Ants That Move the Log:

Crashes, Distorted Beliefs, and Social Transmission

Qian Yang

Michigan State University
yangqia8@msu.edu

December 28, 2022

Motivation

- Conventional view in asset pricing and microstructure:
- Retail investors ≈ Noise traders, uncorrelated, inconsequential (Black, 1986, Kyle, 1985)
- Institutional investors ≈ Marginal investor

$$P = f(Trade_{informed}) + \epsilon \quad (1)$$

Retail Trading Volume Surge

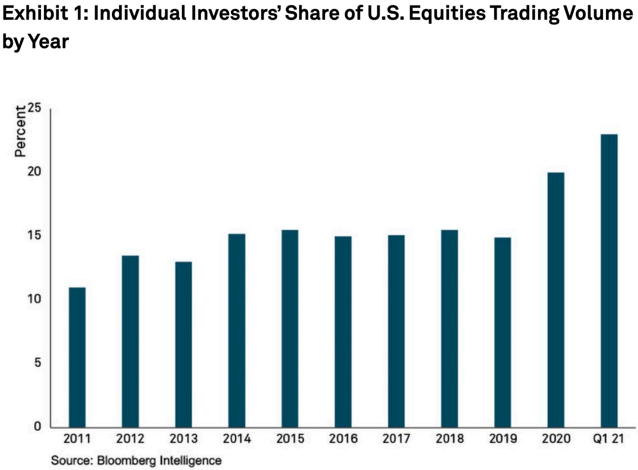


Figure: Retail Share

GameStop Saga 2021

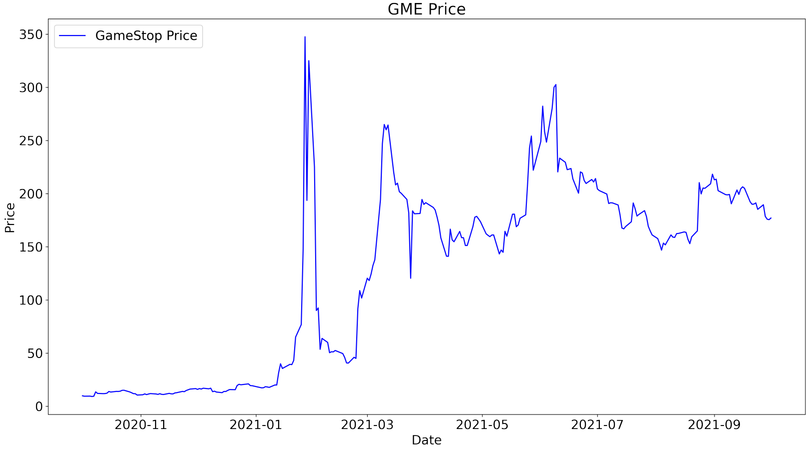


Figure: GameStop Price

With a Little Help from Social Media

Figure: Ants Moving the Log

Image credit: https://www.istockphoto.com/photos/ants-carrying-log-teamwork.

Social Transmission

Presidential Address:
Social Transmission Bias in Economics and Finance
"...a new intellectual paradigm, social economics and finance – the study of the social processes that shape economic thinking and behavior. This emerging field recognizes that people observe and talk to each other. A key, underexploited building block of social economics and finance is social transmission bias: systematic directional shift in signals or ideas induced by social transactions...For example, social transmission bias compounds recursively, which can help explain booms, bubbles, return anomalies, and swings in economic sentiment."

– David Hirshleifer, *Journal of Finance, 2020*

Research Questions

- Can social transmission contribute to stock price crash risk (left-tail risk)?
- Can investor preference help explain the negative price of crash risk in the cross-section?

Why Study Crash Risk?

