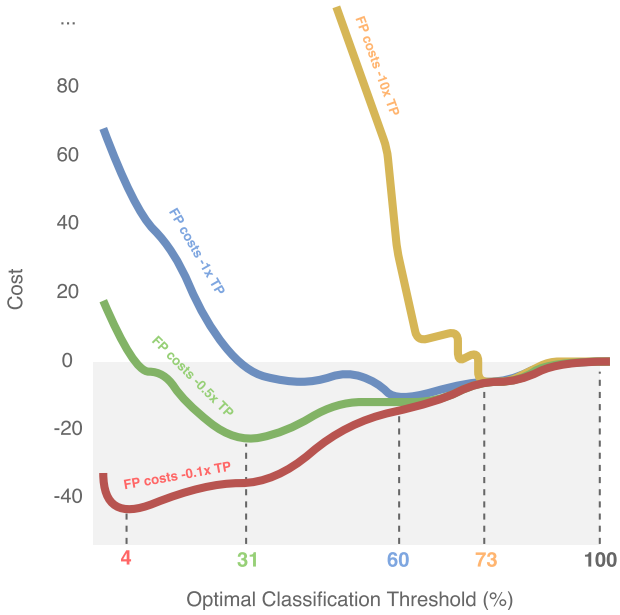

Figure 2: Cost plot displaying the expected return on investment for several false positive costs (-0.1x, -0.5x, -1x, and -10x the cost of a true positive). The dotted lines display the optimal classifier probabilistic threshold that maximizes return on investment.

possible reason for this is the bias of individual judges. Let us assume that judges have a noisy estimate of an idea's potential for survival in the market, but also suffer from a bias that randomly perturbs their judgment. It is reasonable to expect that as the number of judges increases, the difference between their average judgment and the idea's true potential for survival will decrease (Heyde 2006). However, because the number of judges evaluating each idea in the competition is relatively small (4-6 judges per team), the random effects of individual bias confound the nomination and increase the difficulty of robustly modeling the decision process. Furthermore, the information available to our model was a relatively small subset of what was available to the judges including: an extended abstract, knowledge of the team's equity holdings, patents, funding sources, and an (optional) video. It is also possible that the criteria of nomination is more complex than the criteria of survival.

Venture Abstracts

The significance of venture abstract descriptors for the prediction of team survival is surprising because, unlike in the competition setting, the venture abstract is unlikely to be assessed by stakeholders two years after the competition's entry. Indeed, we expect that materials generated for marketing to customers, pitching to investors, or attracting employees will differ substantially from what was submitted to the competition, and may even be generated by different people. However, the significance of the venture abstract descriptors implies that a team's initial style and outlook has implications for their future performance. Indeed, prior work has demonstrated that an entrepreneur's *current* outlook is predictive of their *immediate* venture performance (Keh, Foo, and Lim 2002), but our work implies that a venture's outlook may also be predictive of *future* performance (Cooper, Woo, and Dunkelberg 1988).

The Wisdom of the Crowd

Prominent investment firms pride themselves on having a sense of the consumer market which then qualifies them to select ideas and teams that are likely to thrive within those markets. If the market is in fact driven by the end consumer, it stands to reason that a random sampling of judgment from the consumers may correlate with the judgment of the investment firms. Crowd-funding platforms are built upon this very notion. Indeed, recent studies suggest that the general public can, in the aggregate, do as well as experts, when tasks are structured correctly (Soukhoroukova, Spann, and Skiera 2012; Mollick 2013). Although the crowd generated assessments of entrant competence were not found to be statistically significant, we found that the crowd's rating of the idea exhibited a statistically significant univariate association with the outcome ($p < 0.05$) and a positive association ($p < 0.05$) with survival and nomination after adjusting for other factors. It is remarkable that with only a single sentence describing the idea (as opposed to the full abstract paragraph), the assessment of the crowd workers was still predictive of long-term survival, as well as nomination. If crowd ratings of ideas were taken as probabilities of nomination and survival, they would perform with within 2% absolute (36%, 16 out of 42) of the competition judges' for the prediction of survival (38%, 17 out of 42), after excluding finalists.

References

Amason, A. C.; Shrader, R. C.; and Tompson, G. H. 2006. Newness and novelty: Relating top management team composition to new venture performance. *Journal of Business Venturing* 21(1):125–148.

Beckman, C. M.; Burton, M. D.; and O'Reilly, C. 2007. Early teams: The impact of team demography on vc financing and going public. *Journal of Business Venturing* 22(2):147–173.

