

Machine Learning about Venture Capital Choices

Victor Lyonnet
University of Michigan

Léa H. Stern
University of Washington

IESE Banking Initiative Workshop
March 20, 2025

Motivation

- Venture capital plays a crucial role in **innovation**, economic **growth**, and **job creation**

Motivation

- Venture capital plays a crucial role in **innovation**, economic **growth**, and **job creation**
- The VC industry:
 - VCs rate **firm selection** as their **most important task** (Gompers et al., 2020)
 - Success hinges on identifying companies with the **highest potential**. Hard!
 - No historical data on new firms, limited “hard” information
 - **Swing for the fences** environment, extremely **skewed** outcomes

Motivation

- Venture capital plays a crucial role in **innovation**, economic **growth**, and **job creation**
- The VC industry:
 - VCs rate **firm selection** as their **most important task** (Gompers et al., 2020)
 - Success hinges on identifying companies with the **highest potential**. Hard!
 - No historical data on new firms, limited “hard” information
 - **Swing for the fences** environment, extremely **skewed** outcomes
- VCs report relying on “**gut feeling**” and **pattern matching**

Motivation

- Venture capital plays a crucial role in **innovation**, economic **growth**, and **job creation**
- The VC industry:
 - VCs rate **firm selection** as their **most important task** (Gompers et al., 2020)
 - Success hinges on identifying companies with the **highest potential**. Hard!
 - No historical data on new firms, limited “hard” information
 - **Swing for the fences** environment, extremely **skewed** outcomes
- VCs report relying on “**gut feeling**” and **pattern matching**
- **Question**: Can we leverage machine learning to **understand** VC decision-making?

This Paper: Understand VCs' Choice Behavior using ML

- ① **Which firms have the highest chance of success?** Can a model predict which startups will perform best?

This Paper: Understand VCs' Choice Behavior using ML

- ① **Which firms have the highest chance of success?** Can a model predict which startups will perform best?
- ② **Do VCs back highest potential firms?** Would this model select different firms than VCs?

This Paper: Understand VCs' Choice Behavior using ML

- ① **Which firms have the highest chance of success?** Can a model predict which startups will perform best?
- ② **Do VCs back highest potential firms?** Would this model select different firms than VCs?
- ③ **VCs' choice behavior:** Do VCs' choices reveal patterns of systematic errors? Why?

Outline

- 1 Algorithmic Design and Performance
- 2 Differences Between VC-backed and Best Predicted Performers
- 3 How do VCs Make Decisions?
- 4 Conclusion

Data Sources

Entrepreneur Survey

French
administrative data

4 cohorts every 4 years:
1998-2010

representative sample:
1/3 of all
entrepreneurs

survey questions cover:
demographics
expertise
experience
motivation
expectations
VC-backing

Financial Statements

French
administrative data

Corporate tax filings:
all new firms

Exits

Commercial data

Capital IQ, CB Insights, Crunchbase, Orbis, Pitchbook, Preqin, SDC, VentureXpert:
M&As + IPOs

Pitchbook:
valuation and revenue
at exit (in a different
sample)

Burgiss:
deal-level return and
characteristics (in a
different sample)

Data Sources

Entrepreneur Survey

French
administrative data

4 cohorts every 4 years:
1998-2010

representative sample:
1/3 of all
entrepreneurs

survey questions cover:
demographics
expertise
experience
motivation
expectations
VC-backing

Financial Statements

French
administrative data

Corporate tax filings:
all new firms

Exits

Commercial data

Capital IQ, CB In-
sights, Crunchbase, Or-
bis, Pitchbook, Preqin,
SDC, VentureXpert:
M&As + IPOs

Pitchbook:
valuation and revenue
at exit (in a different
sample)

Burgiss:
deal-level return and
characteristics (in a
different sample)

Key Advantages

No survivorship bias or selection bias → observe VCs' choice set ✓
Observe performance regardless of VC-backed status ✓

Algorithm Design

Features

$t = 0$

Entrepreneur and firm characteristics X_i that would be easily available to VCs in first-pass evaluation

48 questions \rightarrow 140 encoded characteristics

Ignore financing variables (e.g., VC-backing)

Outcome

$t = 1$

$y_i =$ Outcome

- Revenue at age 5
- 1[top 5% revenue]
- Exits (M&A, IPO)
- ...

Algorithm Design

Features

$t = 0$

Entrepreneur and firm characteristics X_i that would be easily available to VCs in first-pass evaluation

48 questions \rightarrow 140 encoded characteristics

Ignore financing variables (e.g., VC-backing)

Outcome

$t = 1$

$y_i = \text{Outcome}$

- Revenue at age 5
- 1[top 5% revenue]
- Exits (M&A, IPO)
- ...

- $n \approx 124\text{k}$ new firms in 4 cohorts (1998, 2002, 2006, 2010)
 \rightarrow drop firms in industries that never receive VC

Algorithm Design

Features

$t = 0$

Entrepreneur and firm characteristics X_i that would be easily available to VCs in first-pass evaluation

48 questions \rightarrow 140 encoded characteristics

Ignore financing variables (e.g., VC-backing)

Outcome

$t = 1$

$y_i =$ Outcome

- Revenue at age 5
- 1[top 5% revenue]
- Exits (M&A, IPO)
- ...

