The Automated Venture Capitalist

Data and Methods to Predict the Fate of Startup Ventures

Mohammad M. Ghassemi
ghassemi.xyz

Christopher Song

Tuka Alhanai
talhanai.com

The Authors

A team of researchers, students, and entrepreneurs

Mohammad Ghassemi, Ph.D.

Is an Assistant Professor of Computer Science at MSU where he develops methods that combine human and machine intelligence to solve complex problems. He is a former BCG consultant with over a decade of technical and strategic experience.

Christopher Song

Is an senior at Johns Hopkins University where he studies both Cognitive Science and Computer Science.

Tuka Alhanai, Ph.D.

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

BCG BOSTON CONSULTING GROUP

MIT Technology Review Innovators Under 35

S&P Global

Massachusetts Institute of Technology

UNIVERSITY OF CAMBRIDGE

NYU ABU DHABI

MICHIGAN STATE UNIVERSITY

JOHNS HOPKINS UNIVERSITY

The Motivating Challenge

It is unclear why some ventures succeed and others fail

The iPad succeeded where the Tablet PC failed

In 2001, Bill Gates took the stage at the COMDEX computer show to reveal Microsoft's new Tablet PC. Gates predicted that within five years, it would be the most popular form of PC sold in America. It wasn't. When Apple released their iPad almost a decade later, it was able to succeed where the Tablet PC failed. Why did the iPad succeed where the Tablet failed?

A very hard problem, even for the best investors

Investors, accelerators, and entrepreneurship competitions all implicitly believe it's possible to spot a winning team and idea, but even the best investors are wrong 75% of the time.

So many factors: team, venture, social, financial …

A team's composition, objectives, evaluators and grit all play a role but success is not deterministic. Market conditions, timing, finances, team chemistry and government regulations also play a role. With so many moving parts, predicting outcomes is immensely challenging.

Research Objectives and Contributions

Can the initial conditions of teams can be used to predict their 2-year survival?

1. Collect and curate a novel competition dataset

2. Predict future survival, given initial conditions

3. Balance model interpretability & performance

4. Release code and data to make extensions easy

Data, Features and Outcomes

There were 374 individuals in 177 competition teams; 54 were nominated, 49 survived 2 years on

Teams collected from MIT$100K competition

Established in 1990, the competition claims credit for the creation of 160 successful companies, 4,600 jobs, and $16 Billion in market capitalization. There were 613 entrants comprising 192 teams.

Collected data from 177 Teams, 374 individuals

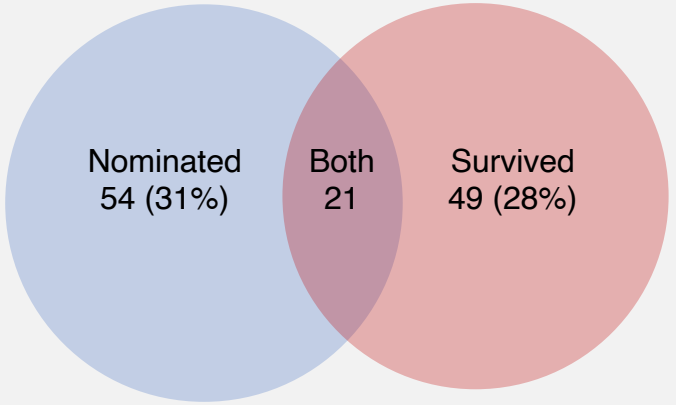
We collected a novel dataset of 177 ventures comprising 374 individuals who entered the MIT$100K in 2016. The dataset contained characteristics of the entrants, free-text descriptions of the ventures, and crowd assessments of venture ideas.

Outcomes: 100K nomination and 2-year survival

The competition committee nominated 54 of the 177 teams on the basis of their value creation, value capture, and ability of leverage the funds. 49 companies remained in operation for two years after entry.

