Rationalizing Entrepreneurs' Forecasts

Nicholas Bloom (Stanford)
Mihai Codreanu (Stanford)
Robert Fletcher (Stanford & Instacart)

July 18th, 2022
NBER Summer Institute, Entrepreneurship

Overview

- Firm expectations–key determinant of investment and production decisions and key input into the design of fiscal and monetary policy
- **RQs: How accurately can entrepreneurs forecast their sales? Can we improve their forecasts? What interventions work?**
- Collect detailed panel revenue forecast data on 7,463 US firms. Cross-check with Stripe.com administrative data.
- Entrepreneurs were paid for accuracy - $25 for forecasting next quarter's sales within 10% of realizations
- We experimented with:
  - Increasing the forecast accuracy reward up to $400
  - Providing them with dashboard information on their current sales
  - Training them on how to use simple forecasting heuristics

⏩ Related literature

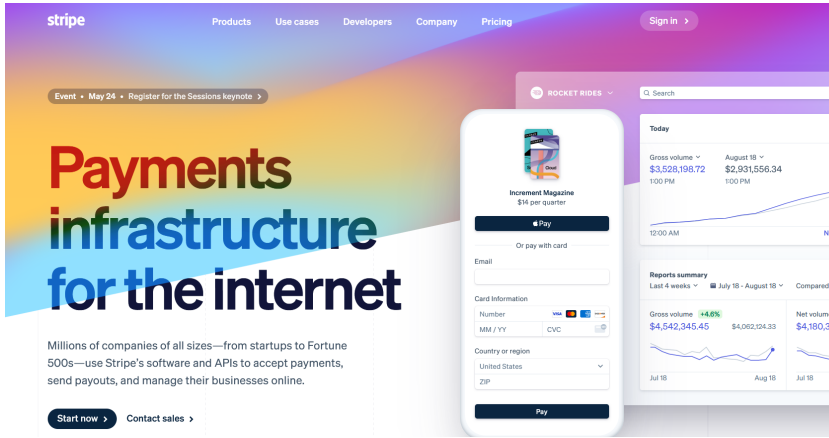
Key results

- **State of forecasting:**
  - Baseline (pre-Covid): only 13% of firms can forecast their sales in the next quarter within 10% of their realised values
  - Random walk benchmark: ≈15% correct
  - Non-systematic errors (noise) trumps over systematic errors (bias): 92.7% of MSE, rest over-confidence
- **Biases and Low Adoption of Forecasting Tools**
  - Widespread over-precision and over-confidence in ability
  - Dunning-Kruger effect on relative forecasting ability (↑ confident $\implies$ ↓ forecaster)
- **RCT Evidence on Forecasting Interventions**
  - Attention: Monetary incentives reduce bias
  - Data: Reviewing historical data reduces noise (& ↓ MSE)
  - Skill: Forecasting training has small effect

Plan for the rest of the talk

Worked with Stripe, leading U.S. payment processing firm

- Fintech with valuation of about $100bn, with 100,000s of firms around the world
- Mostly small firms but some very large firms (note all data presented today is anonymized & winsorized)



Survey Sample

- Panel of 7,463 users, $50 for first survey and $25 per follow-up
- Response Rate: 23% pre-COVID, 17% in all
- $25 for forecasting next quarter's sales within 10% of realizations

Table 1: Survey Rounds Overview

Round	Dates	Responses	Extra Module	Interventions
1	Jan-Apr 2019	3,941	Baseline Characteristics	
2	May-Aug 2019	2,891	Management	
3	Oct 2019-Jan 2020	3,185	Personality	
4	Apr-May 2020	2,446	COVID-19 Part I	
5	Sep-Oct 2020	2,409	COVID-19 Part II	Dashboard + Reward
6	Jan-Apr 2021	1,883	Management	Dashboard
7	Sep-Nov 2021	3,100	Forecasting	Dashboard + Reward + Forecast Training
8	Apr-Aug 2022	1,938	Forecasting Importance	Forecast Training II
9	Sep-Dec 2022	TBC	TBC	TBC
10	Jan-Mar 2023	TBC	End of survey	

⏩ Industries ⏩ Summary stats ⏩ % Rev Stripe ⏩ Comparison ⏩ Geography ⏩ Propensity To Respond

Baseline Survey: Forecast for the next 3 & 12 months

$25 AWARD FOR ACCURACY

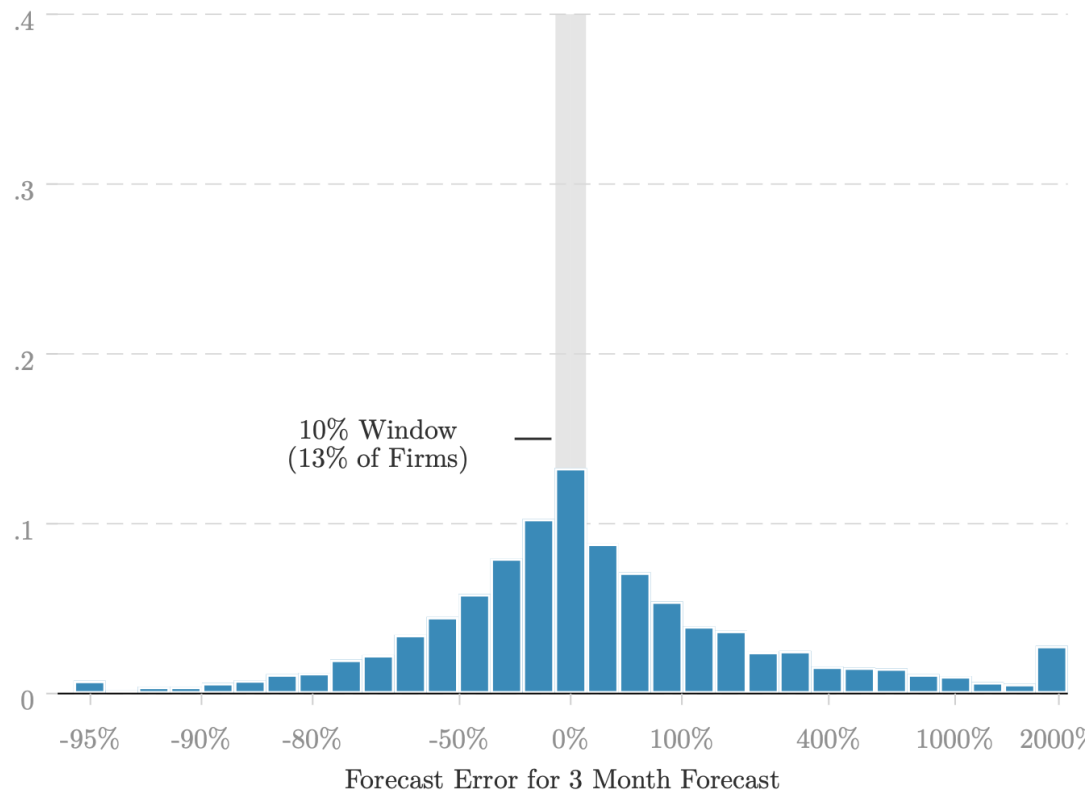
We would like you to make a 3-month prediction for April through June 2021. If your prediction is within 10% of your actual Stripe revenue in 3 months, we'll send you an additional **$25** Amazon gift card.

What do you predict your revenue **on Stripe** will be in April through June 2021?

$[ ].00

- Forecasting Competition for the 3-months
- Limited to Stripe revenue only (check directly rather than report)
- Discussions with managers suggests platform of revenue matters (potentially big differences in fees, payout schedules etc.)
- Results robust to firms with higher and lower % of revenues on Stripe

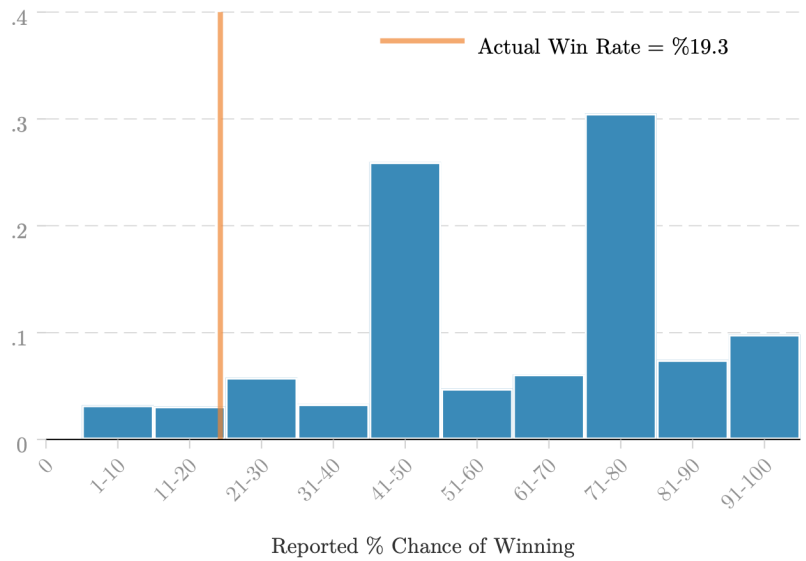
Firms forecast next quarter sales poorly

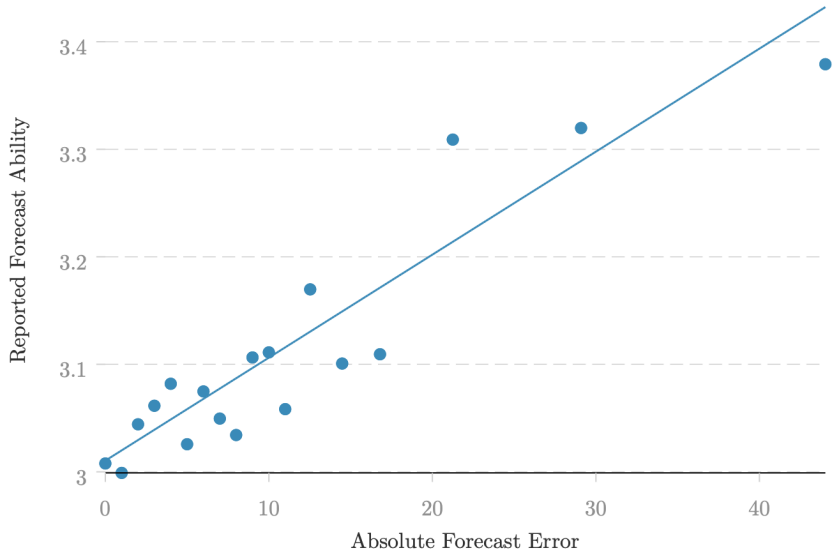


Note: Forecasting error is calculated as log(forecast next quarter sales) – log(realization of next quarter sales). Results for rounds 1-3, 5300 firms. All firms were paid $25 for quarterly sales forecasts within 10% of their actual numbers