- $n \approx 124k$ new firms in 4 cohorts (**1998, 2002, 2006**, 2010)
 \rightarrow drop firms in industries that never receive VC
- **Train** XGBoost model on first 3 cohorts (69% of observations)

Algorithm Design

Features

$t = 0$

Entrepreneur and firm characteristics X_i that would be easily available to VCs in first-pass evaluation

48 questions \rightarrow 140 encoded characteristics

Ignore financing variables (e.g., VC-backing)

Outcome

$t = 1$

$y_i = \text{Outcome}$

- Revenue at age 5
- 1[top 5% revenue]
- Exits (M&A, IPO)
- ...

- $n \approx 124\text{k}$ new firms in 4 cohorts (1998, 2002, 2006, 2010)
 \rightarrow drop firms in industries that never receive VC
- Train *XGBoost* model on first 3 cohorts (69% of observations) \rightarrow
Predict outcome: $\hat{m}(x_i) \rightarrow M(x_i)$ **percentile rank**

Algorithm Design

Features

$t = 0$

Entrepreneur and firm characteristics X_i that would be easily available to VCs in first-pass evaluation

48 questions \rightarrow 140 encoded characteristics

Ignore financing variables (e.g., VC-backing)

Outcome

$t = 1$

$y_i =$ Outcome

- Revenue at age 5
- 1[top 5% revenue]
- Exits (M&A, IPO)
- ...

- $n \approx 124k$ new firms in 4 cohorts (1998, 2002, 2006, **2010**)
 \rightarrow drop firms in industries that never receive VC
- Train *XGBoost* model on first 3 cohorts (69% of observations) \rightarrow Predict outcome: $\hat{m}(x_i) \rightarrow M(x_i)$ percentile rank
- Results **evaluated** in 2010 cohort (31% of observations), never seen by algorithm

Predictive Accuracy

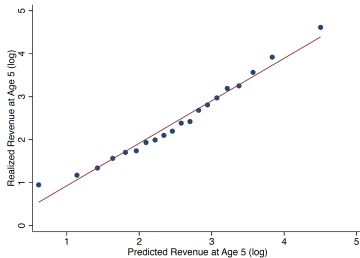


Figure 1: All companies in test set

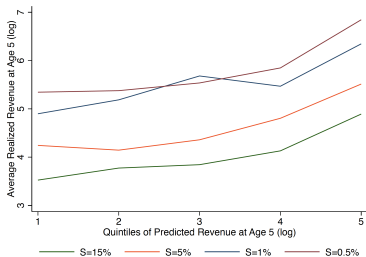
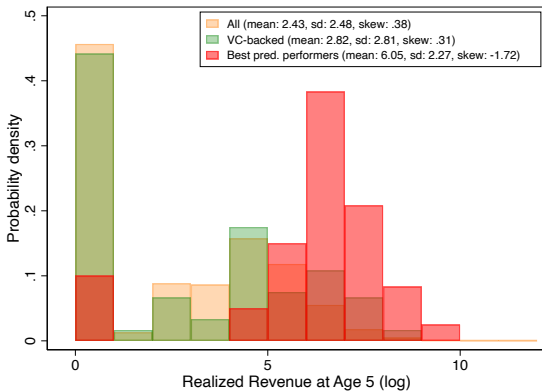
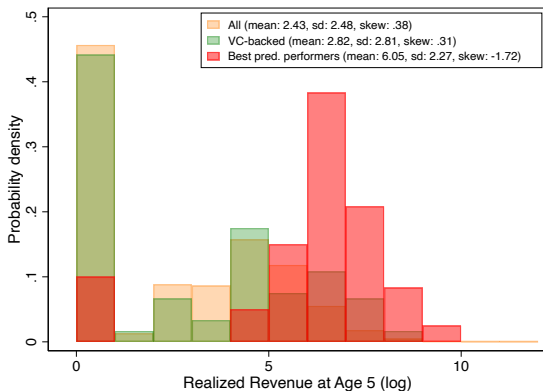


Figure 2: Right tail of predicted performance distribution

Realized Performance: All firms, VC-backed, Best Predicted Performers



Realized Performance: All firms, VC-backed, Best Predicted Performers



- The **best predicted performers** outperform **VC-backed** companies (regardless of the outcome measure).

Sensitivity Analysis: Deal Terms

- Using revenue as performance measure → implicit assumption of constant deal terms
- What if the best predicted performers were “too expensive”?

Sensitivity Analysis: Deal Terms

- Using revenue as performance measure → implicit assumption of constant deal terms
- What if the best predicted performers were “too expensive”?
- $MOIC_{\alpha} - MOIC_h$

$$MOIC_i = \frac{\text{post valuation} * \% \text{acquired} * \text{dilution}}{\text{invested capital}}$$

Sensitivity Analysis: Deal Terms

- Using revenue as performance measure → implicit assumption of constant deal terms
- What if the best predicted performers were “too expensive”?
- $MOIC_{\alpha} - MOIC_h$

$$MOIC_i = \frac{\text{post valuation} * \% \text{acquired} * \text{dilution}}{\text{invested capital}}$$

$$MOIC_i = \frac{\text{revenue}_5 * \text{industry multiple} * \% \text{acquired} * 0.75}{\text{invested capital}}$$

Sensitivity Analysis: Deal Terms

- Using revenue as performance measure → implicit assumption of constant deal terms
- What if the best predicted performers were “too expensive”?
- $MOIC_{\alpha} - MOIC_h$

