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

Supplementary Materials

Nomination Outcome - Judge's Criteria

1. "Evaluate the value creation proposition along a scale of 1 to 7, with 7 indicating a business plan that has clearly identified a significant problem and compellingly articulated a scalable solution to it."
2. "Evaluate the value capture proposition along a scale of 1 to 7, with 7 indicating a business plan that has persuasively articulated how it will earn revenue from the market's adoption of its solution."
3. "Evaluate the capability of the team to leverage prize money and execute their business plan along a scale of 1 to 7, with 7 indicating a team you would personally invest in to launch this proposal."

Limitations

There are two potential limitations of this work. First, our criteria for survival was dependent on teams having a publicly detectable on-line footprint two years following their entry into the competition. There are circumstances where teams may have survived, but gone undetected in our public domain search. Teams may have been operating in "stealth mode", or may have changed their names between entry in May of 2015 and May of 2017. We note however that the very act of competition entry creates public information on the ventures. Hence, it is unlikely that entrants sought to occlude their existence. Second, although entrants came from an impressive variety of backgrounds, it is important to highlight that a majority of entrants (85%) were living and working in the greater Boston area at the time of the competition. Indeed, just as the results of Google's Project Aristotle may not generalize outside the context of Google, so too is our analysis constrained by the environment in which we collected the data. More specifically, it is unclear how well the results of this study will generalize outside the immediate geographic context, and culture of the greater Boston area (or even future competition years within Boston). Having acknowledged this limitation, our method and results have implications for team building beyond entrepreneurship, and may be extended by other members of the community as more data is collected, and more advanced methods are applied.

Acknowledgements

The authors would like to acknowledge Joshua Patterson for supporting the data collection efforts, and Felipe Torres for supporting the data annotation efforts.

Table S1: Grouping of specific academic majors into general major categories.

General Major Category	**Specific Academic Majors**
Social Science	anthropology, political science, media arts and sciences, psychology, social science, education, architecture
Physical/Life Science	biology, cognitive science, chemistry, material science, physics, geophysics
Engineering	aeroastro, biomedical engineering, chemical engineering, civil engineering, industrial engineering,mechanical engineering, bioengineering, nuclear engineering, electrical engineering, computer science, computer engineering, engineering, bioinformatics
Business	business, finance
Law	law
Health	health, medicine
Mathematics	mathematics, economics
Art	film, art, music
Other	agriculture, unknown

Table S2: Grouping of specific job titles into general job categories.

Job Title Category	**Specific Job Titles**
Creative	Musician, Artist, Stylist, Editor, Architect, Designer
Assistant	Assistant, Secretary, Clerk, Receptionist, Babysitter, Volunteer
Engineer	Analyst, Engineer, Technician
Entrepreneur	Entrepreneur, Founder, Owner
Manager	Manager, Board Member, C-level, President Vice President, Investor, Partner, Producer,
Scientist	Professor, Researcher, Scientist
Intern	Intern, Student
Marketer	Marketer, Salesman
Military	Captain, Pilot, Officer, Deputy, Defense
Other Professional	Teacher, Lawyer, Physician

