The Automated Venture Capitalist

Methods and Data to Predict the Fate of Startup Ventures

M.M. Ghassemi
Michigan State University

C. Song
Johns Hopkins University

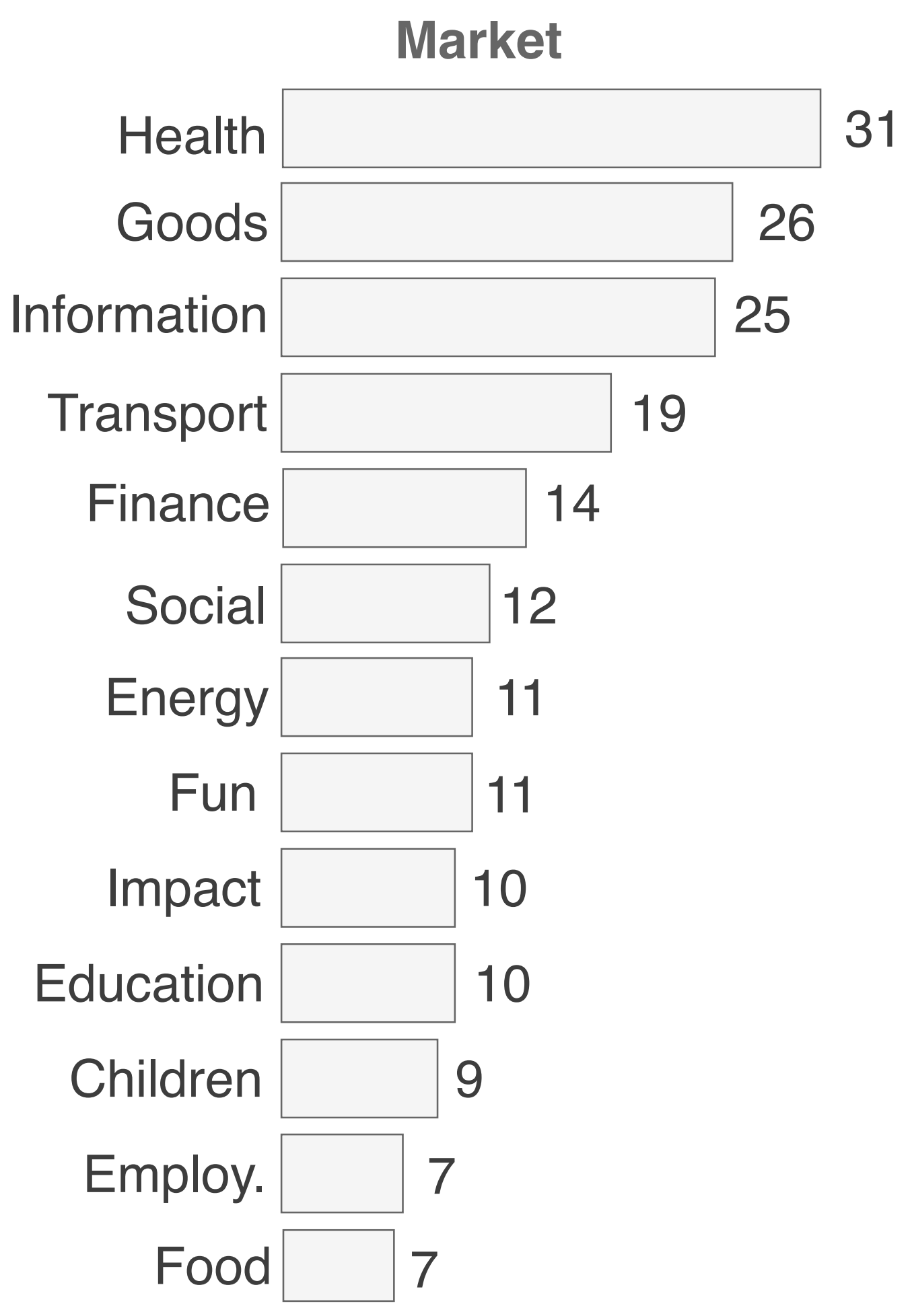
T. Alhanai
New York University AD

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.

Research Question: Do the initial conditions of startup teams and their ideas predict their 2-year survival?