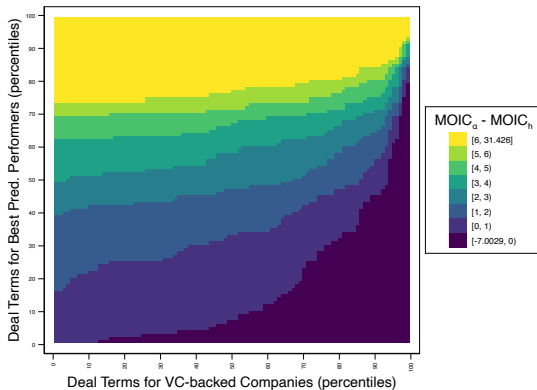
$$MOIC_i = \frac{\text{post valuation} * \% \text{acquired} * \text{dilution}}{\text{invested capital}}$$

$$MOIC_i = \frac{\text{revenue}_5 * \text{industry multiple} * \% \text{acquired} * 0.75}{\text{invested capital}}$$

- Distribution of deal terms in Pitchbook

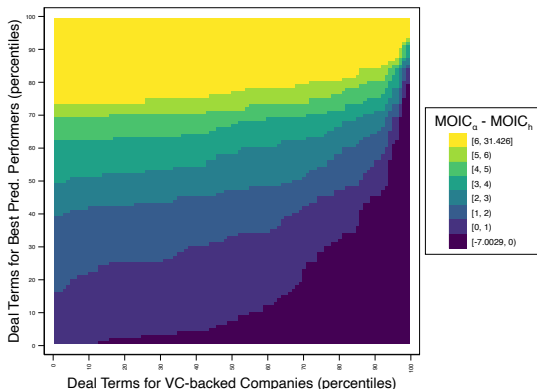
Sensitivity Analysis: Deal Terms

$$MOIC_i = \frac{\text{revenue}_5 * \text{industry multiple} * \% \text{acquired} * 0.75}{\text{invested capital}}$$



Sensitivity Analysis: Deal Terms

$$MOIC_i = \frac{\text{revenue}_5 * \text{industry multiple} * \% \text{acquired} * 0.75}{\text{invested capital}}$$



- Assuming VCs received median terms for their portfolio firms, they would have had to get terms below the 8th pctile on portfolio of best predicted performers for difference to be negative

Sensitivity: Varying Investable Pool \mathcal{D}

Supply and Demand Considerations

→ *Revealed preference approach*

Sensitivity: Varying Investable Pool \mathcal{D}

Supply and Demand Considerations

→ *Revealed preference approach*

1. We restrict the analysis to firms in **industries** that receive VC in our data

Sensitivity: Varying Investable Pool \mathcal{D}

Supply and Demand Considerations

→ *Revealed preference approach*

1. We restrict the analysis to firms in **industries** that receive VC in our data
2. We **adjust** \mathcal{D} to only include founders whose survey responses match those of VC-backed founders on key dimensions

Investable Pool (\mathcal{D})	Revenue at Age 5 (log)	
	Mean	S.D.
Unconstrained	6.05	2.27
Location	5.64	2.3
Industry	5.25	2.88
Growth, innovation and hiring	5.38	2.83
Financially constrained	5.13	2.69
Location and industry	3.9	2.88
Loc., ind., and fin. constrained	3.35	2.96
Growth, fin. cons. and industry	3.91	3
Same revenue at birth	4.89	2.73
Comparison:	Revenue at Age 5 (log)	
	Mean	S.D.
All firms in test set	2.43	2.48
VC-backed firms	2.82	2.81

Sensitivity: Varying Investable Pool \mathcal{D}

Supply and Demand Considerations

→ *Revealed preference approach*

1. We restrict the analysis to firms in **industries** that receive VC in our data
2. We **adjust** \mathcal{D} to only include founders whose survey responses match those of VC-backed founders on key dimensions

▶ ind-loc graph

▶ dd graph

Cost Quantification

- VCs invest in some companies that perform **predictably poorly**

Cost Quantification

- VCs invest in some companies that perform **predictably poorly**
 - **Dropping bottom half** of portfolio companies in terms of predicted performance $\rightarrow \approx$ **50%** increase in imputed portfolio MOIC

Cost Quantification

- VCs invest in some companies that perform **predictably poorly**
 - **Dropping bottom half** of portfolio companies in terms of predicted performance $\rightarrow \approx$ **50%** increase in imputed portfolio MOIC
- VCs fail to invest in companies that perform **predictably well**

Cost Quantification

- VCs invest in some companies that perform **predictably poorly**
 - **Dropping bottom half** of portfolio companies in terms of predicted performance → \approx **50%** increase in imputed portfolio MOIC
- VCs fail to invest in companies that perform **predictably well**
 - Investing exclusively in top 1% of **best predicted performers** → \approx **200%** increase in imputed portfolio MOIC
 - Constrained to same industry + location → \approx 50% increase in imputed portfolio MOIC

Cost Quantification

- VCs invest in some companies that perform **predictably poorly**
 - **Dropping bottom half** of portfolio companies in terms of predicted performance $\rightarrow \approx 50\%$ increase in imputed portfolio MOIC
- VCs fail to invest in companies that perform **predictably well**
 - Investing exclusively in top 1% of **best predicted performers** $\rightarrow \approx 200\%$ increase in imputed portfolio MOIC
 - Constrained to same industry + location $\rightarrow \approx 50\%$ increase in imputed portfolio MOIC

Next

Do VC-backed companies differ from the best predicted performers?