⏩ Forecasting Accuracy Heterogeneity ⏩ Forecasting Accuracy 12-Month

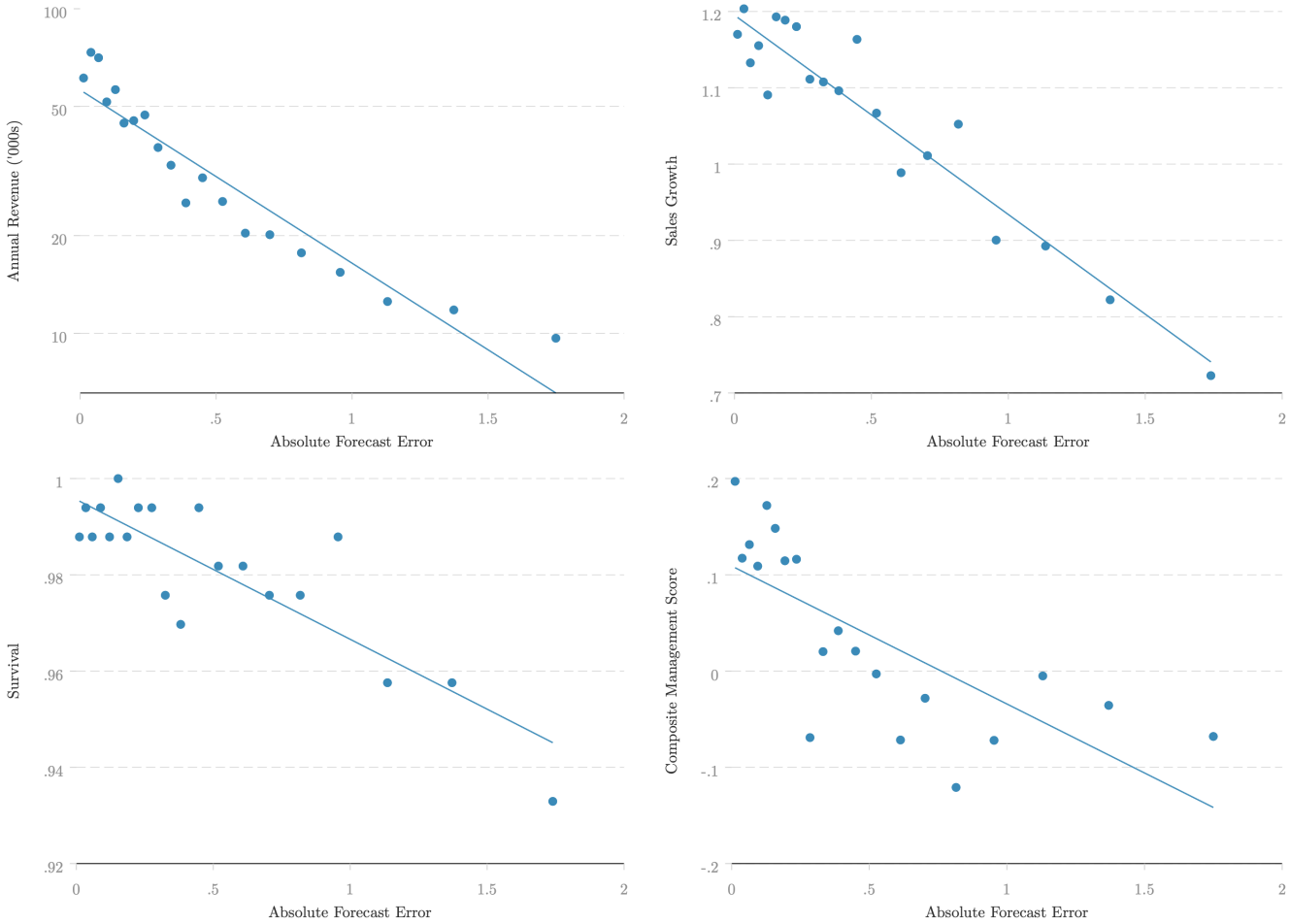
Overconfidence and better forecasters think they are worse (1% significant, t-stat 7.24)





Note: Self-reported abilities were collected in round 7 using a 5-point likert scale with "far below average" as 1 and "far above average" as 5. The absolute error is calculated using the absolute difference between their response and the suggested response in the training module. Reported probabilities of winning were collected in round 8. Win rate reflects the probability to win for the first 495 firms in the sample in round 8.

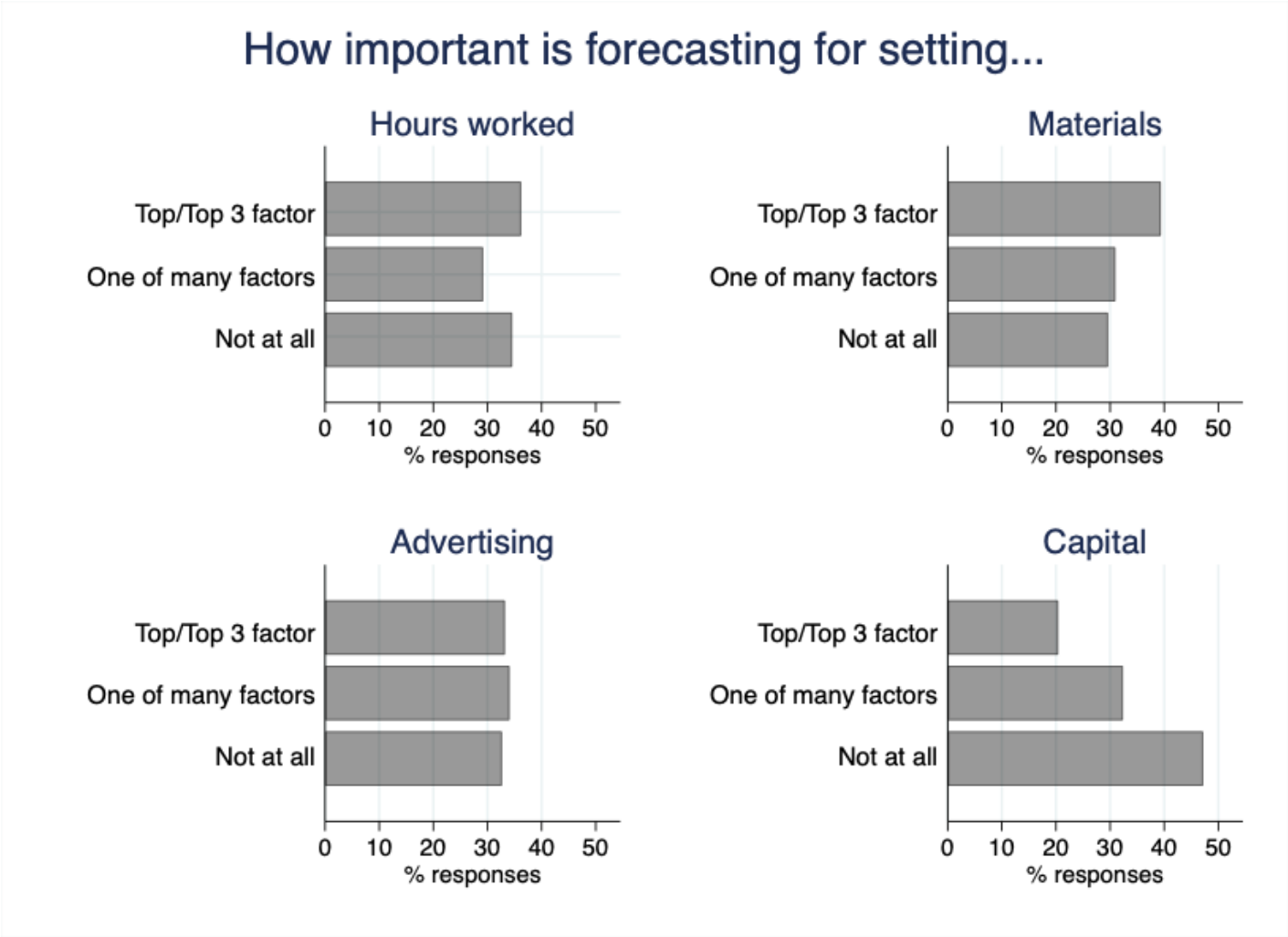
Forecast errors are negatively correlated with performance



Note: Annual revenue, quarterly growth, semi-annual survival, and 12 month forecast error are historical data from rounds 1 through 7, 6,659 firms.

⏩ Training Vs. Performance

And forecasting seems to be important for them...



Note: Self-reported importance of 3-months sales forecasting for the first 1,640 firms participating in wave 8. ⏩ Forecasting Importance Intensity Changes

Do they not pay enough attention when completing the forecast exercise?

- Randomized forecast reward from \$0 (right hand side box) to \$400 (bottom box below)
- Increments of \$25 between \$0-\$50, \$50 above that (between \$50-\$400)

Stanford stripe

What do you predict your revenue **on Stripe** will be in Quarter 4, 2020 (October, November, and December) combined?

\$ _____ .00

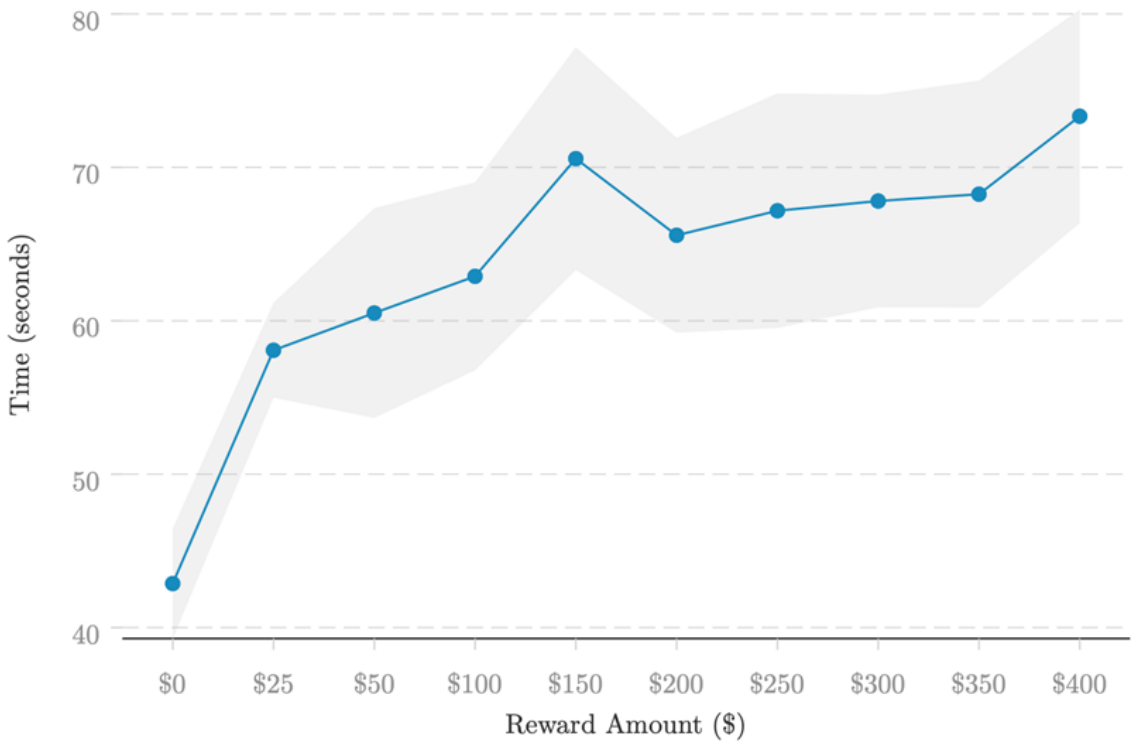
\$400 AWARD FOR ACCURACY

We would like you to make a 3-month prediction for Quarter 4, 2020 (October, November, and December). If your prediction is within 10% of your actual Stripe revenue in 3 months, we'll send you an additional **\$400** Amazon gift card.

What do you predict your revenue **on Stripe** will be in Quarter 4, 2020 (October, November, and December) combined?

\$ _____ .00

Higher rewards led entrepreneurs spend more time on forecasts (Significant at 1%, t-stat of 7.41)



Notes: Time to answer the forecasting question. Times are winsorized at 180 seconds. Sample of 3,177 firms from round 5 and 7.