MIT $100K

Teams = 177



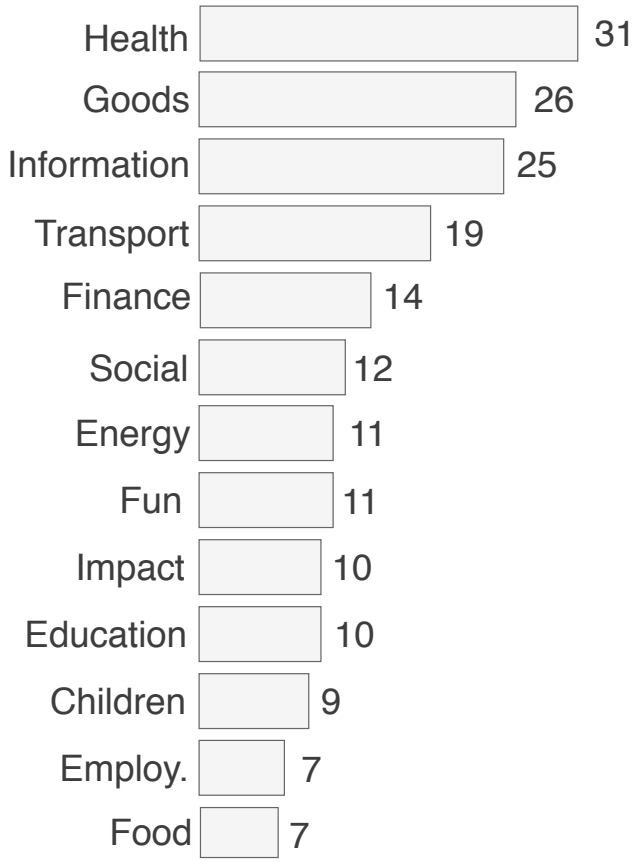
21/177 (12%) teams that were nominated *and* survived

21/54 (39%) teams that survived *given* nomination

28/123 (23%) teams that survived *given* no nomination

Data, Features and Outcomes

The 177 teams spanned many market sectors; health was the most common, food was the least



Exemplary idea

Clothing to detect gunshot wounds

Peer-to-peer market for personal storage

Platform representing legislative statistics

Parking spot rental and management service

Pre-paid debit card for international travel

Social network for the elderly

Gasoline delivery service

Crowd funded independent films

Search engine constrained by social values

Mobile fabrication lab for high school students

Toys to inspire scientific interest

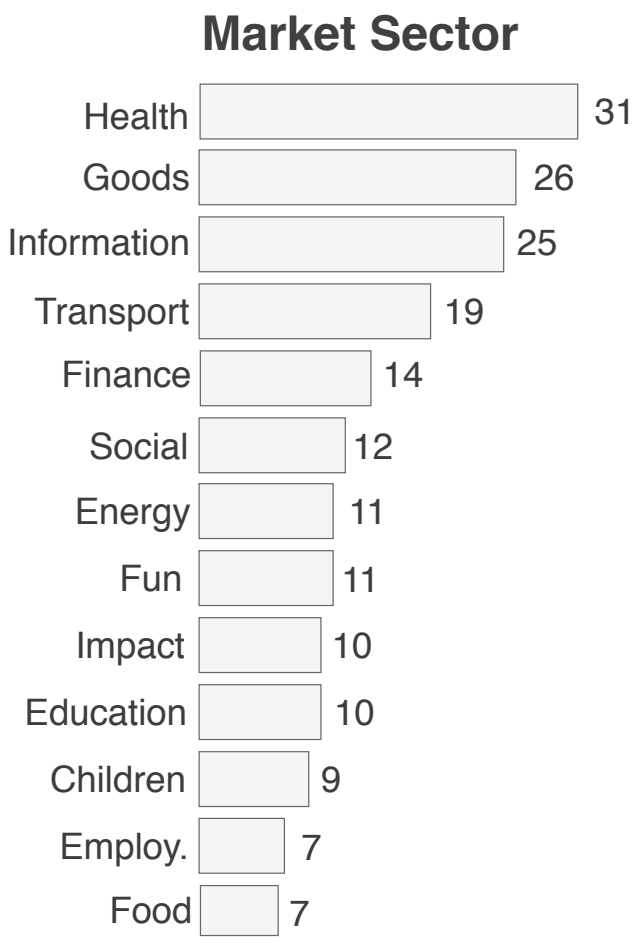
On-demand employment platform

Drones to monitor crop health

Can you guess the winner?