Outline

- 1 Algorithmic Design and Performance
- 2 Differences Between VC-backed and Best Predicted Performers
- 3 How do VCs Make Decisions?
- 4 Conclusion

Differences in Selected Entrepreneurs

Compared to VCs' selections, the model selects:

- **fewer very young** entrepreneurs
- **more female** entrepreneurs
- **fewer** graduates from **elite** schools
- **fewer** firms located in **Paris**

Differences in Selected Entrepreneurs

Compared to VCs' selections, the model selects:

- **fewer very young** entrepreneurs
- **more female** entrepreneurs
- **fewer** graduates from **elite** schools
- **fewer** firms located in **Paris**

→ same when predicting other outcomes

Differences in Selected Entrepreneurs

- Does this mean that VCs are biased against certain entrepreneurs?

Not necessarily!

These are equilibrium outcomes of VCs' decisions based on predictions of firm performance

Differences in Selected Entrepreneurs

- Does this mean that VCs are biased against certain entrepreneurs?

Not necessarily!

These are equilibrium outcomes of VCs' decisions based on predictions of firm performance

- We need to evaluate the quality of VCs' ex-ante predictions

Next

How do VCs form predictions?

Outline

- 1 Algorithmic Design and Performance
- 2 Differences Between VC-backed and Best Predicted Performers
- 3 How do VCs Make Decisions?**
- 4 Conclusion

Can We Predict VCs' Decisions?

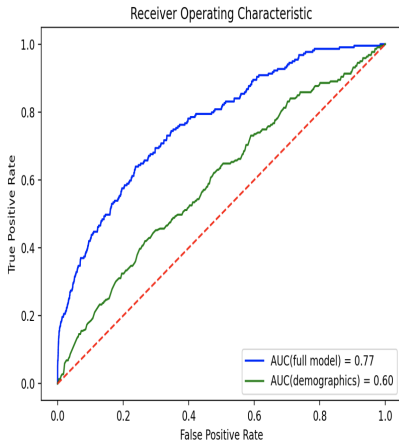


Figure 3: AUC of predictive model of VCs' decisions

- Create a model to predict VCs' backing decision: $\hat{h}(X)$
- Pick two random ventures one *VC-backed* one *not VC-backed*
- Odds that model assigns a higher probability of being VC-backed to the one that indeed is VC-backed is 77%
- Much of the signal is captured by three demographic features: age, education, gender

Can We Predict VCs' Decisions?

	(1)	(2)	(3)	VC-backed		(6)	(7)	(8)
				(4)	(5)			
$\hat{h}(X)$	1.5*** (.043)				1.6*** (.044)	1.5*** (.044)	1.6*** (.044)	1.6*** (.045)
$\hat{m}(X)_{top5_Revenues}$.058*** (.0066)			-.0059 (.0067)			-.0067 (.0087)
$\hat{m}(X)_{exit}$.49*** (.055)			.042 (.056)		.077 (.062)
$\hat{m}(X)_{Log_Revenues}$.0038*** (.00053)			-.0006 (.00053)	-.00047 (.00064)
adj. R^2	.047	.0029	.0029	.0018	.047	.047	.047	.047
Observations	26,440	26,440	26,440	26,440	26,440	26,440	26,440	26,440

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

⇒ Predictability in VC behavior even when accounting for performance predictions!

Do VCs Exaggerate some Features?

Building on Mullainathan and Obermeyer (*QJE*, 2022)

Do VCs Exaggerate some Features?

Building on Mullainathan and Obermeyer (*QJE*, 2022)

- 1 We create new **simple** predictive models of firm performance with the same prediction setup, but **restricting the set of input features** $\rightarrow \hat{m}_{simple}$

Do VCs Exaggerate some Features?

Building on Mullainathan and Obermeyer (*QJE*, 2022)

- 1 We create new **simple** predictive models of firm performance with the same prediction setup, but **restricting the set of input features** $\rightarrow \hat{m}_{simple}$
- 2 We regress VCs' decisions on our **full** model predicting exits + our **simple** models:

$$VC-backed_i = \beta_0 + \hat{m}_{full}(X_i)\beta_1 + \hat{m}_{simple}(X_i)\beta_2 + \epsilon_i \quad (1)$$

Do VCs Exaggerate some Features?

Building on Mullainathan and Obermeyer (*QJE*, 2022)

- 1 We create new **simple** predictive models of firm performance with the same prediction setup, but **restricting the set of input features** $\rightarrow \hat{m}_{simple}$
- 2 We regress VCs' decisions on our **full** model predicting exits + our **simple** models:

$$VC-backed_i = \beta_0 + \hat{m}_{full}(X_i)\beta_1 + \hat{m}_{simple}(X_i)\beta_2 + \epsilon_i \quad (1)$$

- 3 Controlling for predicted perf. that **accounts for all features and their interactions/non-linearities**, we **isolate** the role of individual characteristics on VCs' choices, **over and above** their effect on predicted performance
- 4 Simple models should **not** predict VCs' decisions over and above full model of firm performance!
 - $\beta_2 = 0$: the features in $\hat{m}_{simple}(\cdot)$ do not matter for VCs' decisions *over and above their effect on firm performance*

Do VCs Exaggerate some Features?