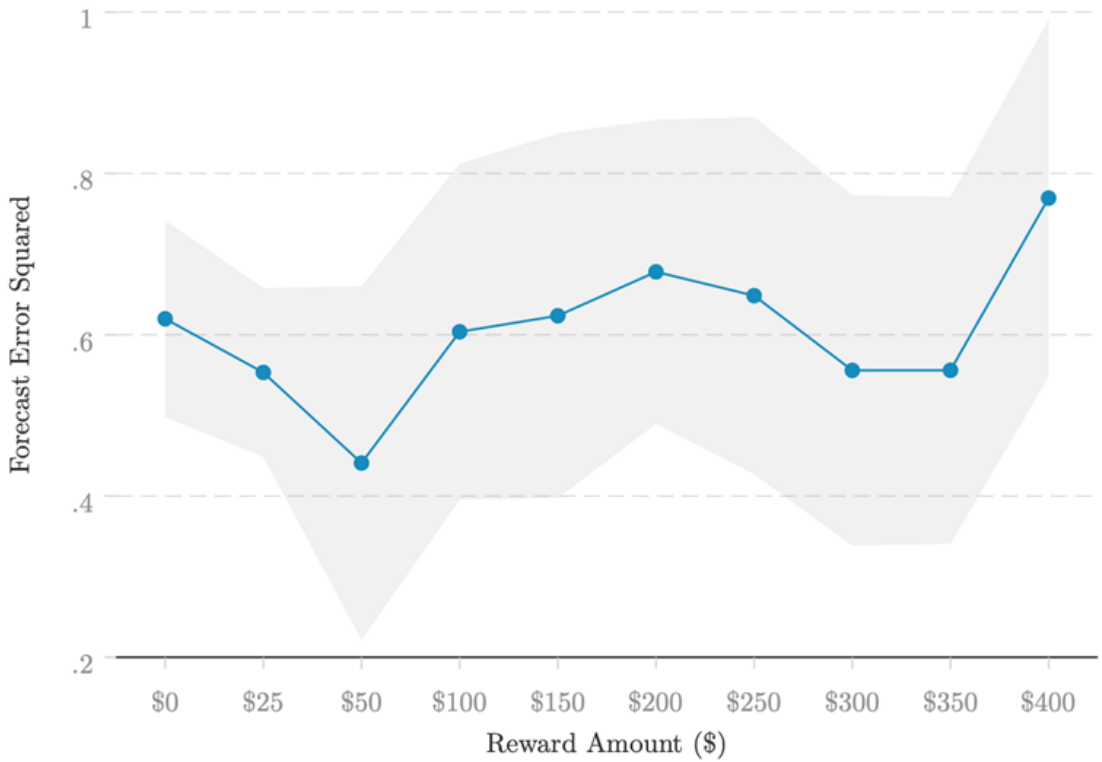
⏩ Timing Regression ⏩ Timing Regression Other

Some impact of reward on forecast error – bias reduced

	Report Err.	(Report Err.)^2	Forecast Err.	(Forecast Err.)^2
Reward '00s	-0.008	0.002	-0.034**	-0.032
	(0.010)	(0.022)	(0.014)	(0.026)
Time FEs	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes
Dep. Mean	0.030	0.414	0.127	0.768
Observations	6659	6659	6659	6659

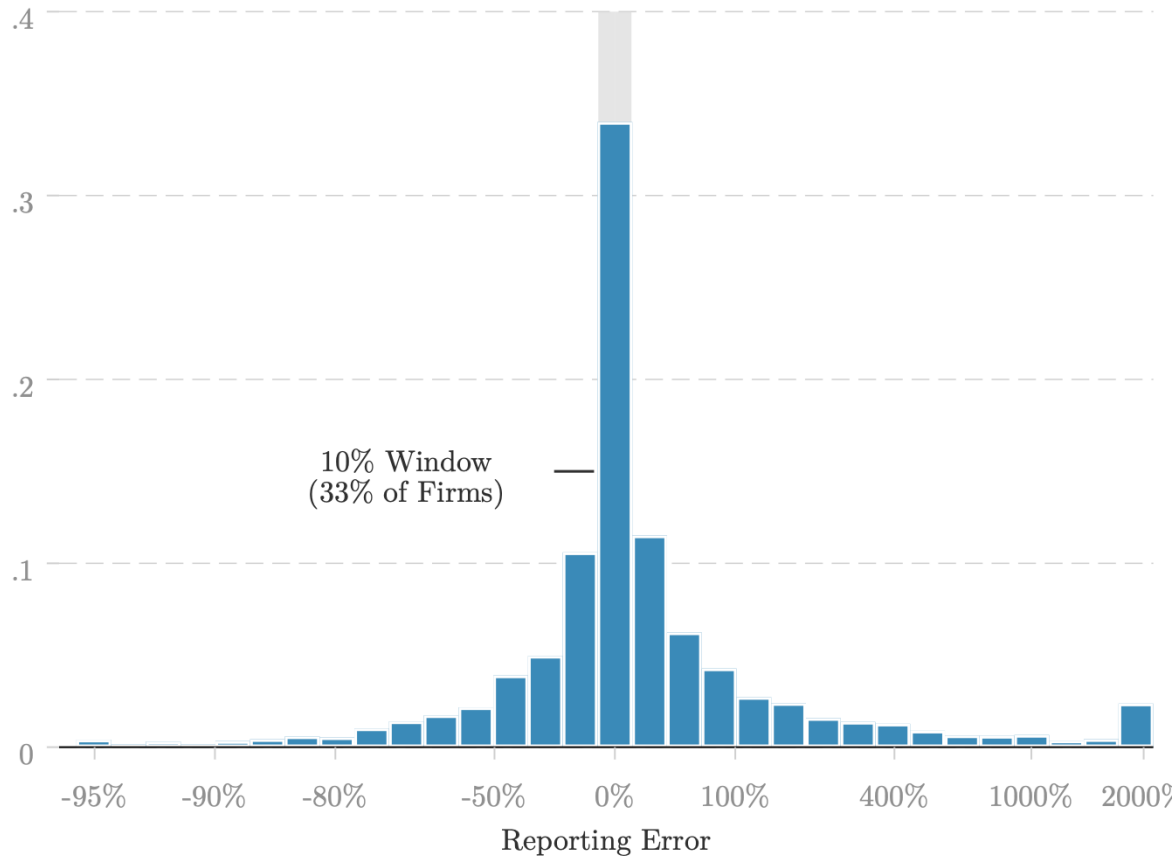
Note: Regression of e.g. log(forecast next quarter sales) – log(realization of next quarter sales) on the reward payment for forecasts within 10% of actual. Data from rounds 1 through 7, with standard errors clustered at the firm level.

But the noise... still there



Notes: Binscatter of squared value of log of (forecast next quarter sales) – (realization of next quarter sales) on the reward payment for forecasts within 10% of actual. Sample of 3,177 firms from round 5 and 7.

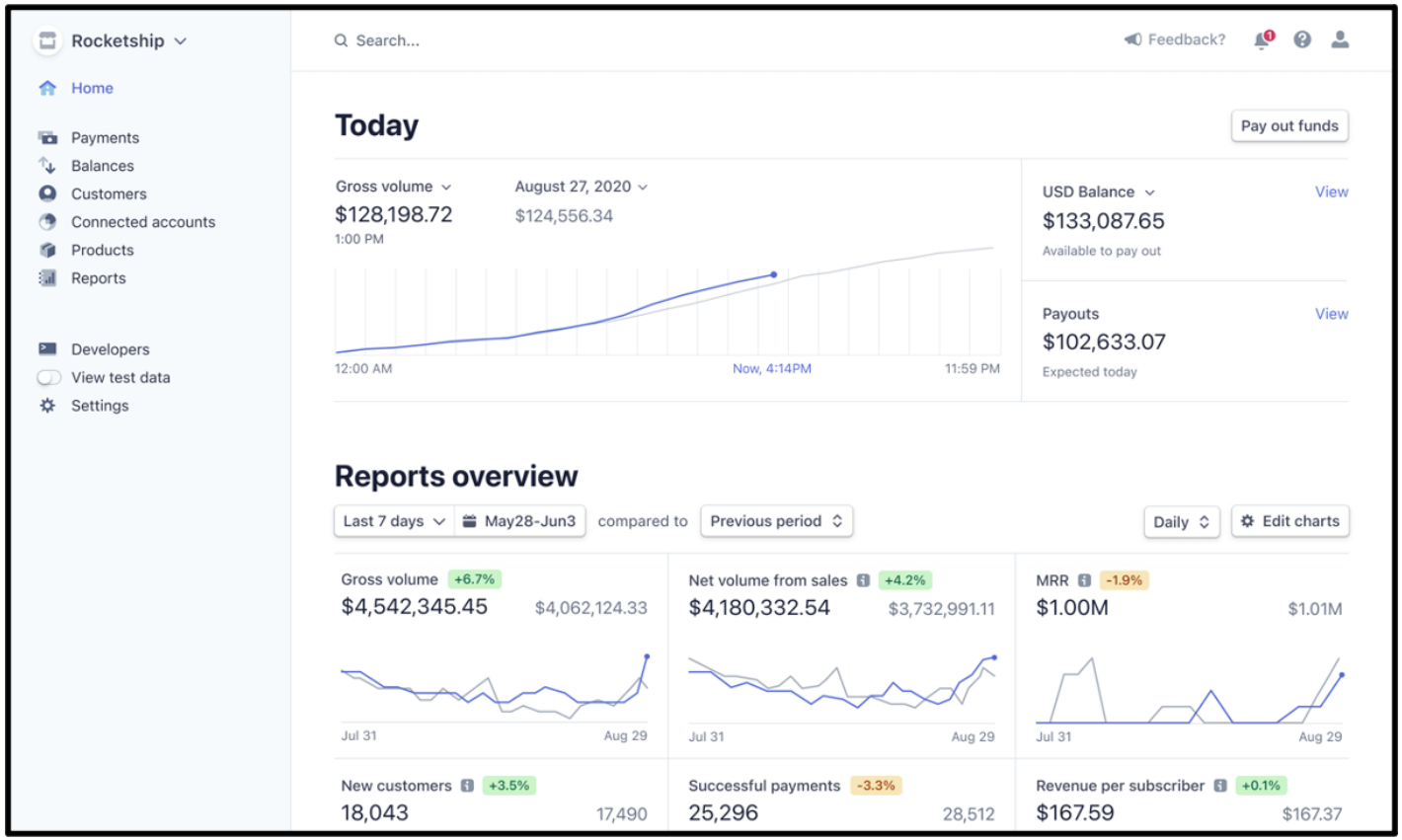
Errors in past sales reporting, could showing data help?