Data, Features and Outcomes

The 177 teams spanned many market sectors; health was the most common, food was the least



Exemplary idea

Clothing to detect gunshot wounds

Peer-to-peer market for personal storage

Platform representing legislative statistics

Parking spot rental and management service

Pre-paid debit card for international travel

Social network for the elderly

Gasoline delivery service

Crowd funded independent films

Search engine constrained by social values

Mobile fabrication lab for high school students

Toys to inspire scientific interest

On-demand employment platform

Drones to monitor crop health

Data, Features and Outcomes

16 features describe background of teams, properties of business abstracts, and crowd feelings

Team Features (n = 7)

Academic Institution	(1) M.I.T., (2) Harvard
Time Since Graduation	(3) Years
Academic Degree	(4) M.B.A., (5) Ph.D.
Academic Major	(6) S.T.E.M.
Non-Technical Skills	(7) Number

Venture Features (n = 7)

Target Market	(1) fundamental (2) periodic (3) inessential
Language Style	(4) posessive pronouns, (5) questions (6) adjectives, (7) sentiment

Crowd Features (n = 2)

Perceived Competence	(1) Consensus
Idea Rating	(2) Popular

Example Team Features

Dr. Linda Rosales, PhD
Education: *Postdoc, Computer Science, MIT*
Previous Jobs: *Engineering (12 months), Management (6 mo)*
Skills: *Engineering, Research, Algorithms, Comms.*
Honors/Awards: 4
Publications: 10 (2 prestigious)
Crowd Would Hire: ☐☐☐

Dr. Jennifer Baxley, MD
Education: *Graduate, Neurology, Harvard Medical School*
Previous Jobs: *Management (3 months), Medicine (12 mo.)*
Skills: *Medicine, Communications, Management*
Honors/Awards: 6
Publications: 2
Crowd Would Hire: ☐☐☐

Nicholás Faustus Aulet
Education: Undergraduate, *Criminology, NMSU*
Previous Jobs: *Military (24 mo.), Marketer (6 months)*
Skills: *Communications, Creative, Design*
Honors/Awards: 1
Publications: 0
Crowd Would Hire: ☐☐☐

Data, Features and Outcomes

16 features describe background of teams, properties of business abstracts, and crowd feelings

Team Features (n = 7)

Academic Institution	(1) M.I.T., (2) Harvard
Time Since Graduation	(3) Years
Academic Degree	(4) M.B.A., (5) Ph.D.
Academic Major	(6) S.T.E.M.
Non-Technical Skills	(7) Number

Venture Features (n = 7)

Target Market	(1) fundamental (2) periodic (3) inessential
Language Style	(4) posessive pronouns, (5) questions (6) adjectives, (7) sentiment

Crowd Features (n = 2)

Perceived Competence	(1) Consensus
Idea Rating	(2) Popular

Example Venture Features

Category Health

Rating

Abstract "*Foodborn Illness costs an estimated $20 billion in US healthcare costs annually. FoodStrip LLC has created a cheap, disposable strip that allows you to test your food or drink for illness causing bacteria in 30 seconds. Dip our patented strip into food or drink before eating, and learn if it's safe to consume.*"

Natural Language Processing

Bag-of-words: 1 1 1 1 1 2

Sentiment: 1 0 1 1 0 (-- - + ++)

Named Entity: 1 1 0 ($ Location Time)

POS Tag: 7 6 3 (Noun Verb Period)

Methods and Results

Compared many models on unseen data; logistic regression had the best performance

Tried many modeling approaches

We compared the classification performance of: 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.

Compared models using LOOCV AUC

We compared models to each-other using AUC across test subjects during leave one-outcross validation (LOOCV).

Selected logistic regression

Logistic regression with our 16 features had the highest AUC of the tested approaches. It also had the highest TPR (12%) when FPR was held at 0%.