Building on Mullainathan and Obermeyer (QJE, 2022)

- 1 We create new **simple** predictive models of firm performance with the same prediction setup, but **restricting the set of input features** $\rightarrow \hat{m}_{simple}$
- 2 We regress VCs' decisions on our **full** model predicting exits + our **simple** models:

$$VC-backed_i = \beta_0 + \hat{m}_{full}(X_i)\beta_1 + \hat{m}_{simple}(X_i)\beta_2 + \epsilon_i \quad (1)$$

- 3 Controlling for predicted perf. that **accounts for all features and their interactions/non-linearities**, we **isolate** the role of individual characteristics on VCs' choices, **over and above** their effect on predicted performance
- 4 Simple models should **not** predict VCs' decisions over and above full model of firm performance!
 - $\beta_2 = 0$: the features in $\hat{m}_{simple}(\cdot)$ do not matter for VCs' decisions *over and above their effect on firm performance*
 - $\beta_2 > 0$: the features matter **too much** in VC decisions
 - $\beta_2 < 0$: the features matter **too little** in VC decisions

Do VCs Exaggerate some Features?

Panel A: Entrepreneurs' features

	VC-backed									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\hat{m}(X)(\text{Exit})$.72*** (.037)	.7*** (.038)	.72*** (.037)	.71*** (.037)	.7*** (.037)	.64*** (.034)	.72*** (.037)	.72*** (.037)	.73*** (.04)	.71*** (.037)
$\hat{m}_{simple}(\text{Personal Characteristics})$.25** (.11)								
$\hat{m}_{simple}(\text{Age})$.068 (.16)							
$\hat{m}_{simple}(\text{Gender})$.91*** (.33)						
$\hat{m}_{simple}(\text{Graduate Degree})$.56*** (.16)					
$\hat{m}_{simple}(\text{Elite School})$.85*** (.16)				
$\hat{m}_{simple}(\text{French Nationality})$.13 (1.2)			
$\hat{m}_{simple}(\text{Relatives})$								-.6 (.66)		
$\hat{m}_{simple}(\text{Optimism})$									-.038 (.1)	
$\hat{m}_{simple}(\text{Serial Entrepreneur})$.86*** (.32)
Adj. R^2	.01	.01	.01	.01	.01	.012	.01	.01	.01	.01
Observations	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353	37,353

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

- Relative to uncond. backing rate, founders with certain characteristics are more likely to be VC-backed **than justified by the impact of these characteristics on exit predictions, accounting for feature interactions and non-linearities:**
 - **male** founders: **1.4x**
 - **elite schools grads**: **3x**
 - **serial** founders: **2x**

Why do VCs Exaggerate some Features?

Representativeness of Success

- A feature is **representative** (or **stereotypical**) of success if it is *more frequent among the best* performing entrepreneurs relative to the other ones (Tversky and Kahneman, 1974; Bordalo et al., 2016)

Why do VCs Exaggerate some Features?

Representativeness of Success

- A feature is **representative** (or **stereotypical**) of success if it is *more frequent among the best performing entrepreneurs* relative to the other ones (Tversky and Kahneman, 1974; Bordalo et al., 2016)

Feature	Top 1%	Bottom 99%	Representativeness of best performers $\frac{Pr(X_i Top1)}{Pr(X_i Bottom99)}$
	(1)	(2)	(3)
Male	80.95	69.57	1.16
Graduate Degree	18.11	10.53	1.72
Elite School	3.91	1.76	2.22
Optimism	53.02	19.63	2.7
Serial Entrepreneur	13.02	3.68	3.53
Paris-based	16.92	10.14	1.67
High-tech Ind.	5.92	4.13	1.43

Why do VCs Exaggerate some Features?

Representativeness of Success

- A feature is **representative** (or **stereotypical**) of success if it is *more frequent among the best performing entrepreneurs* relative to the other ones (Tversky and Kahneman, 1974; Bordalo et al., 2016)

Feature	Top 1%	Bottom 99%	Representativeness of best performers $\frac{Pr(X_i \text{Top1})}{Pr(X_i \text{Bottom99})}$
	(1)	(2)	(3)
Male	80.95	69.57	1.16
Graduate Degree	18.11	10.53	1.72
Elite School	3.91	1.76	2.22
Optimism	53.02	19.63	2.7
Serial Entrepreneur	13.02	3.68	3.53
Paris-based	16.92	10.14	1.67
High-tech Ind.	5.92	4.13	1.43

→ VCs over-index on entrepreneurs with characteristics **stereotypical** of the best performing entrepreneurs (consistent with VCs swinging for the fences)

Why do VCs Exaggerate **Stereotypical** Features?

Bridging Belief Formation Literature and ML approach

- **Distortion in estimated odds of success Ψ** (Bordalo et al., 2023):

$$\Psi = \frac{\hat{\pi}_{Success_{\mathcal{F}}}}{\hat{\pi}_{NonSuccess_{\mathcal{F}}}} / \frac{\pi_{Success_{\mathcal{F}}}}{\pi_{NonSuccess_{\mathcal{F}}}}$$

where for a set of entrepreneurs with vector of features \mathcal{F} :

$\pi_{Success_{\mathcal{F}}}$ = success prob. in this set

$\hat{\pi}_{Success_{\mathcal{F}}}$ = estimated prob. of success for firms in this set: proxied with VC-backing rates in this set

Why do VCs Exaggerate **Stereotypical** Features?