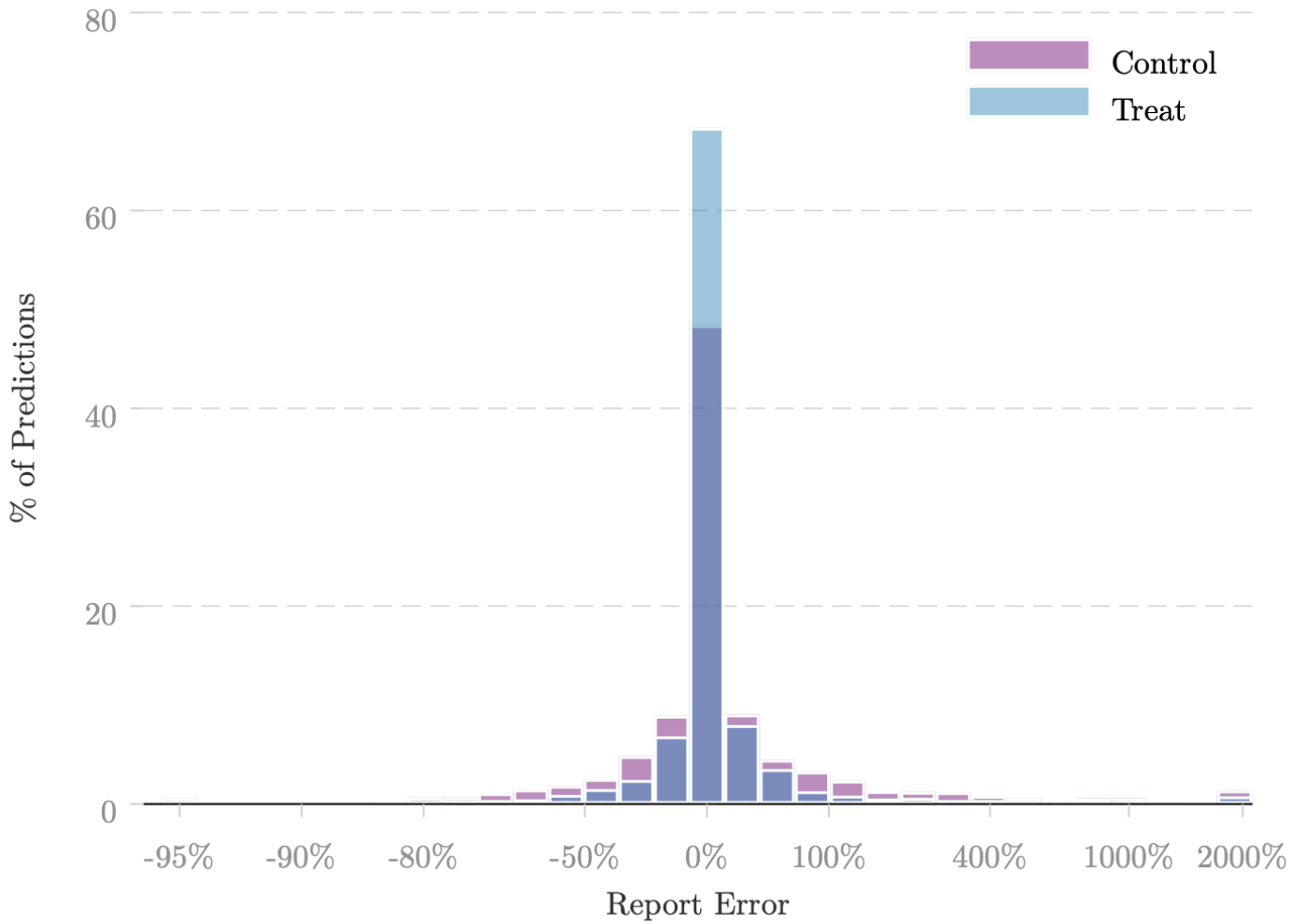
Note: Reporting error is calculated as log(reported last quarter sales) – log(last quarter sales). Results for rounds 1-3, 5300 firms. ⏩ Heterogeneity ⏩ Accuracy 12-Month

We ran an experiment using the dashboard

- The dashboard is a simple presentation of firms' revenue data offering comprehensive information on past sales, customers, payouts, etc.
- They were asked in the survey to use the dashboard to report their ID number and last quarter's revenue



Dashboard led to more accurate revenue reporting



Note: Reporting error is calculated as log(reported last quarter sales) – log(last quarter sales). Data is from rounds 5 through 7, 3,975 firms.

Treatment reduction in reporting and forecast error

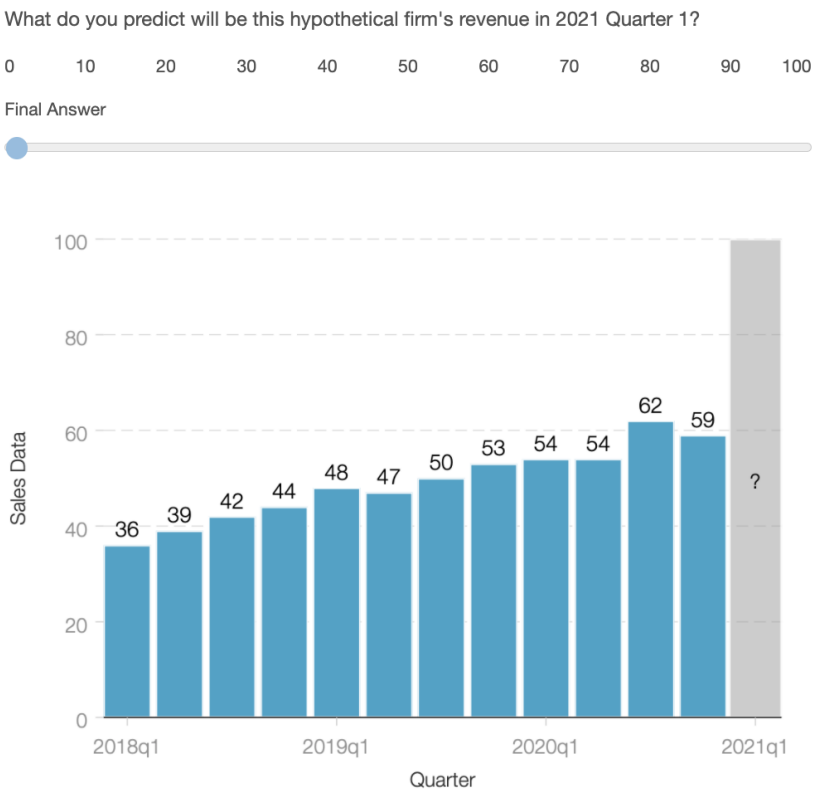
	Report Err.	(Report Err.)^2	Forecast Err.	(Forecast Err.)^2
Dashboard	-0.022	-0.217***	-0.012	-0.114**
	(0.022)	(0.042)	(0.029)	(0.054)
Time FEs	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes
Dep. Mean	0.107	0.695	0.099	1.034
Observations	6659	6659	6659	6659

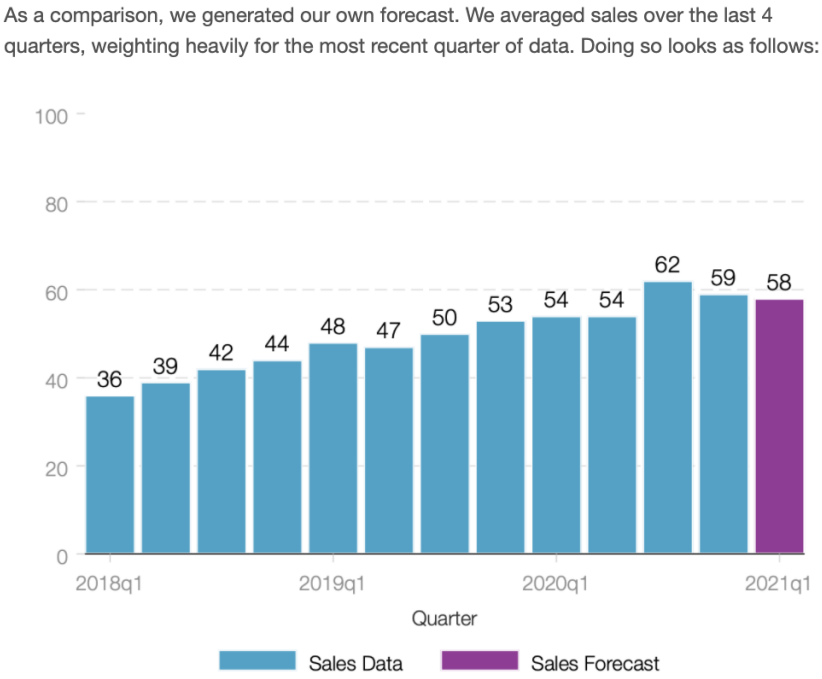
Note: Regression of e.g. log(forecast next quarter sales) – log(realization of next quarter sales) on the dashboard treatment for forecasts within 10% of actual. Data from rounds 1 through 7, with standard errors clustered at the firm level.

- Some heterogeneity too: main reduction in forecasting error from dashboard use occurs in smaller firms ⏩ Dashboard Treatment Effect By Size

⏩ Dashboard Treatment Effect By Views ⏩ Dashboard Treatment Effect By Stripe Usage

We ran an experiment training forecasting heuristics

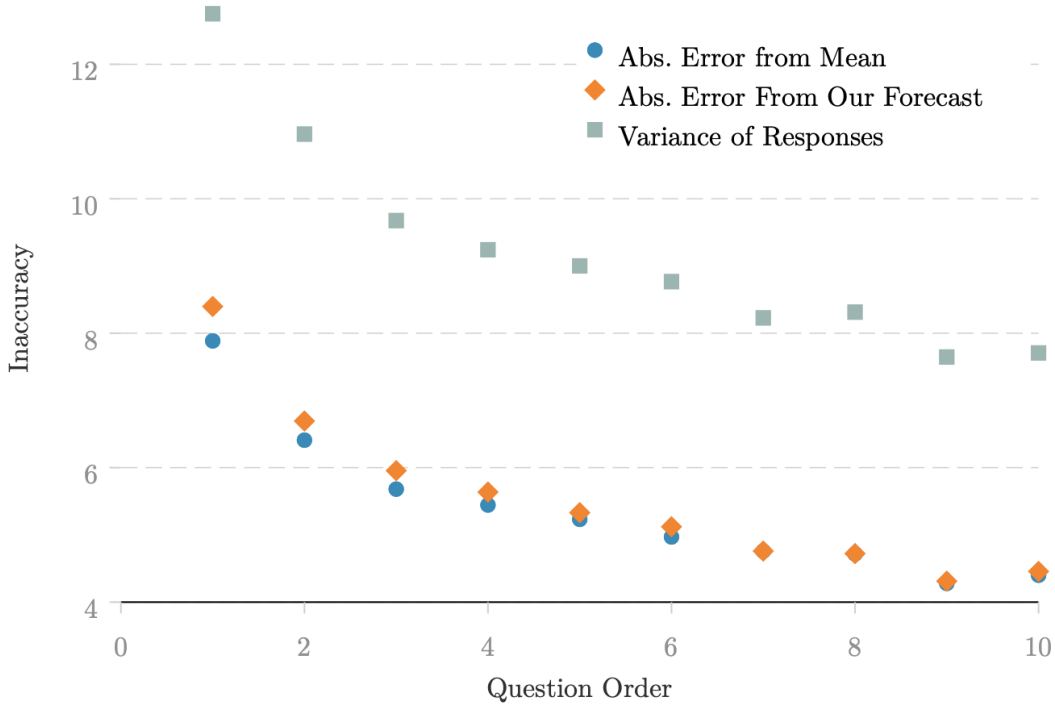




Note: These suggested forecasts were created using an autoregressive model of next quarter's revenue on the previous four quarters using Stripe firms in the sample.

Our Forecast

Respondents learned from training



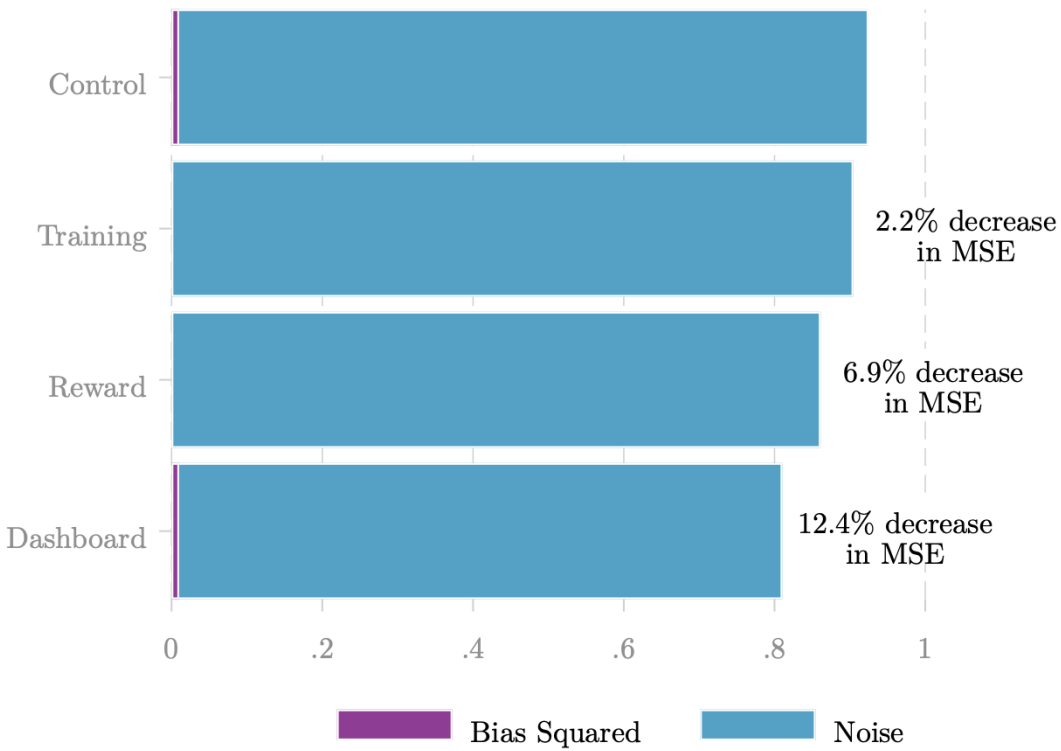
Note: Absolute error from the mean is calculated as the absolute difference between the forecast and the average forecast. Variance of responses is the variance of all responses for a given question number. Survey participants answered all 10 questions in a random order. Data from 2,953 firms in Round 7. ⏩ Forecasting Vs. Training

Training effect on accuracy is null

	Report Err.	(Report Err.)^2	Forecast Err.	(Forecast Err.)^2
Training	-0.012	0.053	-0.029	-0.020
	(0.041)	(0.081)	(0.055)	(0.097)
Time FEs	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes
Dep. Mean	0.096	0.592	0.090	0.946
Observations	6659	6659	6659	6659

Note: Regression of e.g. log(forecast next quarter sales) – log(realization of next quarter sales) on the dashboard treatment for forecasts within 10% of actual. Data from rounds 1 through 7, with standard errors clustered at the firm level.