Model ModelForm	**AUC**	**TPR at FPR**		
		10%	**5%**	**0%**
Logistic Regression	0.72	31%	20%	12%
SVM				
Linear	0.71	18%	12%	4%
Cubic	0.61	14%	6%	2%
Qudratic	0.62	12%	6%	4%
Medium Gaussian	0.67	24%	12%	0%
Coarse Gaussian	0.67	10%	6%	0%
Discriminant Analysis				
Linear	0.68	31%	14%	2%
Quadratic	0.71	33%	8%	0%
Ensemble				
RUSBoosted Trees	0.65	33%	27%	0%
Boosted Trees	0.65	35%	22%	0%
Bagged Trees	0.62	31%	20%	2%
Neural Networks				
1 Layer, 4 nodes	0.69	29%	10%	0%
2 Layers, 5x4 nodes	0.66	27%	16%	0%

* A subset of the tested models are shown here. For the full list, see here.

Methods and Results

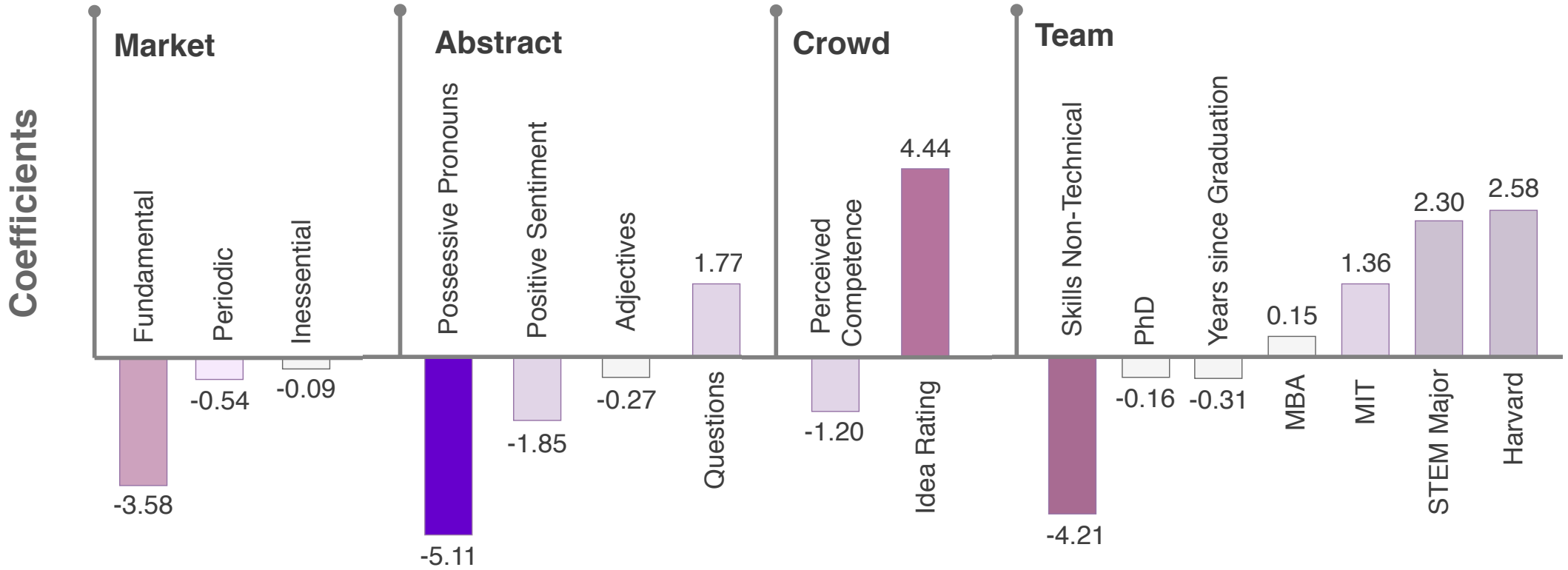
Model coefficients reveal that various factors work together to predict outcomes

Ventures that targeted fundamental markets (finance and energy) were less likely to survive.

Venture free-text abstracts that used possessive pronouns ('you', 'your', etc.) were less likely to survive.

Ideas that the crowd consistently preferred were 80 times as likely to survive than ideas the crowd did not consistently prefer.

Teams composed entirely of STEM major(s) were nearly 10 times as likely to survive as teams composed entirely of non-STEM majors.