Bridging Belief Formation Literature and ML approach

- **Distortion in estimated odds of success Ψ** (Bordalo et al., 2023):

$$\Psi = \frac{\hat{\pi}_{Success_{\mathcal{F}}}}{\hat{\pi}_{NonSuccess_{\mathcal{F}}}} / \frac{\pi_{Success_{\mathcal{F}}}}{\pi_{NonSuccess_{\mathcal{F}}}}$$

where for a set of entrepreneurs with vector of features \mathcal{F} :

$\pi_{Success_{\mathcal{F}}}$ = success prob. in this set

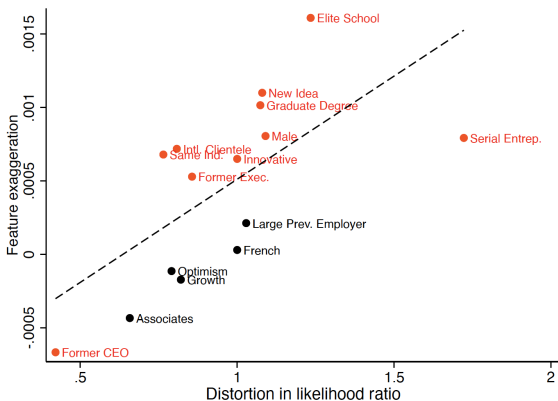
$\hat{\pi}_{Success_{\mathcal{F}}}$ = estimated prob. of success for firms in this set: proxied with VC-backing rates in this set

- For instance, for **elite school** entrepreneurs:

$$\Psi = \frac{Pr(VC-backed|elite school)}{Pr(Not VC-backed|elite school)} / \frac{Pr(Success|elite school)}{Pr(Not Success|elite school)}$$

Why do VCs Exaggerate Stereotypical Features?

Bridging Belief Formation Literature with ML approach



→ (ML-derived) feature **exaggeration** ($\hat{\beta}_2$) increases with the **distortion in estimated odds** (Ψ)

→ Consistent with VCs **mispredicting** success rates for some types of entrepreneurs due to beliefs driven by **representativeness heuristic**

Conclusion

- We use predictive methods to study how VCs make early-stage investment decisions
 - Requires observing VCs' **choice set** and rich data, including outcomes, for VC-backed and **non-VC-backed** firms
- Using an ex ante approach that **accounts for uncertainty** as of decision making, we **identify, quantify** and **explain** patterns of systematic errors
- Some of the patterns could emerge from VCs' more limited access to some firms, e.g. network effects. Still VCs spend considerable resources sourcing deals, and are incentivized to find most promising ventures.
- High uncertainty + swinging for the fences environment → **“kernel of truth” stereotypes** that overemphasize traits associated with success
- Results consistent with **misprediction** arising from **representativeness-driven distortion in estimated odds** of success
 - shed light on the **root cause** of documented investment patterns and the **narrowness** of the VC industry (Lerner and Nanda, 2020)

Appendix

Summary Statistics

Variable	Training					
	Mean	SD	p50	p90	p99	N
Outcomes						
Revenue at Age 5 (log), k euros	2.31	2.46	2.20	5.67	7.68	84,583
Revenue at Age 5, k euros	157.55	1,420.22	8.00	289.00	2,155.00	84,583
Alive at Age 5	0.62	0.48	1.00	1.00	1.00	84,583
Demographics						
Entrepreneur's Age	37.78	10.00	37.00	52.00	63.00	84,583
Female	0.28	0.45	0.00	1.00	1.00	84,583
Entrepreneur's Nationality (FR)	0.90	0.30	1.00	1.00	1.00	84,583
Entrepreneurial Family	0.69	0.46	1.00	1.00	1.00	84,583
Professional Background						
Self-employed	0.37	0.48	0.00	1.00	1.00	84,583
Previously Employed	0.51	0.50	1.00	1.00	1.00	84,583
Part-time Entrepreneur	0.18	0.39	0.00	1.00	1.00	84,583
Same Prior Industry	0.54	0.50	1.00	1.00	1.00	84,583
Serial Entrepreneur	0.04	0.19	0.00	0.00	1.00	84,583
Previously Employed in Small Firm	0.45	0.50	0.00	1.00	1.00	84,583
Previously Inactive	0.10	0.30	0.00	0.00	1.00	84,583
Below High School Degree	0.38	0.48	0.00	1.00	1.00	84,583
Undergraduate Degree	0.21	0.41	0.00	1.00	1.00	84,583
Graduate Degree	0.11	0.31	0.00	1.00	1.00	84,583
Grande Ecole	0.04	0.21	0.00	0.00	1.00	33,806
Completed Required Training	0.21	0.41	0.00	1.00	1.00	84,583
Motivation and Expectations						
Expectation: Growth	0.52	0.50	1.00	1.00	1.00	84,583
Expectation: Sustain	0.27	0.45	0.00	1.00	1.00	84,583
Expectation: Rebound	0.07	0.25	0.00	0.00	1.00	84,583
Motivation: Peer Entrepreneurs	0.11	0.31	0.00	1.00	1.00	84,583
Expect to Hire	0.24	0.43	0.00	1.00	1.00	84,583
Motivation: New Idea	0.18	0.38	0.00	1.00	1.00	84,583
Motivation: Opportunity	0.32	0.47	0.00	1.00	1.00	84,583
Innovation	0.34	0.47	0.00	1.00	1.00	84,583