Noise reductions drive reductions in MSE



Note: Data is from rounds 5 through 7, 3,975 firms. Reward effects are calculated for the average payment value in our experiment of $200.

Summary

- Presence of biases in entrepreneurial forecasting: over-confidence, over-precision, Dunning-Kruger effect on forecasting ability
- Staggering in our setting: almost all of forecasting errors caused by noise, not bias. Understanding and mitigating uncertainty is key
- We tried to improve their forecasts:
  - Attention: Monetary incentives reduces bias
  - Data: Reviewing historical data reduces noise, $\downarrow$ MSE
  - Skill: Forecasting training has small effect
  - Entrepreneurs do not seem to understand the benefits of data usage
- Overall effects of interventions are small! **We might have been over-confident ourselves...**

Thank you!

You can contact me at:

 mihaic@stanford.edu
 /in/mihai-alexandru-codreanu
 m_codreanu
 profiles.stanford.edu/mihai-codreanu

Or, find more about my research at:

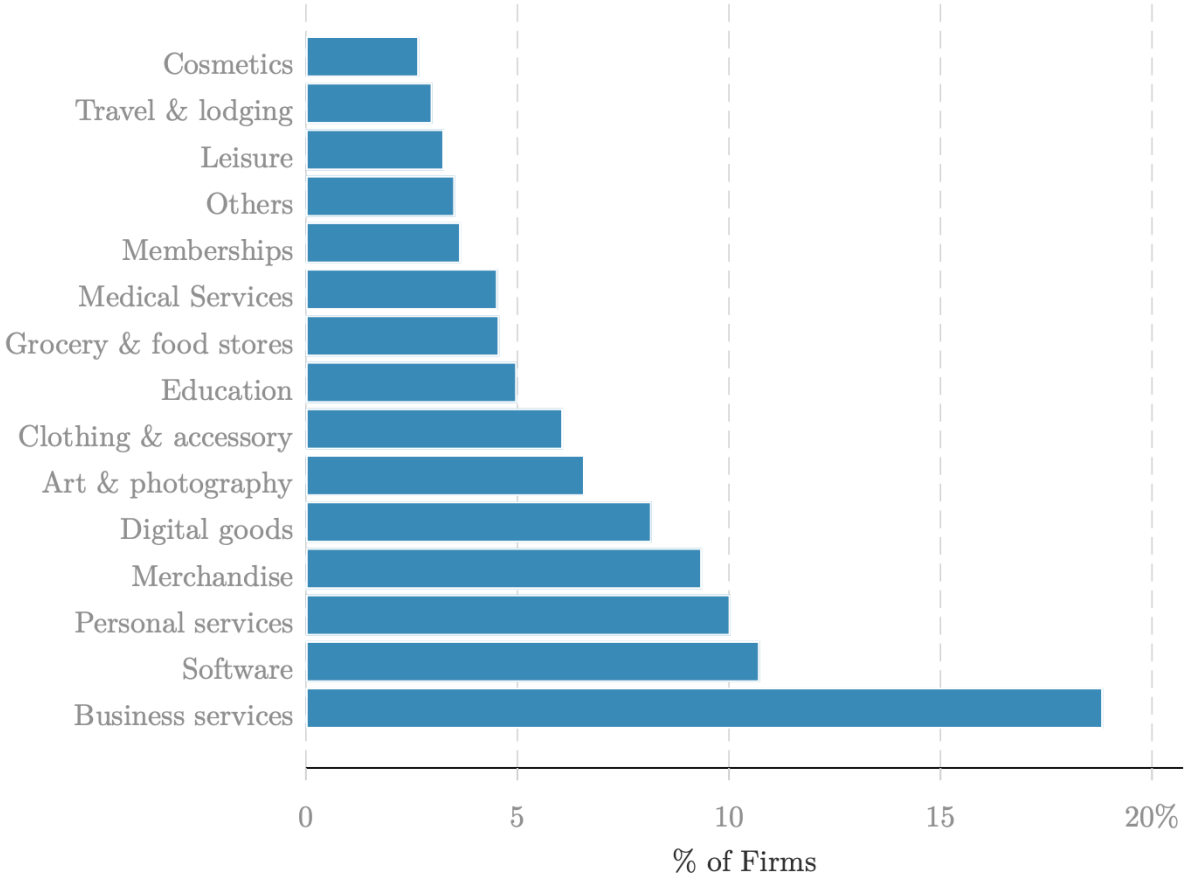
http:// www.mihaicodreanu.net

Related literature

- Firms have rational expectations? Probably not (Gennaioli, Ma, and Shleifer, 2016)
- New literature suggests they could be over-confident (e.g Malmendier and Tate, 2015). Bloom et al. (2019): more productive/better managed firms have improved forecast accuracy (concurs with Massenot and Pettinicchi, 2018; Bachmann and Elstner, 2015)
- Our paper somewhat similar to Mellers et al. (2014) and Satopää et al. (2021), which analyze the effect of various interventions on global event forecasting performance. Recommendations: training, teaming, and tracking, as well as "wisdom of the crowds"
- However, intuitively much harder to do this at micro level for firm-specific events

⏩ Overview

Industries



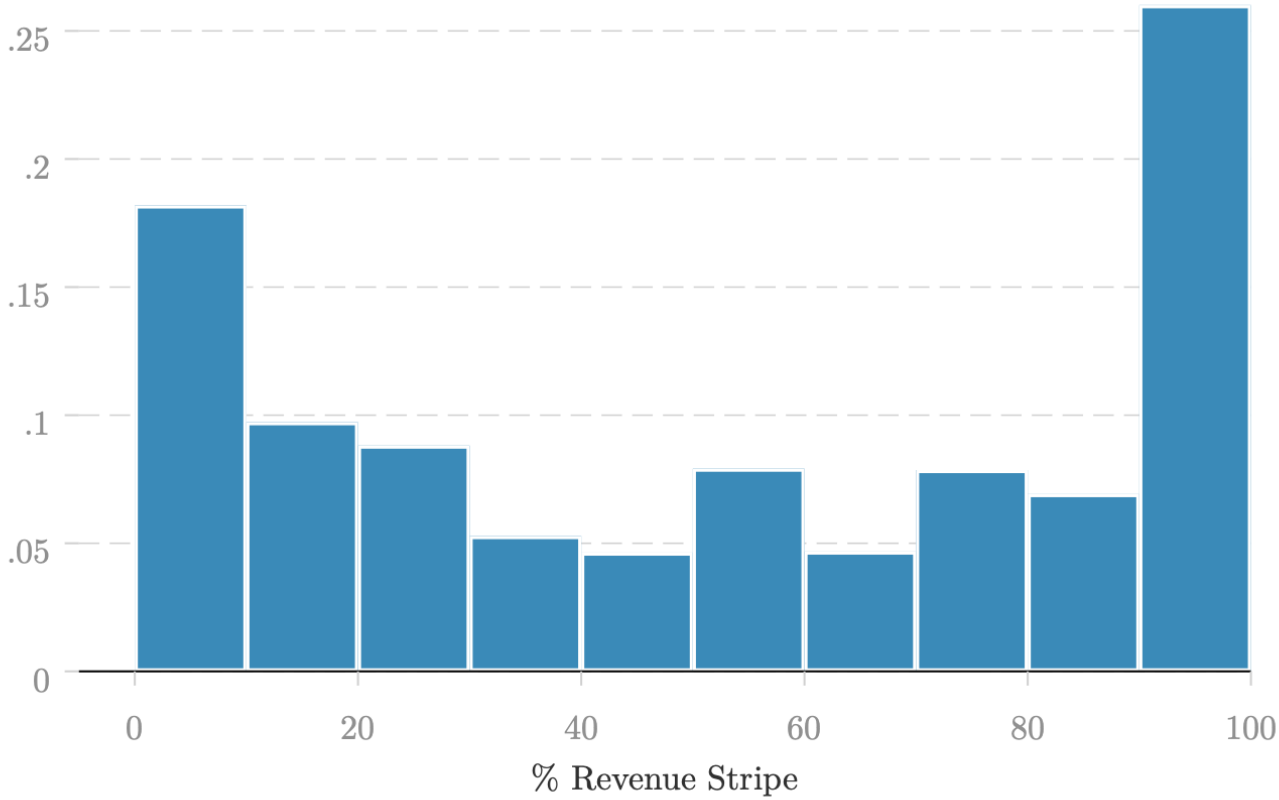
Note: Data for firms comes from 7,463 survey responses on the Stanford-Stripe Study of Internet Entrepreneurship. ⏩ Survey sample

Summary Statistics at Entry

	Sample size	Average	Median	Std. Dev.	Min	Max
Firm Characteristics						
Number Founders	6630	1.5	1	0.8	1	5
Number Employees	6630	10.4	2	203.4	1	16000
% Revenue Online	6630	67.5	90	37.6	0	100
% Revenue TechCo	6630	52.1	50	36.1	0	100
% Revenue International	4586	8.6	0	19.0	0	100
Revenue past 12 mo. ('000)	6630	403.8	80	795.2	2	3070
TechCo Revenue past 12 mo. ('000)	6630	151.7	24	337.6	0	1400
Firm Age	6579	5.9	4	6.0	0	81
Funded Flag	6630	0.20	0	0.40	0	1
Entrepreneur Characteristics						
Age	6630	39.2	37	10.7	16	100
Hours worked (per week)	6630	40.3	40	22.2	0	100
Earnings from firm past 12 mo.	6630	51.5	30	60.3	0	215
Number Businesses Owned	6630	1.5	1	0.8	1	5
Number Previous Businesses	6630	1.0	0	1.3	0	5
Has Other Job Flag	6630	0.3	0	0.4	0	1
Total sample size	7463					
Valid sample size	6630					

Note: Data for firms comes from 7,463 survey responses on the Stanford-Stripe Study of Internet Entrepreneurship. ⏩ Survey sample