- Extreme returns (jumps) account for almost all daily returns (Kapadia and Zekhnini, 2019)
- 80% of equity risk premium represents compensation for shocks that coincide with returns lower than -10% (Beason and Schreindorfer, 2022)
- Ex-ante, "Less" endogenous than studying simple returns
- Crash risk is strongly linked to overvaluation (Bollen and Whaley, 2004, Kim and Zhang, 2014, Kim et al., 2016, Van Buskirk, 2011)
- High crash risk stocks **resemble** "lottery" (positive loading on *MAX*, *Tskew*, *IVOL*, etc.)

Results

- Social transmission enables retail investors to "causally" increase crash risk
  - During the first 4 months when users started to chat about a stock on "Wallstreetbets", the monthly crash risk increased by **10%**
  - At daily frequency, a one-standard-deviation increase in chatters about a stock is associated with **2%** increase in crash risk
- Retail investors (Robinhood traders) tend to buy high-crash-risk stocks, while institutions tend to sell
- Consistent with Brunnermeier et al. (2007), the price of crash risk is more negative when lagged sentiment is high
- Propose a measure of ex-ante crash risk estimated via machine learning

Literature

- Crash risk/left-tail risk: negatively associated with expected returns (Atilgan et al., 2020, Conrad et al., 2014, Jang and Kang, 2019)
- Retail investors and stock returns: attention or herding forecast subsequent returns (Barber and Odean, 2008, Barber et al., 2021); reduce market quality (Eaton et al., 2022); increase volatility (Foucault et al., 2011)
- Social transmission and returns (Bali et al., 2021, Han et al., 2022, Hu et al., 2021)
- Preference and beliefs (Barberis and Huang, 2008, Brunnermeier et al., 2007)
- Machine learning in asset pricing (Bianchi et al., 2021, Feng et al., 2020, Gu et al., 2020, Kozak et al., 2020)

Crash Risk

$$CrashRisk_{i,t} = E[P(r_{i,t} < -20\%)|X_{i,t-j}] \quad (2)$$

- Following literature (Conrad et al., 2014, Jang and Kang, 2019), crashes $\approx$ 5% of total obs
- Binary response $\rightarrow$ probabilities
- X include 204 stock characteristics (Chen and Zimmermann, 2021), 1996 – 2020
- Use both logit and machine learning side-by-side
- Monthly frequency with rolling 6-month windows
- Ex-ante, as compared to e.g. *VaR* (Atilgan et al., 2020)

Pricing

Table: Decile High-Minus-Low Portfolio Alphas

		Logit		EEC-Adaboost	
	Pricing model	Alpha	T-stat	Alpha	T-stat
VW	CAPM	-1.852	-3.730	-1.967	-4.393
	FF3	-1.842	-4.440	-1.963	-5.456
	FF4	-1.533	-3.531	-1.775	-4.636
	FF5	-0.874	-2.834	-1.120	-3.947
	FF6	-0.696	-2.263	-1.023	-3.442
EW	CAPM	-2.470	-5.571	-2.458	-5.325
	FF3	-2.461	-7.941	-2.452	-7.573
	FF4	-2.106	-7.161	-2.173	-7.005
	FF5	-1.656	-5.637	-1.783	-6.093
	FF6	-1.438	-5.788	-1.614	-5.947

Fama-MacBeth Regressions

Table: Fama-MacBeth Cross-Sectional Regressions

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Returns in %				
Crash Risk (Logit)	-0.491***	-0.453***			
	(0.080)	(0.077)			
Crash Risk (EEC)			-0.507***	-0.459***	
			(0.097)	(0.086)	
VaR1%		-0.123		-0.097	-0.246***
		(0.082)		(0.074)	(0.083)
Controls	YES	YES	YES	YES	YES
Observations	545,367	545,290	545,367	545,290	564,466
R-squared	0.083	0.086	0.083	0.085	0.084