Data

Venture Features

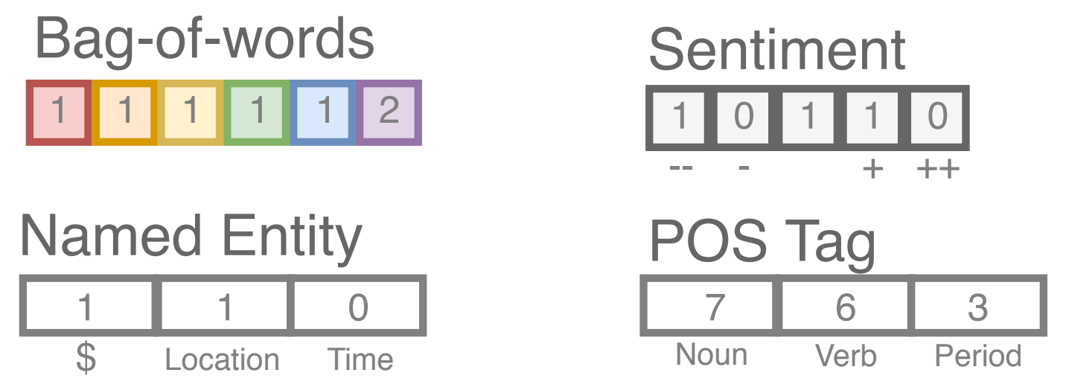


Category Health

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

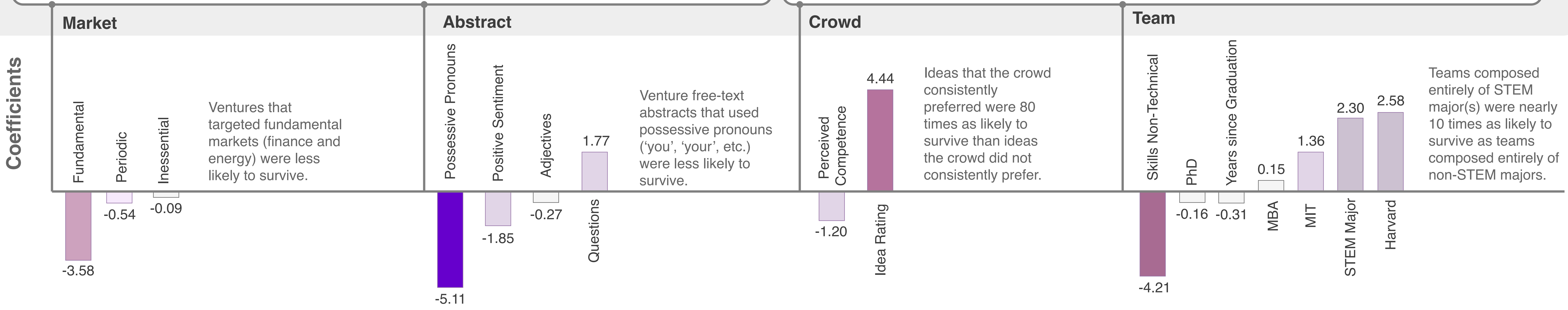


Team Features

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: ▨▨▨



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

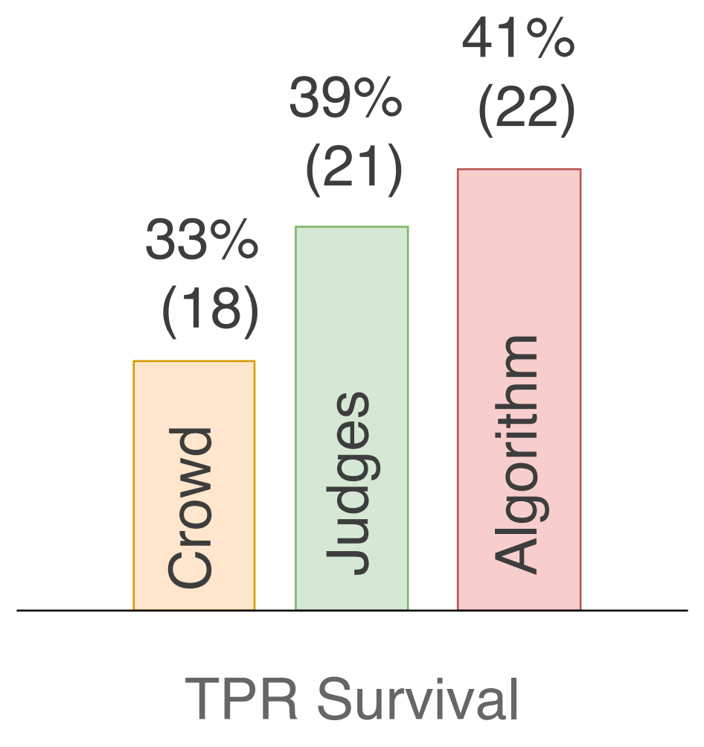
Models: We compared the classification performance of: Decision Trees, Discriminant Analysis, Logistic Regression, Support Vector Machines, k-Nearest Neighbors (k-NN), Ensemble Learning, and Neural Networks. All Neural Networks were feed-forward, and topology optimized using grid search. The best performing approach was Logistic Regression.