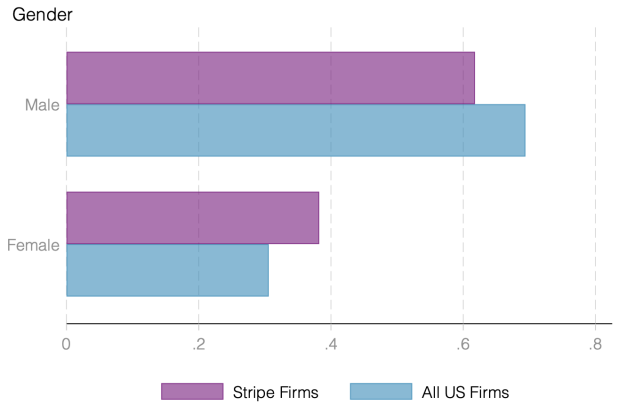
% of Revenue on Stripe

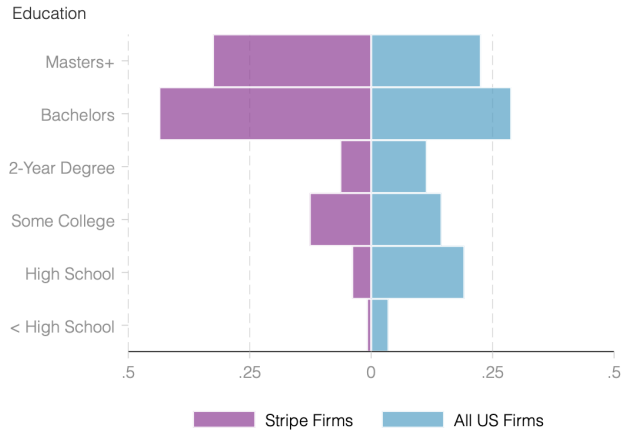


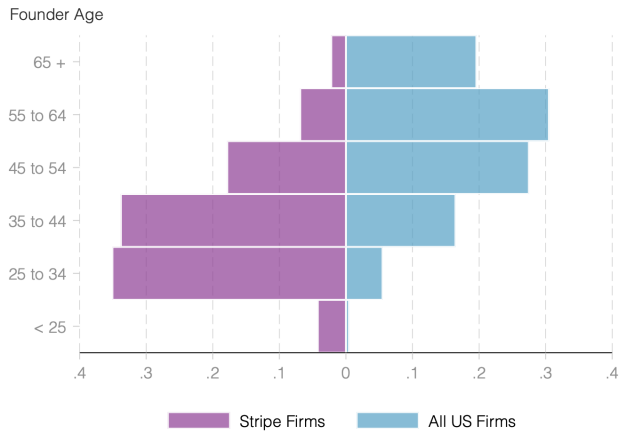
Note: Data for online firms comes from survey responses to the Stanford-Stripe Study of Internet Entrepreneurship.

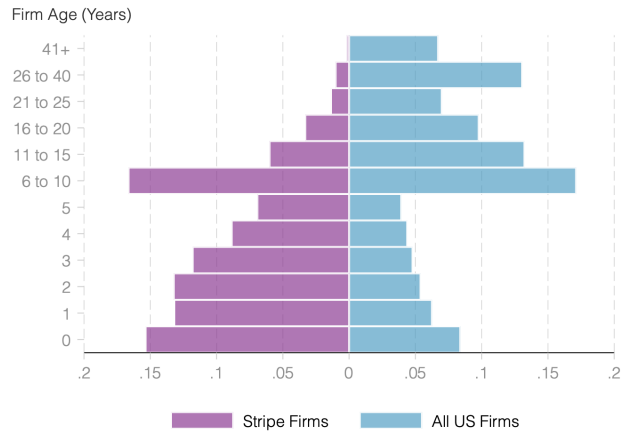
Note: Self-reported data on 7,463 firms participating in Rounds 1-7. ⏩ Survey sample

Comparison of Sample Users Vs. U.S. Businesses



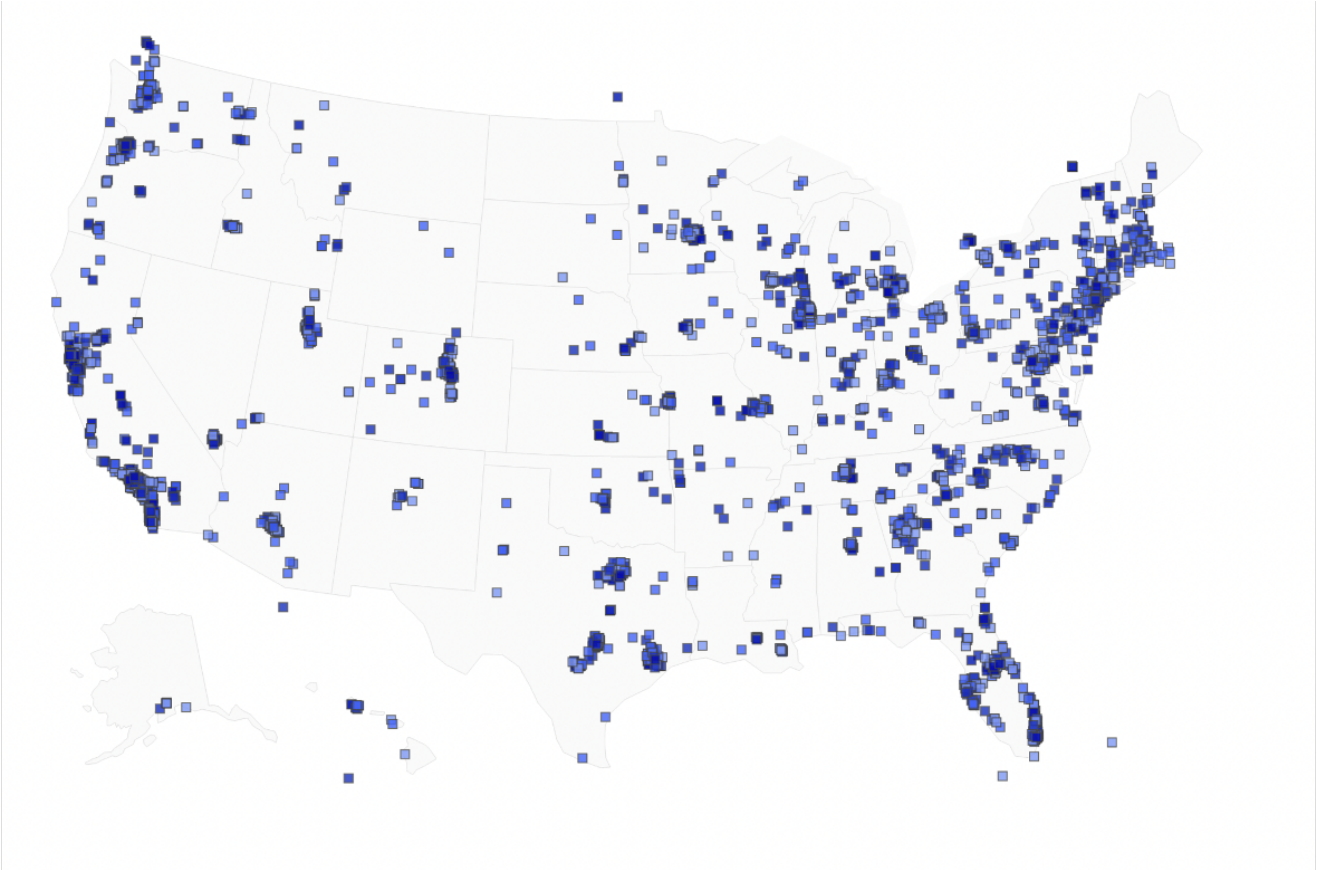






⏩ Survey sample

Geography of Businesses in Our Sample



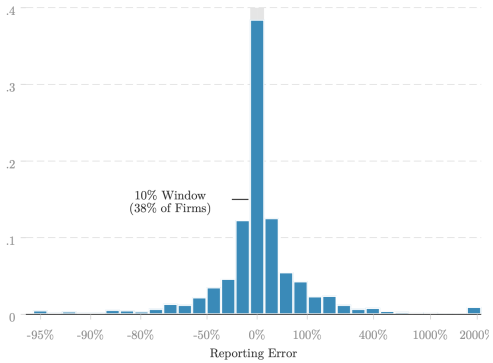
Note: Data for firms comes from rounds 1-6, 5,291 survey responses on the Stanford-Stripe Study of Internet Entrepreneurship. ⏩ Survey sample

Propensity to Respond

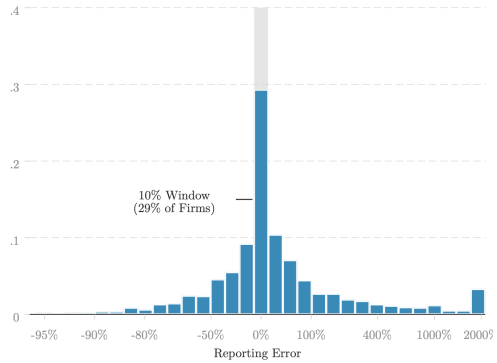
	Finished	Finished	Finished	Finished
Log Revenue	-0.003***	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Funded		-0.048***	-0.046***	-0.044***
		(0.006)	(0.006)	(0.006)
Industry FEs			Yes	Yes
Region FEs				Yes
F-test Industry			0.000	0.000
F-test Region				0.001
R2	0.001	0.003	0.009	0.010
Adj R2	0.000	0.003	0.008	0.008
Dep. Mean	0.227	0.227	0.227	0.227
# Obs	23069	23069	23069	23060

Notes: Data for firms comes from 7,463 survey respondents in the Stanford-Stripe Study of Internet Entrepreneurship. Finishing corresponds with ever completing a survey. ⏩ Survey sample