Summary Statistics

Variable	Training					
	Mean	SD	p50	p90	p99	N
Venture Characteristics						
Paris-based	0.10	0.30	0.00	1.00	1.00	84,583
Marseille-based	0.02	0.14	0.00	0.00	1.00	84,583
Lyon-based	0.02	0.13	0.00	0.00	1.00	84,583
Bordeaux-based	0.02	0.14	0.00	0.00	1.00	84,583
Business Services Industry	0.16	0.36	0.00	1.00	1.00	85,119
Health and Social Work Industry	0.04	0.20	0.00	0.00	1.00	85,119
Construction Industry	0.18	0.39	0.00	1.00	1.00	85,119
High tech Industry	0.01	0.12	0.00	0.00	1.00	85,119
Energy Industry	0.00	0.02	0.00	0.00	0.00	85,119
B2B	0.33	0.47	0.00	1.00	1.00	84,583
B2C	0.63	0.48	1.00	1.00	1.00	84,583
International Customers	0.07	0.25	0.00	0.00	1.00	84,583
Local Customers	0.53	0.50	1.00	1.00	1.00	84,583
Domestic Customers	0.14	0.35	0.00	1.00	1.00	84,583
Venture Organization						
Co-founders	0.12	0.32	0.00	1.00	1.00	84,583
Outsourcing: Accounting	0.64	0.48	1.00	1.00	1.00	84,583
Number of Employees	1.59	1.52	1.00	3.00	8.00	84,583
10+ Clients	0.63	0.48	1.00	1.00	1.00	84,583
Number of Paid Managers	0.15	0.46	0.00	1.00	2.00	84,583
Customers from Prior Job	0.30	0.46	0.00	1.00	1.00	84,583
Suppliers from Prior Job	0.23	0.42	0.00	1.00	1.00	84,583
Help from Professionals	0.03	0.17	0.00	0.00	1.00	84,583
Help from Family	0.27	0.44	0.00	1.00	1.00	84,583
No External Help	0.44	0.50	0.00	1.00	1.00	84,583

Selective Labels and VC Treatment Effect

Selective labels

- We cannot observe VCs' counterfactual investment returns on non-VC backed companies: $[r_i | h_i = 0]$

Selective Labels and VC Treatment Effect

Selective labels

- We cannot observe VCs' counterfactual investment returns on non-VC backed companies: $[r_i | h_i = 0]$
- VCs' returns are a function of their portfolio companies' performance
- Admin data: performance available for **all** firms

Selective Labels and VC Treatment Effect

Selective labels

- We cannot observe VCs' counterfactual investment returns on non-VC backed companies: $[r_i | h_i = 0]$
- VCs' returns are a function of their portfolio companies' performance
- Admin data: performance available for **all** firms
- *quasi-labels* $[y_i | h_i = 0]$ (account for z_i)
→ effect of unobservables limited to VC treatment effect

Selective Labels and VC Treatment Effect

Selective labels

- We cannot observe VCs' counterfactual investment returns on non-VC backed companies: $[r_i|h_i = 0]$
- VCs' returns are a function of their portfolio companies' performance
- Admin data: performance available for **all** firms
- *quasi-labels* $[y_i|h_i = 0]$ (account for z_i)
→ effect of unobservables limited to VC treatment effect
- **Assumption on VC treatment effect:**

$$\underbrace{(y_i|h_i = 1) - (y_i|h_i = 0)}_{\text{treatment effect}} > - \underbrace{((y_i|h_i = 0) - (y_j|h_j = 1))}_{\text{performance gap(>0)}}. \quad (2)$$

→ VC treatment effect for firms selected by the alternative policy cannot be more negative than the observed (positive) performance gap

Why Use Machine Learning?

- **No unified theory** on which entrepreneurs are more promising ex-ante
 - some covariates likely matter in some cases but not others, and may interact in nonlinear ways
 - no obvious choice of regression model to make out-of-sample predictions of new ventures' performance
- **Machine learning** methods
 - rely on a rigorous data-driven model selection that maximizes out-of-sample predictive accuracy
 - are well adapted for structured data

Omitted Payoffs Analysis

Algorithm trained on	Algorithm evaluated on							
	Revenue ₅ (log)	Revenue ₇ (log)	Top 5% Revenue ₅	Top 5% Revenue ₇	Imputed Valuation (log)	Revenue Growth	Exits	
Revenue ₅ (log)	6.05	5.58	0.60	0.53	1.48	0.15	4	
Revenue ₇ (log)	5.62	5.19	0.48	0.44	1.29	0.13	3	
Top 5% Revenue ₅	5.50	4.94	0.57	0.52	1.36	0.11	3	
Top 5% Revenue ₇	5.53	5.15	0.58	0.55	1.42	0.11	3	
Imputed Valuation (log)	4.16	3.97	0.23	0.19	0.83	0.13	3	
Revenue Growth	2.73	2.71	0.00	0.00	0.30	0.12	0	
Exit (IPO/M&A)	4.09	3.50	0.32	0.28	0.87	0.10	10	
Comparison: Average performance measures								
	Revenue ₅ (log)	Revenue ₇ (log)	Top 5% Revenue ₅	Top 5% Revenue ₇	Imputed Valuation (log)	Revenue Growth	Exits	
All firms in test set	2.43	2.02	0.05	0.05	0.28	-0.05	118	
VC-backed firms	2.82	2.46	0.15	0.16	0.45	-0.01	10	