Performance Metrics and Validation: All models in this study were assessed using leave one-outcross validation (LOOCV). The classification performances of all models were measured using the Area Under the Receiver Operator Characteristic Curve (AUC). We compared against judges using True Positive Rates.

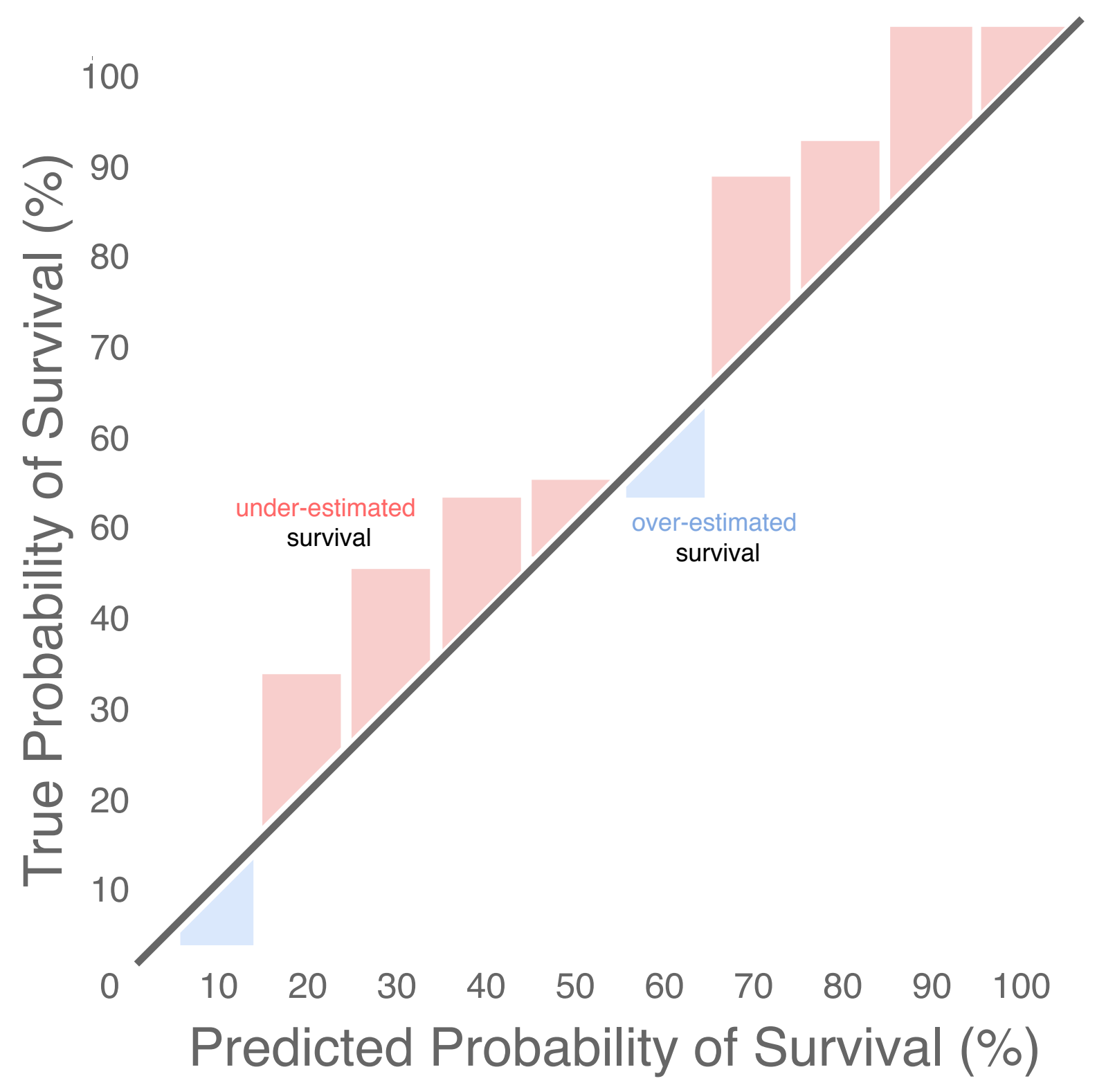
Proposed Cost Matrix: We identified a positive prediction rate that minimized the overall model cost across several different penalties, where the cost of a false positives was -0.1x, -0.5x, -1x, and -10x the cost of a true positive. We assumed the cost of a true negative to be 0, and the cost of a false negative to be 0.