Table S3: Entrant Skill Set

Skill Class	n	Specific Skills
algorithms	176	algorithms, analysis, analytics artificial intelligence, big data, business analysis, co degeneration, competitive analysis, data analysis, database design, data center, data mining data structures, distributed systems, econometrics, economics, intelligence analysis, machine learning, malware analysis, mathematical modeling, mathematics, modeling, natural language processing, numerical analysis, numerical simulation, optimization, pattern recognition, probability, programming, root cause analysis, scientific computing, signal processing, simulations, spatial analysis, statistics, ubiquitous computing, wireless
communication	138	advertising, argumentation, blogging, brandmanagement, broadcast, chinese, commercialization, communityoutreach, consulting, continuousimprovement, corporatecommunications, customerexperience, customerservice, digitalcommunication, digitalmarketing, digitalmedia, digitaltv, documentaries, environmentalawareness, fluentinspanish, foreignlanguages, french, fundraising, german, grants, grantwriting, inboundmarketing, integratedmarketing, japanese, legalwriting, mandarin, marketentry, marketing, marketingmanagement, marketingstrategy, marketplanning, marketresearch, mediarelations, multimedia, networkcommunications, onlineadvertising, onlineretail, processconsulting, productmarketing, proposalwriting, publicspeaking, retailsales, sales, scienceoutreach, serviceclient, socialmedia, socialmediamarketing, socialnetworking, spanish, spanish speaking, spoken chinese mandarin, strategiccommunications, strategicconsulting, telecommunications, television, turkish, volunteerrecruiting, webservices, writing
creative	30	3dmodeling, 3dprinting, adobecreativesuite, art, creativeconceptdesign, creativeservices, creativewriting, culinaryskills, drawing, dyeing, editing, english, featurefilms, filmproduction, fineart, music, photography, postproduction, sculpture, sketching, textiles, thinfilms, videoediting, videoproduction, violin, virtualreality
design	60	adobecreativesuite, architecture, autocad, cad, circuitdesign, creativeconceptdesign, curriculumdesign, databasedesign, design, designengineering, designofexperiments, designresearch, designstrategy, designthinking, engineeringdesign, experimentaldesign, gamedesign, gameplay, humancomputerinteraction, illustrator, indesign, industrialdesign, instructionaldesign, integratedcircuitdesign, interactiondesign, knitwear, networkdesign, objectorienteddesign, organizationaldesign, pcbdesign, pencilrendering, photoshop, productdesign, rhino, softwaredesign, solutionarchitecture, sustainabledesign, textiledesign, urbandesign, urbanplanning, userexperience, userexperiencedesign, userinterface, userinterfacedesign, webdesign
education	21	activelearning, arthistory, childcare, curriculumdesign, educationalconsulting, educationaltechnology, educationreform, highereducation, medicaleducation, publiceducation, teaching, training, tutoring
engineering	109	acoustics, arcgis, automation, automotive, aviation, biomedicalengineering, chemicalengineering, circuitdesign, ciscotechnologies, civilengineering, cleantech, cloudcomputing, comkatcompartmentmodeling, computergraphics, computerscience, computersecurity, computervision, designengineering, electricalengineering, electricalsafety, electronics, embeddedsystems, energyefficiency, engineering, engineeringdesign, engineeringmanagement, gis, helicopters, imageanalysis, imageprocessing, informatics, informationarchitecture, informationretrieval, informationtheory, itdevelopment, leadgeneration, leanmanufacturing, logistics, manufacturing, mechanicalengineering, networkcommunications, networkdesign, networksecurity, oil/gas, operationsresearch, petroleum, processengineering, processimprovement, processoptimization, projectengineering, propulsion, reactionengineering, reactor, semiconductors, sixsigma, softwareengineering, storage, supplychain, systemsengineering, telecommunications, tensiletesting, testing, troubleshooting, unmannedvehicles, waterresources, webdevelopment
finance	140	accounting, alternativeinvestments, assetmanagement, banking, bloomberg, bookkeeping, capitalmarkets, corporatefinance, cpa, creditrisk, derivatives, ecommerce, equities, equityresearch, equityvaluation, finance, financialanalysis, financialmarkets, financialmodeling, financialreporting, financialstructuring, fixedincome, fixedincomeanalysis, forecasting, hedgefunds, humancapital, internetbanking, investmentbanking, investments, loans, microfinance, mobilebanking, options, payroll, portfoliomanagement, privateequity, proprietarytrading, realestatedevelopment, securities, tax, technicalaccountingresearch, tradefinance, trading, upstream, valuation, venturecapital