Methods and Results

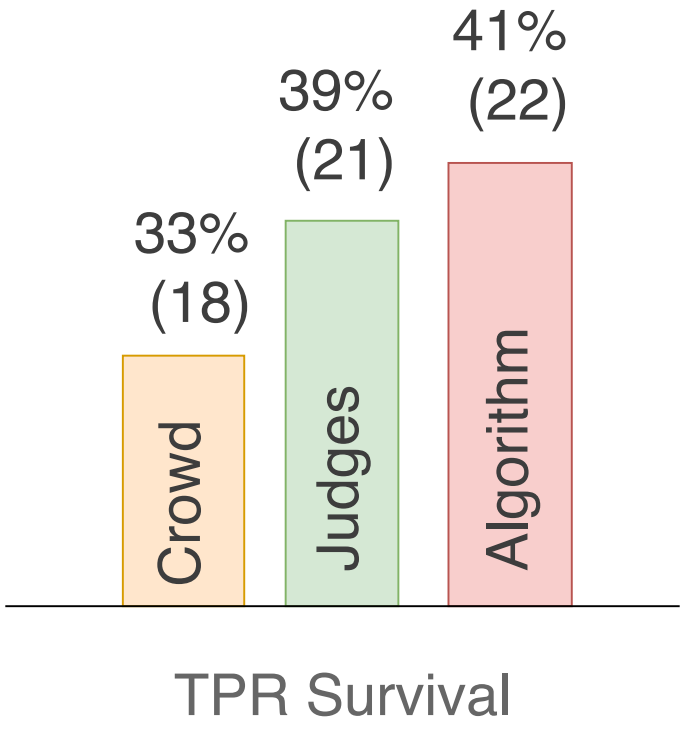
Our Algorithm out-performed the judges by 2%

Crowd detects survival from single sentence

The crowd workers only had access to a single sentence summary of the venture's idea, yet they still performed at 5% above the baseline survival rate of 28%.

Our Algorithm out-performed judges by 2%

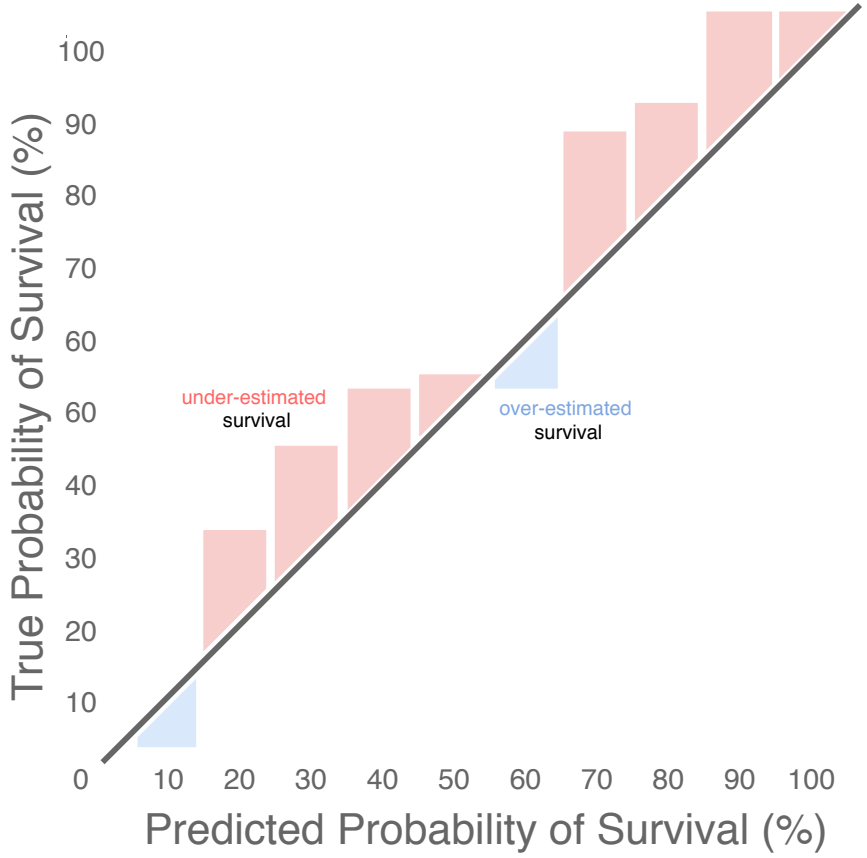
Our model outperformed the judges by 2%. This is remarkable given the informational discrepancy between our model and the judges.