▶ [back](#)

Choice of Predicted Performance Measure

- Using Pitchbook data (France + U.S.): we show a strong correspondence between **exit revenue** and **exit value**

Step 1. For each sector:

- take the **top performers** in terms of **exit revenue**
- calculate their percentile rank in terms of **exit value**
- calculate the mean and median rank of **top performers**

Step 2. Take the average across sectors

	Percentile rank in exit value distribution (averaged across sectors)	
	Average venture	Median venture
Top 1% revenue	90th	96th
Top 5% revenue	88th	95th
Top 10% revenue	84th	91th

Accounting for Deal Terms using Imputed MOIC

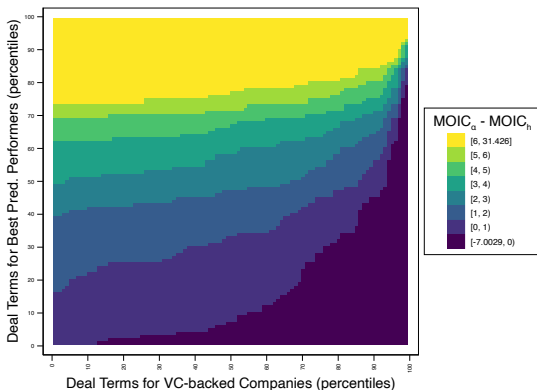
$$MOIC_i = \frac{\delta_i * M_s * y_i}{k_i}$$

- M_s : median revenue multiple at exit for each sector
- δ_i : dilution (75%)
- k_i : initial capital
- y_i : realized revenue

▶ Back

Accounting for Deal Terms using Imputed MOIC

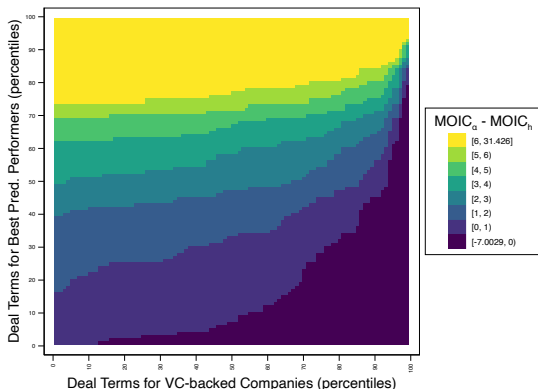
$$MOIC_i = \frac{\delta_i * M_s * y_i}{k_i}$$



▶ Back

Accounting for Deal Terms using Imputed MOIC

$$MOIC_i = \frac{\delta_i * M_s * y_i}{k_i}$$



VCs would have had to get terms below the 8th pctile for companies in \mathcal{A}_s to do worse than VC-backed companies

► Back

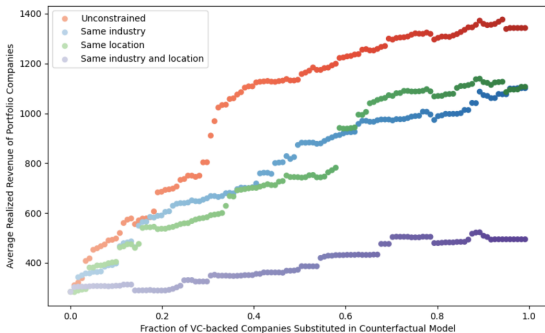
Cost of VCs' Errors

Panel A: Cost of Picking Bad Pred. Performers					
Dropping VC-backed w/ pred. perf:	# Port. Companies	Multiple (survivors only)	Multiple	% Increase	
	(1)	(2)	(3)	(4)	
bottom 10%	108	1.33	.76	9.2	
bottom 25%	90	1.32	.79	13.3	
bottom 50%	60	1.55	.98	39.9	

Panel B: Cost of Passing On Best Pred. Performers					
	# Port. Companies	Multiple (survivors only)	Multiple	% Increase	
	(1)	(2)	(3)	(4)	
top 1%	373	2.52	2.05	193	
top 0.5%	186	2.8	2.38	240	
top 120	120	2.99	2.54	262.8	

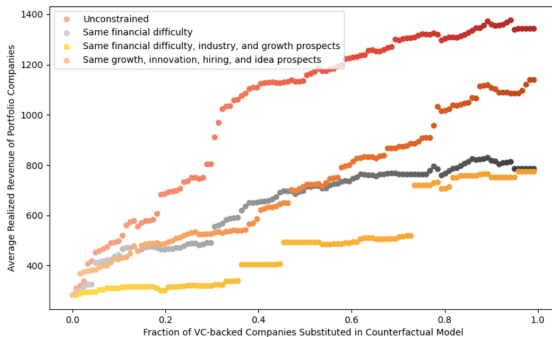
▶ Back

Counterfactuals Accounting for VCs' Industry and Location Preferences



▶ Back

Counterfactuals Accounting for Demand Side



▶ Back

References I

Gompers, Paul A, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev. 2020.
“How do venture capitalists make decisions?” *Journal of Financial Economics*,
135(1): 169–190.