Note: *p<0.1; **p<0.05; ***p<0.01

Aggregate Crash Risk

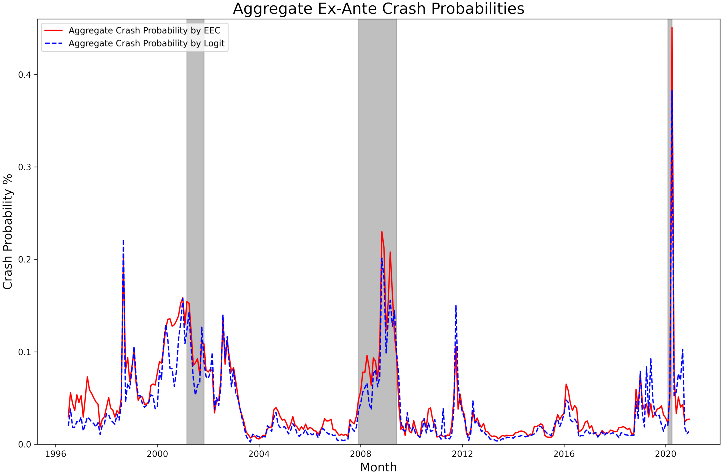


Figure: Aggregate Crash Risk

Intuition

- Investors are unable to distinguish “noise” from “signal”
- A sender shares his/her trading strategies, and receivers follow these strategies (Han et al., 2022)
- This induces herding and trading in the same direction, and thus exacerbates overvaluation

Empirical Designs

- With the caveat of unobservable trading data, we look at the direct impact of social transmission on crash risk
- Data: **ALL** Reddit comments 2012 – 2020
- Design I:
- Explore the first time (month) that every stock was mentioned on "Wallstreetbets"
- Stacked "diff-in-diffs" (Cengiz et al., 2019)
- Design II:
- Daily number of comments on "Wallstreetbets" instrumented by non-economic/financial comments

Tickers Mentioned on “Wallstreetbets”

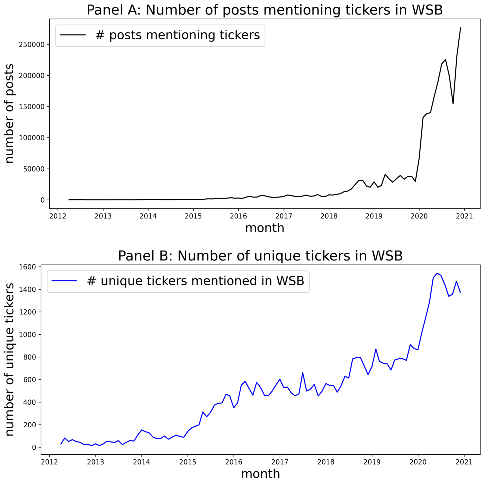


Figure: Number of Comments about Tickers & Firms

Design I: First Mentioning

- Endogenous?
- Assumption: people are less likely to buy high-crash risk stocks if they "know"
- Counterfactual: control for "lottery" characteristics (skewness, idiosyncratic risk, MAX, etc.)
- Check "parallel trends"

$$
\begin{aligned}
Crash\,Risk_{i,c,t} = {} & \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} \\
& + \sum_{p} \beta_p Control_{p,i,t-1} + \epsilon_{i,t}
\end{aligned}
$$

Diff-in-Diffs Results

Table: Debut of Stock Tickers on "Wallstreebets" and Crash Risk

VARIABLES	(1)	(2) Crash Risk (Logit)	(3)	(4)	(5) Crash Risk (EEC)	(6)
Treated	1.032***	0.560***		0.674***	0.303***	
	(0.103)	(0.129)		(0.054)	(0.064)	
Month -3			0.009			0.001
			(0.160)			(0.082)
Month -2			-0.041			0.041
			(0.140)			(0.074)
Month 0			0.464***			0.152**
			(0.136)			(0.076)
Month +1			0.326*			0.152
			(0.185)			(0.097)
Month +2			0.689***			0.478***
			(0.199)			(0.095)
Month +3			0.735***			0.508***
			(0.218)			(0.105)
Observations	208,502	125,734	125,734	208,502	125,734	125,734
R-squared	0.874	0.909	0.909	0.921	0.946	0.946
Cohort × Units FE	YES	YES	YES	YES	YES	YES
Cohort × Month FE	YES	YES	YES	YES	YES	YES