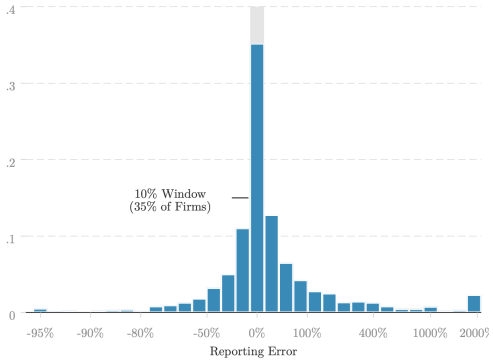
Reporting Accuracy by Business Type



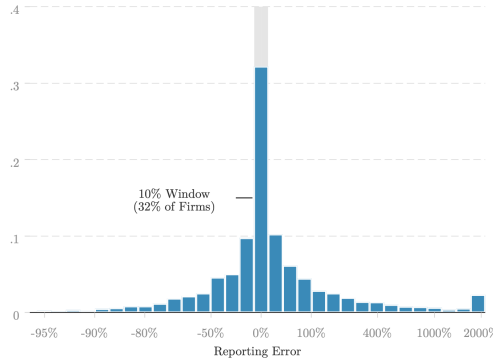
(a) Big Business



(b) Small Business



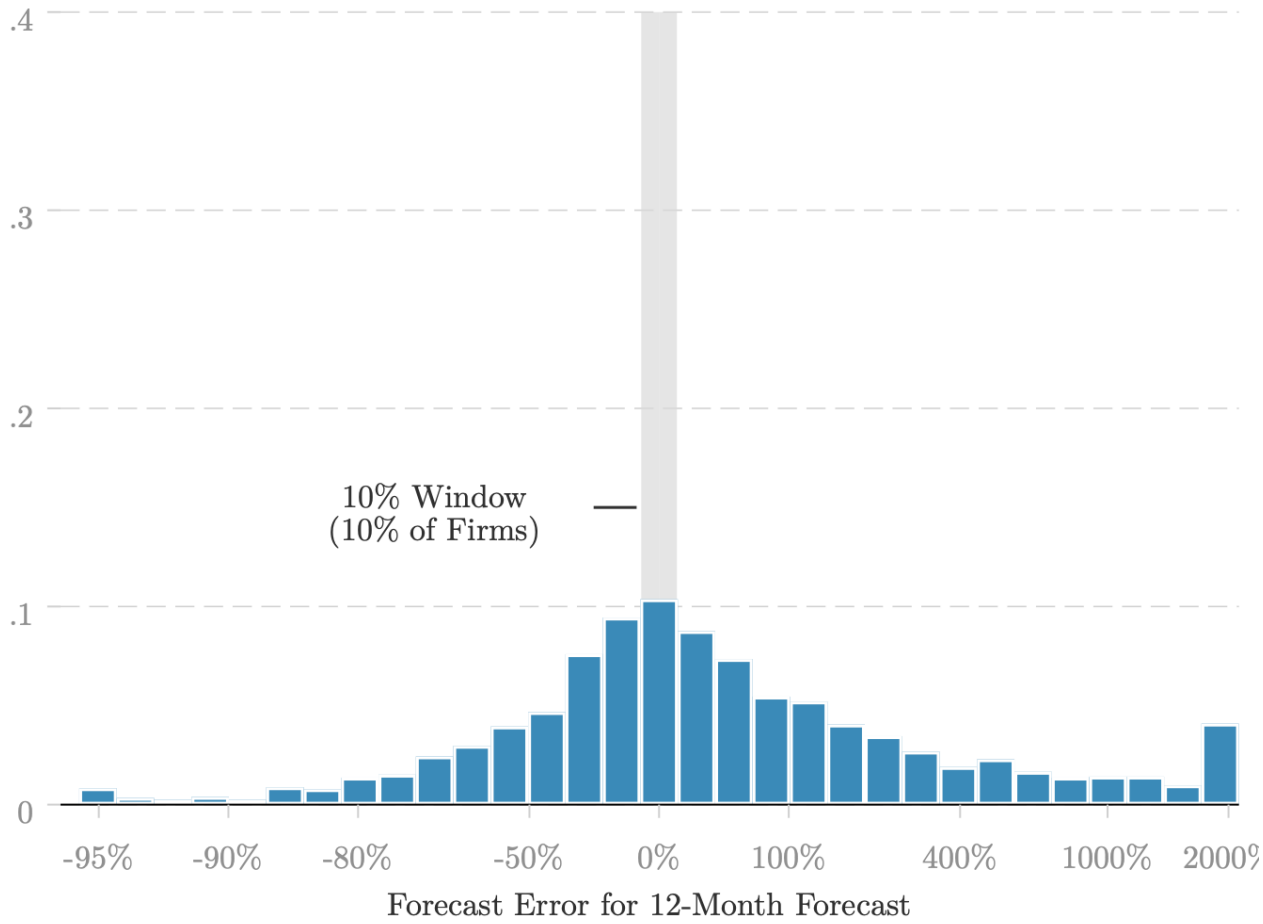
(c) High Stripe User



(d) Low Stripe User

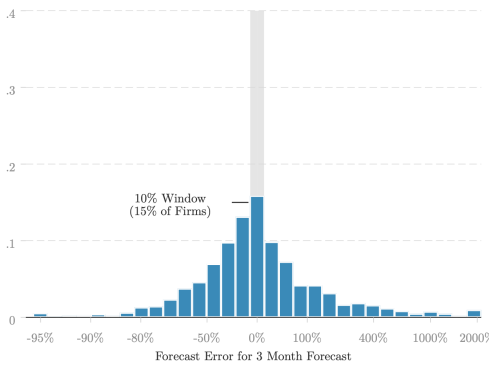
Big and small are defined using our strata definition. High and low Stripe user are above and below 50% of sales on Stripe, respectively. ⏩ Reporting Accuracy

Forecasting accuracy for 12-Months isn't any better

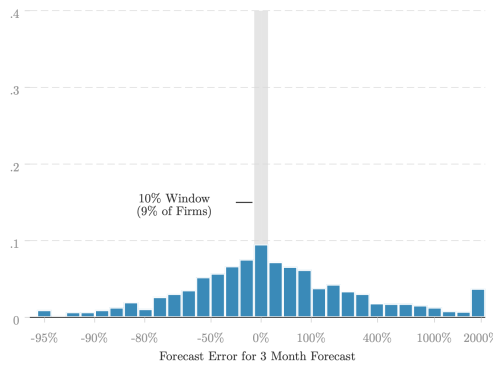


Note: Forecasting error is calculated as log(forecast next year's sales) – log(realization of next year's sales). Results for rounds 1-3, 5,300 firms (note Covid effect). ⏩ Forecasting Accuracy

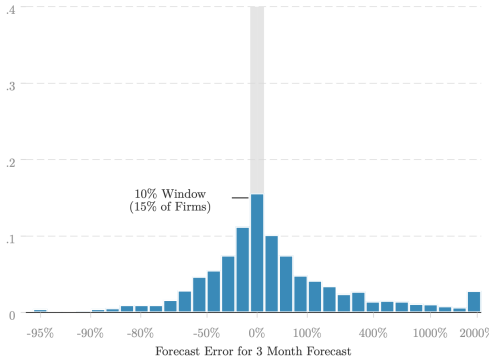
Forecasting Accuracy by Business Type



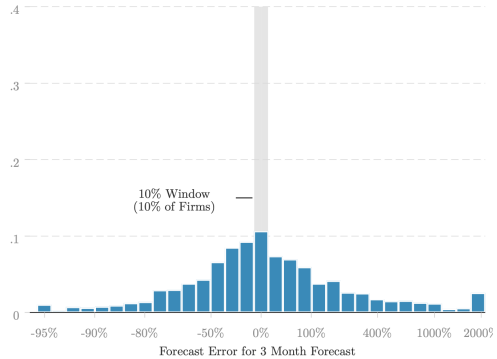
(a) Big Business



(b) Small Business



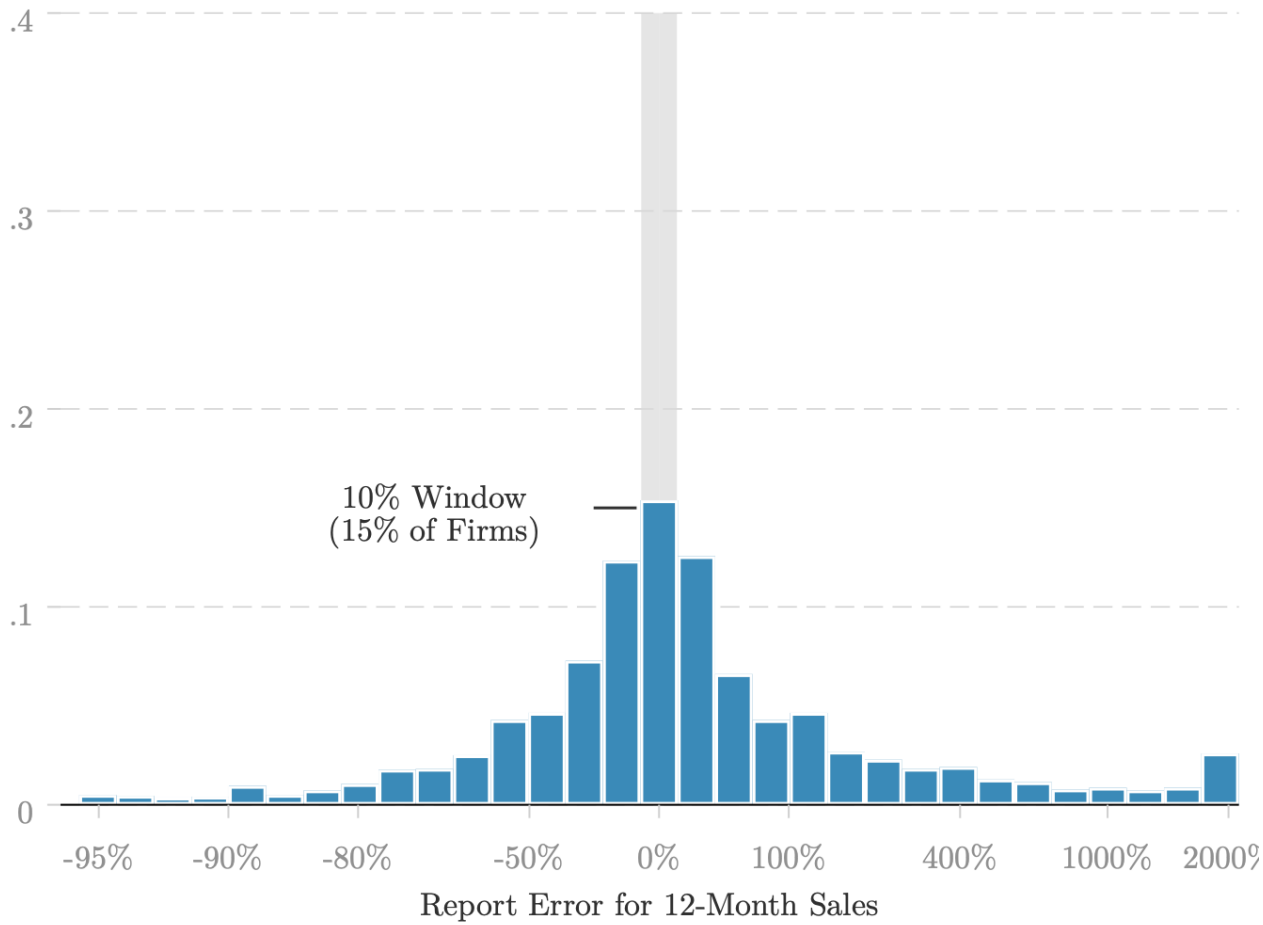
(c) High Stripe User



(d) Low Stripe User

Big and Small are defined using our strata definition. High and low Stripe user are above and below 50% of sales on Stripe, respectively. ⏩ Forecasting Accuracy

Reporting accuracy for 12-Months isn't any better



Note: Reporting error is calculated as log(reported last year sales) – log(last year sales). Results for rounds 1-3, 5,300 firms. ⏩ Reporting Accuracy

Autoregression of Sales for Suggested Forecast

	AsinhRev	AsinhRev	AsinhRev	AsinhRev
L1AsinhRev	0.997***	0.793***	0.740***	0.729***
	(0.000)	(0.003)	(0.004)	(0.005)
L2AsinhRev		0.204***	0.174***	0.153***
		(0.003)	(0.005)	(0.005)
L3AsinhRev			0.084***	0.037***
			(0.003)	(0.004)
L4AsinhRev				0.080***
				(0.003)
Dep. Mean	9.570	9.620	9.665	9.699
Coef. Sum	0.997	0.997	0.998	0.999
R-Squared	0.986	0.989	0.989	0.990
Adj R-Squared	0.986	0.989	0.989	0.990
# Obs	309896	274688	241655	210676

Note: Calculated using all 26,000 firms that were sampled prior to round 6

⏩ Training Introduction

Reward increases time on prediction

	Time (s) Prediction	Time (s) Prediction	Time (s) Prediction
Reward '00s	5.762***		5.763***
	(0.778)		(0.778)
Dash Treat		-0.406	-0.462
		(1.562)	(1.545)
Dep. Mean	56.369	56.369	56.369
Observations	3177	3177	3177

Notes: Time has been trimmed to drop respondents who went through the survey in times too short to have read and comprehended the questions.

⏩ Reward Impact

Effect of Reward treatment on Other Question Timing

	Time (s) Past Sales	TimingGoodBad3Months	TimingProb
Reward '00s	-1.206	0.678***	0.306
	(1.443)	(0.255)	(0.234)
Dep. Mean	70.470	22.615	23.899
Observations	1167	3528	3525

Note: Past sales are asked about prior to the reward treatment, while question on good and bad cases and probabilities of outcomes occur after. ⏩ Reward Impact

Dashboard Effect by Firm Size

	Report Err.	(Report Err.)^2	Forecast Err.	(Forecast Err.)^2
Low Rev. X Dashboard	-0.065*	-0.354***	-0.055	-0.207**
	(0.035)	(0.072)	(0.050)	(0.096)
High Rev. X Dashboard	0.010	-0.083*	0.022	-0.045
	(0.024)	(0.045)	(0.030)	(0.051)
F-Test Revenue	0.054	0.001	0.151	0.104
Time FEs	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes
Dep. Mean	0.107	0.695	0.099	1.034
Observations	6101	6101	6435	6435

Regression of log(forecast next quarter sales) – log(realization of next quarter sales) on the dashboard treatment for forecasts within 10% of actual. Data from rounds 2 through 6, with standard errors clustered at the firm level.