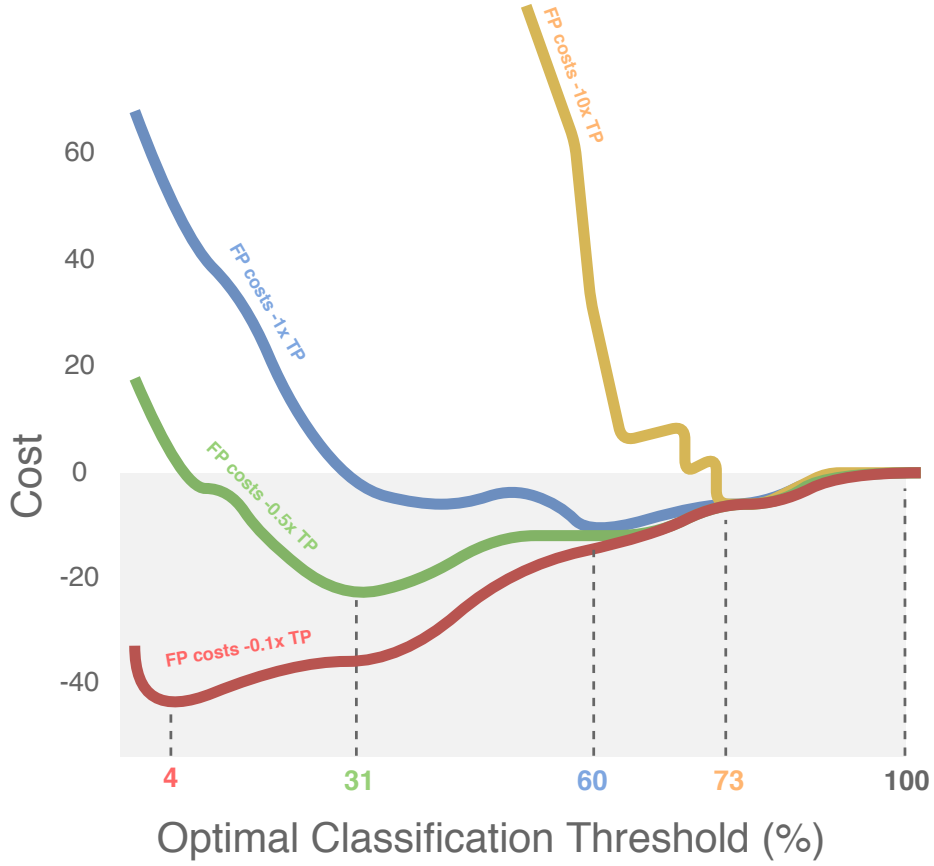
Methods and Results

The calibration of the model implies it's potential utility for risk assessment

Calibration implies risk-scoring capabilities



Varying risk tolerances accommodated



Conclusions

The initial conditions of teams can be used to predict their 2-year survival

1. Collect and curate a novel competition dataset

We collected a novel dataset of 177 ventures, comprising 374 individuals who entered the MIT$100K in 2016. The dataset contained characteristics of the entrants, free-text descriptions of the ventures, and crowd assessments of venture ideas.

2. Predict future success, given initial conditions

We extracted 16 features that described the ventures, the teams, and the crowd's assessment of the teams. We compared the ability of several models to predict the 2-year survival of teams against the competition judges.

3. Balance model interpretability & performance

The best performing approach was logistic regression, which has the added advantage of being highly interpretable. Ventures operating in non-saturated markets, that were liked by the crowd, had highly technical teams, and detailed ideas were most likely to survive.

4. Release code and data to make extensions easy

We made all the code and data available so others can recreate our model, figures and tables. This will make it easier for others to build on and improve this work.

https://github.com/ghamut/automated-venture-capitalist

ghamut

For more information

Paper Code Poster Contact