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

First Appearances of Stocks on “Wallstreetbets”

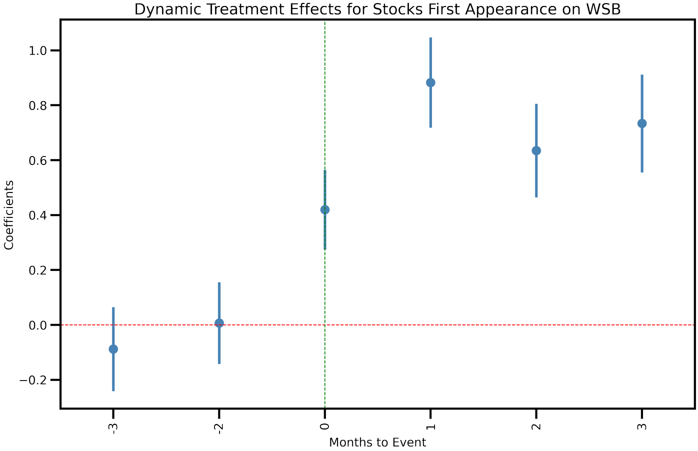


Figure: Dynamic stacked diff-in-diffs

Cross-Sectional Results: Size & IO

Table: Debut of Stock Tickers on "Wallstreebets" and Crash Risk: Size & IO

VARIABLES	(1)	(2)	(3)	(4)
	Crash Risk (Logit)			
Treated	1.501***	1.038***	1.539***	0.988***
	(0.182)	(0.326)	(0.177)	(0.310)
Treated $\times D_{size}$	-0.930***	-0.743**		
	(0.205)	(0.343)		
Treated $\times D_{io}$			-1.082***	-0.689**
			(0.202)	(0.330)
Controls	NO	YES	NO	YES
Observations	208,502	125,734	208,502	125,734
R-squared	0.874	0.909	0.874	0.909
Cohort×Units FE	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Cross-Sectional Results: Influencers

Table: Debut of Stock Tickers on "Wallstreebets" and Crash Risk: Influencers

VARIABLES	(1)	(2)	(3)	(4)
	Crash Risk (Logit)		Crash Risk (EEC)	
Treated	1.116***	0.529***	0.800***	0.313***
	(0.155)	(0.195)	(0.077)	(0.095)
Treated $\times D_{influencer}$	-0.138	0.045	-0.175*	0.028
	(0.196)	(0.236)	(0.103)	(0.125)
Controls	NO	YES	NO	YES
Observations	206,566	124,201	206,566	124,201
R-squared	0.875	0.909	0.921	0.946
Cohort×Units FE	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Cross-Sectional Results: Sentiment

Table: Debut of Stock Tickers on "Wallstreebets" and Crash Risk: Sentiment

VARIABLES	(1) Crash Risk (Logit)	(2)	(3) Crash Risk (EEC)	(4)
Treated	1.116***	0.487***	0.723***	0.301***
	(0.115)	(0.142)	(0.061)	(0.072)
Treated× Sentiment	-0.364**	0.303	-0.213**	0.009
	(0.180)	(0.211)	(0.091)	(0.103)
Controls	NO	YES	NO	YES
Observations	208,502	125,734	208,502	125,734
R-squared	0.874	0.909	0.921	0.946
Cohort×Units FE	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Trade Volume and Volatility