⏩ Dashboard Treatment Effect

Dashboard Effect by Dashboard Usage

	Report Err.	(Report Err.)^2	Forecast Err.	(Forecast Err.)^2
Low Views X Dashboard	-0.056*	-0.292***	-0.038	-0.131
	(0.031)	(0.066)	(0.047)	(0.088)
High Views X Dashboard	0.003	-0.146***	0.008	-0.102*
	(0.027)	(0.048)	(0.032)	(0.056)
F-Test Views	0.116	0.055	0.377	0.763
Time FEs	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes
Dep. Mean	0.107	0.695	0.099	1.034
Observations	6310	6310	6659	6659

Regression of log(forecast next quarter sales) – log(realization of next quarter sales) on the dashboard treatment for forecasts within 10% of actual. Data from rounds 2 through 6, with standard errors clustered at the firm level.

⏩ Dashboard Treatment Effect

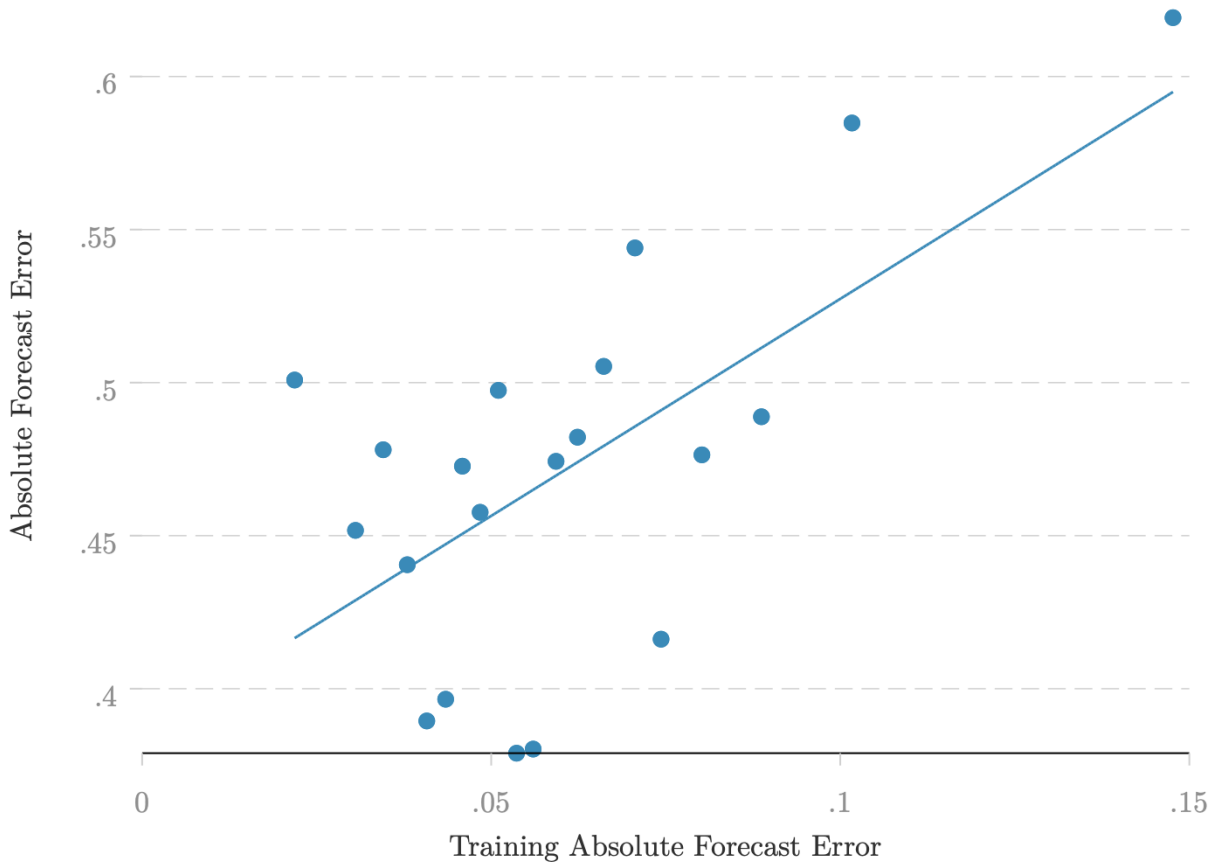
Dashboard Effect by Stripe Usage

	Report Err.	(Report Err.)^2	Forecast Err.	(Forecast Err.)^2
Low Stripe User X Dashboard	-0.053*	-0.329***	-0.062	-0.206**
	(0.031)	(0.062)	(0.046)	(0.082)
High Stripe User X Dashboard	-0.001	-0.143***	0.022	-0.054
	(0.025)	(0.049)	(0.033)	(0.060)
F-Test Stripe Use	0.152	0.008	0.101	0.086
Time FEs	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes
Dep. Mean	0.107	0.695	0.099	1.034
Observations	6813	6813	6658	6658

Regression of log(forecast next quarter sales) – log(realization of next quarter sales) on the dashboard treatment for forecasts within 10% of actual. Data from rounds 2 through 6, with standard errors clustered at the firm level.

⏩ Dashboard Treatment Effect

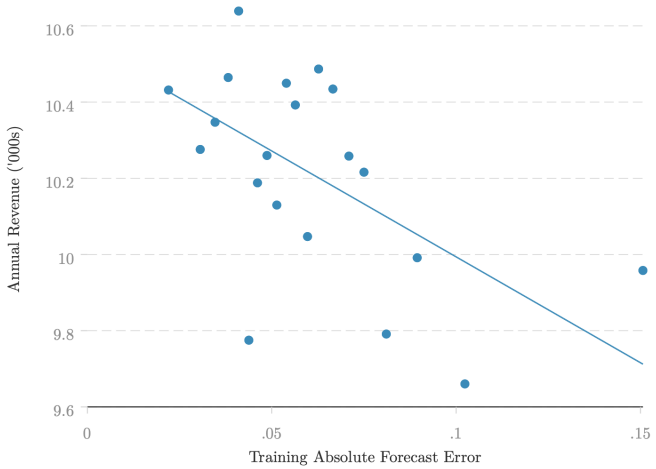
Forecasting Errors Vs. Training Errors



Note: Forecast error are historical data from rounds 1 through 7. Training forecast errors are from the training module in round 7.

Training Learning

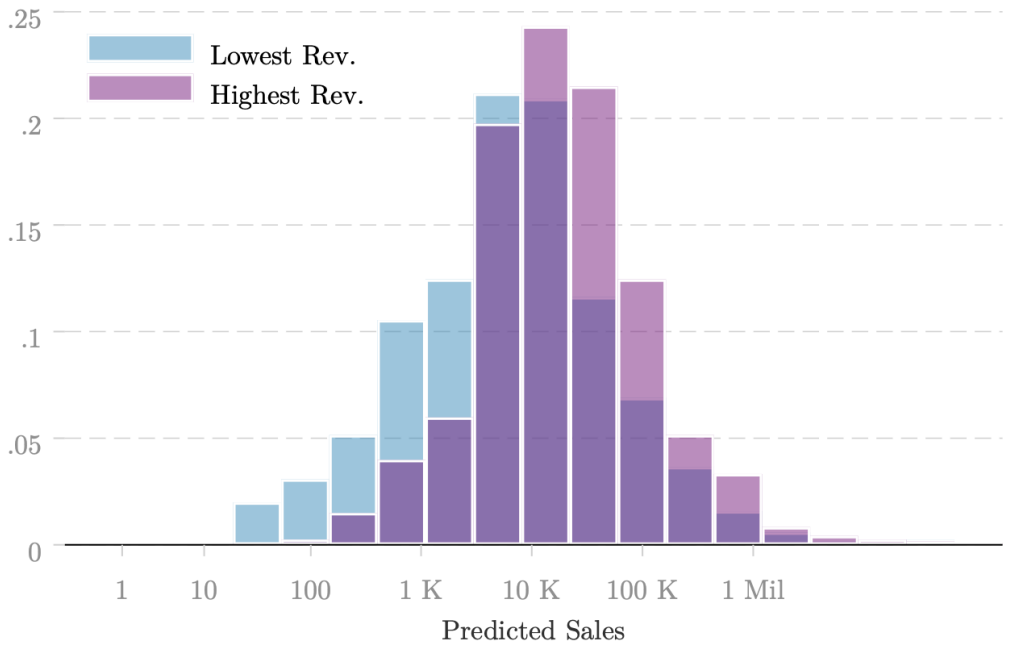
Errors negatively correlated with training performance



Note: Annual revenue, quarterly growth, semi-annual survival, and annual forecast error are historical data from rounds 1 through 7. Training forecast errors are from the training module in round 7.

⏩ Forecasting Vs. Performance

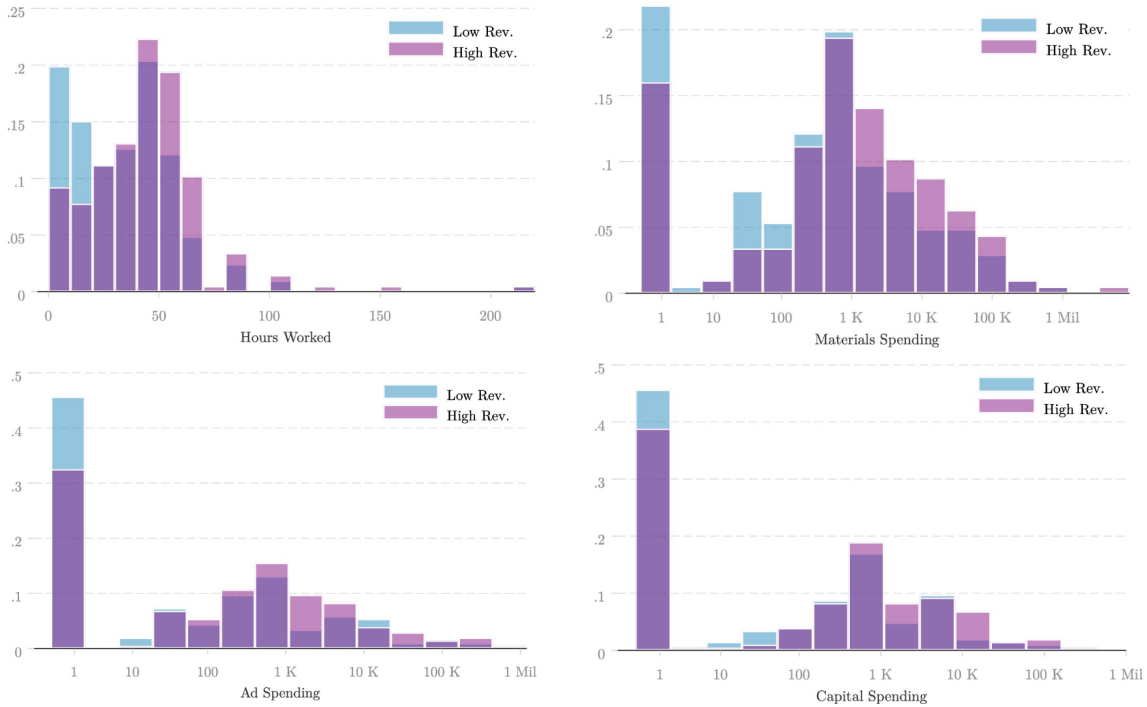
Highest and lowest cases sales predictions



Note: Data for online firms comes from survey responses to the Stanford-Stripe Study of Internet Entrepreneurship.

Notes: Changes between self-reported highest and lowest sales for the next quarter scenarios. Only first 1,640 firms participating in wave 8.

Changes in inputs in different states



Notes: Changes between highest and lowest sales scenarios for hours worked (ULS) and materials (URS), advertising (DLS) and capital (DRS). Only first 1,640 firms participating in wave 8. ⏩ Forecasting Importance