hard science	93	aerodynamics, aerospace, airfreechemistry, analyticalchemistry, appliedmathematics, appliedphysics, catalystdevelopment, chemistry, climatechange, computationalfluiddynamics, electromagnetics, energy, experimentaldesign, experimentation, finiteelementanalysis, fundamentalanalysis, geography, heattransfer, hydrology, materialsscience, microscopy, moleculardynamics, molecularmodeling, nanomaterials, nanoparticles, nanotechnology, nuclear, optics, pharmaceuticalindustry, photonics, physicalchemistry, physics, polymercharacterizationwithnmr,gpc,dsc,maldi, polymers, reactionkinetics, renewableenergy, scanningelectronmicroscopy, science, silicon, solarenergy, steel, surfacechemistry, sustainableenergy, ultrasonics, wind
hardware	41	arduino, cnclathe, cncmill, digitalfabrication, instrumentation, integratedcircuitdesign, medicaldevices, microcontrollers, microfabrication, microprocessors, mobiledevices, nanofabrication, nmr, pcbdesign, powerelectronics, rapidprototyping, rfid, rfidapplications, roboticfabrication, robotics, sensors, servers, waterjet, welding, westernblotting, wirelessnetworking, woodworking, wordpress
health	41	autismspectrumdisorders, clinicaldevelopment, clinicalresearch, clinicaltrials, criticalcare, developmentaldisabilities, epidemiology, healthcare, healthcareinformationtechnology, healthpolicy, hospitals, immunology, internalmedicine, medicine, publichealth, speechtherapy, sports, surgery
legal	55	administrativelaw, corporategovernance, courts, energypolicy, environmentalcompliance, environmentalimpactassessment, environmentalpolicy, government, humanrights, internationaldevelopment, internationalsecurity, ip, legalresearch, legalwriting, litigation, patentprosecution, policy, politicalscience, politics, publicpolicy, socialjustice, sustainability, tcp/ip
life science	80	assaydevelopment, biochemistry, biofuels, bioinformatics, biomass, biomechanics, biomedicalengineering, biophysics, biostatistics, biotechnology, cancer, cellbiology, chemicalbiology, cognition, cognitivescience, computationalbiology, drugdelivery, drugdiscovery, flowcytometry, fluorescencemicroscopy, gelelectrophoresis, genetics, immunofluorescence, immunohistochemistry, invitro, invivo, lifesciences, massspectrometry, medicalimaging, medicalresearch, molecularbiology, molecularcloning, neuroscience, organicchemistry, organicsynthesis, photolithography, spectroscopy, systemsbiology
management	413	agilemethodologies, assetmanagement, brandmanagement, businessdevelopment, businessintelligence, businessplanning, businessprocess, businessstrategy, changemanagement, commissioning, corporatedevelopment, demandgeneration, economicdevelopment, emergingmarkets, engineeringmanagement, entrepreneurship, eventplanning, globalbusinessdevelopment, goaloriented, innovation, integration, internationallogistics, internationalprojectmanagement, itmanagement, itstrategy, leadership, management, managementconsulting, marketingmanagement, mergers&acquisitions, motivatedselfstarter, nationalsecurity, newbusinessdevelopment, nonprofitadministration, nonprofits, operationalplanning, operationsmanagement, organizationaldesign, personnelmanagement, portfoliomanagement, procurement, productdevelopment, productmanagement, programmanagement, projectcoordination, projectmanagement, projectplanning, riskmanagement, socialentrepreneurship, stakeholdermanagement, startups, strategicpartnerships, strategicplanning, strategy, stronglyselfmotivated, structuredproducts, successdriven, supplychainmanagement, sustainabledevelopment, teammanagement, testdrivendevelopment, timemanagement, vendormanagement, watermanagement
military	25	army, defense, military, militaryexperience, militaryoperations, navy, security, submarines
relationship	35	crossfunctionalteamleadership, highlyselfmotivated, internationalrelations, negotiation, networking, publicrelations, teambuilding, teamleadership, teamwork
research	49	clinicalresearch, designresearch, equityresearch, legalresearch, marketresearch, medicalresearch, operationsresearch, qualitativeresearch, qualityassurance, quantitativeresearch, research, technicalaccountingresearch
software	354	android, c, comsol, comsolmultiphysicssoftware, css, django, eclipse, enterprisesoftware, finalcutpro, fortran, fpga, git, googleanalytics, gps, html, html5, iosdevelopment, java, javascript, joomla, jquery, labview, latex, linux, mathematica, matlab, microsoftexcel, microsoftoffice, microsoftsqlserver, microsoftword, mobileapplications, mongodb, msexcelpivottables, multithreading, mysql, objectivec, objectorienteddesign, perl, php, powerpoint, python, rubyonrails, saas, schematiccapture, shellscripting, simulink, sketchup, softwaredefinedradio, softwaredesign, softwareengineering, solidworks, sql, stata, verilog, vhdl, visio, visualc, vpn, windows, windowsserver, xml