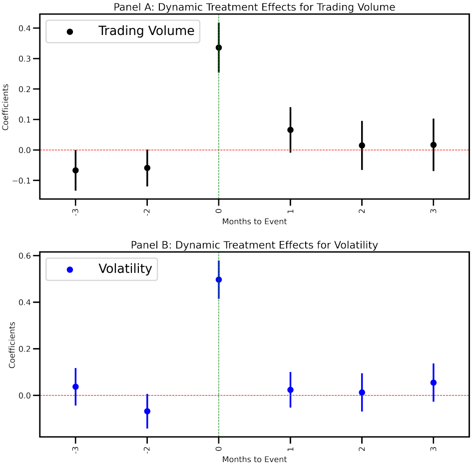


Figure: Trade Volume and Volatility

Design II: Daily Number of Comments

$$SKEW_{i,t} = ImpliedVol_{i,t}^{OTM-Put} - ImpliedVol_{i,t}^{ATM-Call} \quad (3)$$

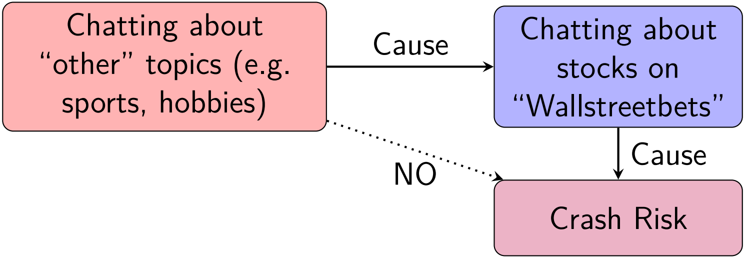
- Use *SKEW* as crash risk (Xing et al., 2010)
- Conversation is endogenous, consider IV

$$WSB\_Posts_{i,t-1} = \alpha_0 + \beta_Z Non\_Finance\_Posts_{i,t-1} + \epsilon_{i,t-1} \quad (4)$$

$$SKEW_{i,t} = \alpha_1 + \beta_X WSB\_Posts_{i,t-1} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + u_{i,t} \quad (5)$$

Intuition for IV

- Assumption: people that are active on other topics are more likely to chat about stocks
- Example: today, pre-trade hours, two persons A and B talk about $AAPL, sum all comments A and B posted on non-economic/financial “Subreddits” on Reddit



Identifying Non-Economic/Financial Subreddits

- Use natural language processing (textual analysis) on titles of Subreddits
- Follow Li et al. (2021), choose a list of "seed words" ('finance', 'stock-market', 'stocks', 'wall-street', 'trading', 'forex', 'options', 'investment', 'bond-market', 'bonds')
- Find out the top 50 words/phrases similar to each of the "seed words" (in total 371 words/phrases) via GloVe (Pennington et al., 2014) and cosine similarity:

$$CosineSim_{1,2} = \frac{V_1 \cdot V_2}{||V_1|| \cdot ||V_2||} \tag{6}$$

- Drop all "Subreddits" that contain these keywords/phrases

IV Results

Table: IV Estimation: “WSB” Posts and Crash Risk (*SKEW*)

VARIABLES	(1) Panel	(2) Panel	(3) IV
Number of “Wallstreetbets” Posts	0.070***	0.067***	0.193***
	(0.019)	(0.018)	(0.035)
Number of Non-Finance Posts		0.005	
		(0.004)	
Controls	YES	YES	YES
Observations	2,655,209	2,655,209	2,655,209
R-squared	0.089	0.089	0.042
Day FE	YES	YES	YES
Firm Cluster	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Why are crash risk negatively priced?

- REH ⇒ positive return correlation
- If stock in bubble (high crash prob), institutions less likely to arbitrage if costly (Jang and Kang, 2019)
- Investors underestimate left-tail risk (Atilgan et al., 2020)
- Underlying assumption:
- Retail traders over-buy crash-prone stocks

Do Retail Investors Buy Crash Risk?