Bercovitz, J., and Feldman, M. 2011. The mechanisms of collaboration in inventive teams: Composition, social networks, and geography. *Research Policy* 40(1):81–93.

2014. Startup death data. cbinsights.

Cooper, A. C.; Woo, C. Y.; and Dunkelberg, W. C. 1988. Entrepreneurs' perceived chances for success. *Journal of business venturing* 3(2):97–108.

Costello, F. J., and Keane, M. T. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science* 24(2):299–349.

Dean, D. L.; Hender, J. M.; Rodgers, T. L.; and Santanen, E. 2006. Identifying good ideas: constructs and scales for idea evaluation.

Delmar, F., and Shane, S. 2006. Does experience matter? the effect of founding team experience on the survival and sales of newly founded ventures. *Strategic Organization* 4(3):215–247.

Der Foo, M.; Wong, P. K.; and Ong, A. 2005. Do others think you have a viable business idea? team diversity and judges' evaluation of ideas in a business plan competition. *Journal of Business Venturing* 20(3):385–402.

Duhigg, C. 2016. What google learned from its quest to build the perfect team. *The New York Times Magazine* 26:2016.

Eisenhardt, K. M., and Schoonhoven, C. B. 1990. Organizational growth: Linking founding team, strategy, environment, and growth among us semiconductor ventures, 1978-1988. *Administrative science quarterly* 504–529.

Eppler, M. J., and Hoffmann, F. 2012. Does method matter? an experiment on collaborative business model idea generation in teams. *Innovation* 14(3):388–403.

Gage, D. 2012. The venture capital secret: 3 out of 4 startups fail. *Wall Street Journal* 20.

Golshan, B.; Lappas, T.; and Terzi, E. 2014. Profit-maximizing cluster hires. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1196–1205. ACM.

Guzman, J., and Stern, S. 2016. The state of american entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 15 us states, 1988-2014. Technical report, National Bureau of Economic Research.

Hackman, J. R. 2002. Why teams don't work. In *Theory and research on small groups*. Springer. 245–267.

Heyde, C. 2006. Central limit theorem. *Encyclopedia of Actuarial Science*.

Hindle, K., and Mainprize, B. 2006. A systematic approach to writing and rating entrepreneurial business plans. *The journal of private equity* 9(3):7–22.

Keh, H. T.; Foo, M. D.; and Lim, B. C. 2002. Opportunity evaluation under risky conditions: The cognitive processes of entrepreneurs. *Entrepreneurship theory and practice* 27(2):125–148.

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

Lingard, L.; Espin, S.; Evans, C.; and Hawryluck, L. 2004. The rules of the game: interprofessional collaboration on the intensive care unit team. *Critical care* 8(6):R403.

Machin, S., and Van Reenen, J. 1998. Technology and changes in skill structure: evidence from seven oecd countries. *The Quarterly Journal of Economics* 113(4):1215–1244.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, 55–60.

Matheson, R. 2017. Mit $100k winner's optical chips perform ai computations at light speed. MITNews.

McNally, L.; Brown, S. P.; and Jackson, A. L. 2012. Cooperation and the evolution of intelligence. In *Proc. R. Soc. B*, rspb20120206. The Royal Society.

Mollick, E. R. 2013. Swept away by the crowd? crowdfunding, venture capital, and the selection of entrepreneurs.

Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T. R.; and Feinstein, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology* 49(12):1373–1379.

Soukhoroukova, A.; Spann, M.; and Skiera, B. 2012. Sourcing, filtering, and evaluating new product ideas: An empirical exploration of the performance of idea markets. *Journal of Product Innovation Management* 29(1):100–112.

Sudek, R. 2006. Angel investment criteria. *Journal of Small Business Strategy* 17(2):89.

Thomas, J. B., and McDaniel, R. R. 1990. Interpreting strategic issues: Effects of strategy and the information-processing structure of top management teams. *Academy of Management journal* 33(2):286–306.

Visintin, F., and Pittino, D. 2014. Founding team composition and early performance of university—based spin-off companies. *Technovation* 34(1):31–43.

Wiltbank, R.; Read, S.; Dew, N.; and Sarasvathy, S. D. 2009. Prediction and control under uncertainty: Outcomes in angel investing. *Journal of Business Venturing* 24(2):116–133.

Xu, H.; Yu, Z.; Yang, J.; Xiong, H.; and Zhu, H. 2016. Talent circle detection in job transition networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 655–664. ACM.