Results

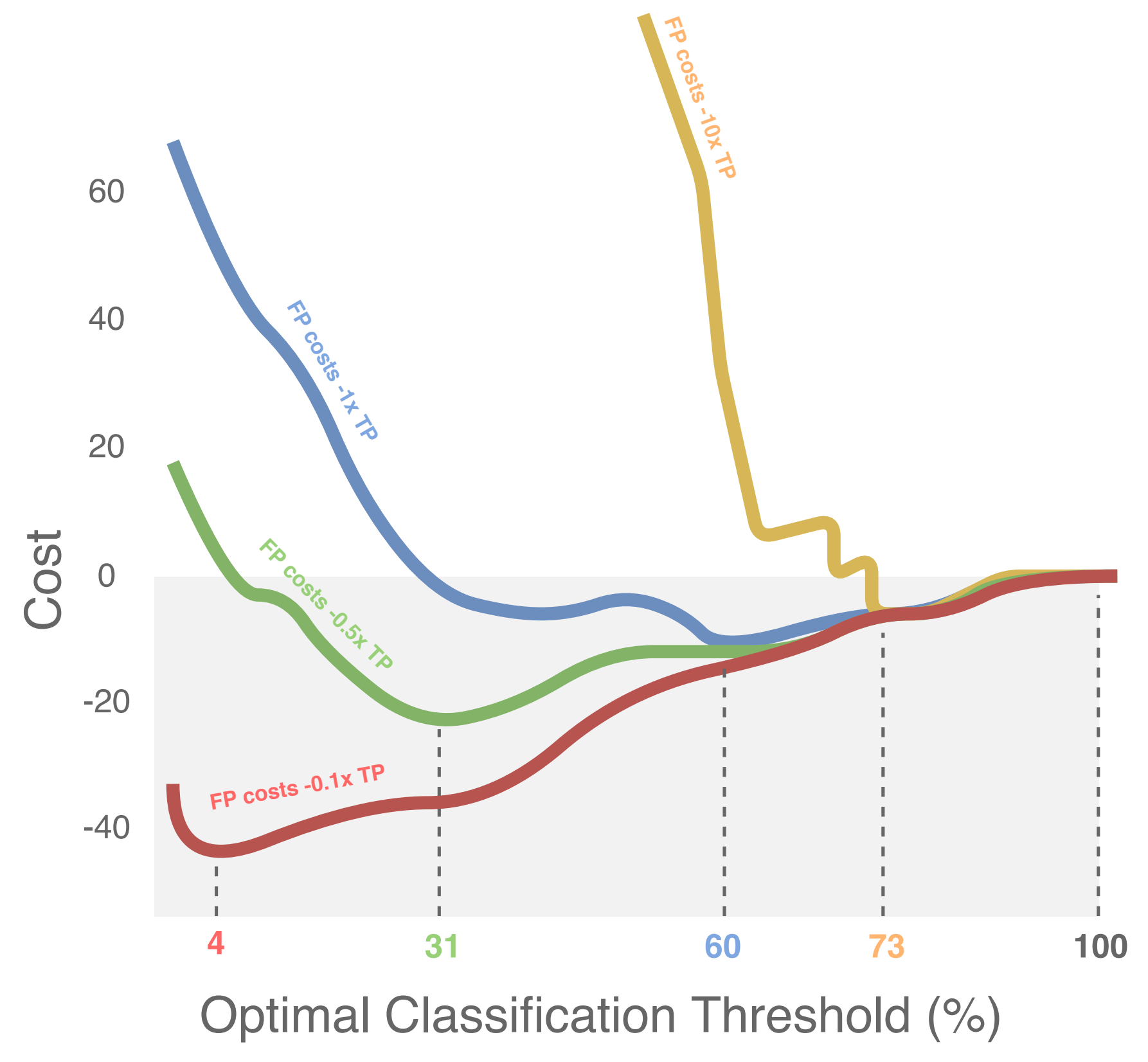
The best performing model exceeded the performance of the competition judges in predicting venture survival.



True Positive Rates: Predictive performance of judges, crowd popularity and model (algorithm). Baseline incidence of survival was 28%. The Algorithm identified the largest number of winning teams. Our Algorithm's TPR was 12% at an FPR of 0%.



Model Calibration: Calibration plot of the 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 = $p > 0.05$).



Expected Costs/Gains: 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 classification threshold that maximizes ROI.

ghamut

For questions, or collaboration interest, please email: research@ghamut.com
Data and code to replicate this study are available online: https://github.com/ghamut