Table S4: *Model Results.* The performance of all modeling frameworks for the prediction of venture survival, given access to our 16 descriptors (crowd, venture, and team).

Model ModelForm	**AUC**	**TPR at FPR**			**FPR at TPR**		
		10%	**5%**	**0%**	**90%**	**95%**	**100%**
Logistic Regression	0.72	31%	20%	12%	62%	73%	100%
SVM							
Linear	0.71	18%	12%	4%	65%	77%	87%
Cubic	0.61	14%	6%	2%	77%	95%	95%
Qudratic	0.62	12%	6%	4%	82%	94%	95%
Medium Gaussian	0.67	24%	12%	0%	82%	92%	97%
Coarse Gaussian	0.67	10%	6%	0%	67%	71%	88%
Discriminant Analysis							
Linear	0.68	31%	14%	2%	61%	86%	88%
Quadratic	0.71	33%	8%	0%	50%	69%	77%
Decision Tree							
Simple	0.45	0%	0%	0%	92%	100%	100%
Medium	0.51	6%	0%	0%	100%	100%	100%
Complex	0.48	6%	0%	0%	100%	100%	100%
Ensemble							
RUSBoosted Trees	0.65	33%	27%	0%	83%	91%	95%
Boosted Trees	0.65	35%	22%	0%	71%	95%	100%
Bagged Trees	0.62	31%	20%	2%	80%	89%	94%
k-NN							
Fine	0.63	0%	0%	0%	100%	100%	100%
Medium	0.65	6%	6%	0%	92%	92%	100%
Coarse	0.62	4%	4%	2%	69%	90%	94%
Cosine	0.63	8%	2%	0%	91%	100%	100%
Cubic	0.65	8%	4%	0%	89%	89%	100%
Weighted	0.69	18%	14%	0%	82%	88%	100%
Neural Networks							
1 Layer, 4 nodes	0.69	29%	10%	0%	63%	78%	97%
2 Layers, 5x4 nodes	0.66	27%	16%	0%	83%	87%	100%

Table S5: *Descriptor Exploration.* Our model's performance for the prediction of competition nomination, and venture survival given access to different descriptor subsets.

Outcome Predicted Descriptors	**AUC**	**TPR at FPR**			**FPR at TPR**		
		10%	**5%**	**0%**	**90%**	**95%**	**100%**
Survival							
Team, Venture, Crowd	0.72	31%	20%	12%	62%	73%	100%
Venture, Crowd	0.65	24%	12%	4%	66%	74%	100%
Team, Crowd	0.61	24%	8%	0%	80%	89%	92%
Crowd	0.54	8%	0%	0%	78%	80%	92%
Nomination							
Team, Venture, Crowd	0.63	24%	15%	0%	79%	84%	100%
Venture, Crowd	0.58	22%	19%	0%	87%	95%	100%
Team, Crowd	0.58	17%	13%	0%	85%	92%	100%
Crowd	0.54	13%	13%	2%	89%	89%	96%
Nomination & Survival							
Team, Venture, Crowd	0.63	24%	10%	0%	72%	100%	100%