- Retail traders have preference for "lottery-like" stocks
- Use Robintrack user change as proxy for retail trading:

$$Change\,in\,Log(\#User_{i,t}) = \log(\#User_{i,t}) - \log(\#User_{i,t-1}) \quad (7)$$

- Also use percentage change (Barber et al., 2021):

$$\%Change\#User_{i,t} = \#User_{i,t}/\#User_{i,t-1} - 1 \quad (8)$$

- Finally institutional trading:

$$IO\_Change_{i,t} = IO_{i,t} - IO_{i,t-1} \quad (9)$$

Trading Results

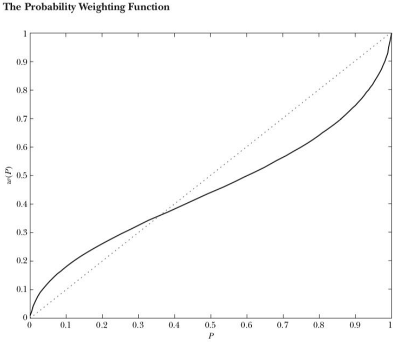
Table: Investor Trading and Crash Risk

VARIABLES	(1) Change in Log(User)	(2) User%Change	(3) IO Change
Crash Risk	0.093***	0.154***	-0.026***
	(0.010)	(0.020)	(0.002)
Controls	YES	YES	YES
Observations	63,692	63,692	375,339
R-squared	0.241	0.191	0.500
Firm & Time FE	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Cumulative Prospect Theory (CPT)

- Barberis and Huang (2008) → overweight tail probabilities → under-buy left tail → positive price



Notes: The graph plots the probability weighting function proposed by Tversky and Kahneman (1992) as part of cumulative prospect theory, namely $w(P) = P^{\delta}/(P^{\delta} + (1-P)^{\delta})^{1/\delta}$, where P is an objective probability, for two values of δ. The solid line corresponds to $\delta = 0.65$, the value estimated by the authors from experimental data. The dotted line corresponds to $\delta = 1$, in other words, to linear probability weighting.

Figure: CPT

Optimal Expectations Theory (OET)

- Brunnermeier et al. (2007) → underestimate left tail, overestimate right tail → over-buy left tail → negative price

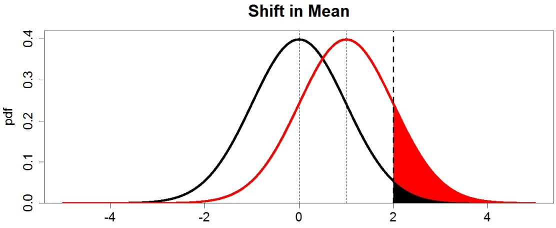


Figure: OET

Crash Risk and Sentiment

Table: Sentiment and Crash Risk Returns

	(1)	(2)	(3)	(4)
	FMB		Panel	
VARIABLES	Low Sent	High Sent	Return	Return
Crash Risk	-0.405***	-0.619***	-0.335***	-0.135**
	(0.108)	(0.141)	(0.050)	(0.062)
SentimentD×Crash Risk				-0.374***
				(0.063)
Controls	YES	YES	YES	YES
Observations	240,805	269,577	545,227	510,260
R-squared	0.078	0.085	0.168	0.159

Note: *p<0.1; **p<0.05; ***p<0.01

Discussion

- Supercharged by social media, retail investors are a force to be reckoned with
- Heightened crash risk might feed back into corporate decisions (higher risk but cheap funding)
- Firms can afford more risky projects (GameStop invested in crypto/NFT; AMC bought a gold mine)
- Future research: the real impact of “meme frenzy”

Intuition for Imbalanced Sample Problem

- Take two classes: crash and plain. Logit loss function:

$$logLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (10)$$

- If we rewrite the loss function as follows:

$$\begin{aligned} logLoss = &-\frac{1}{N_{plain}+N_{crash}}[\sum_{i=1}^{N_{plain}} \log(p_i^{plain})] \\ &-\frac{1}{N_{plain}+N_{crash}}[\sum_{i=1}^{N_{crash}} \log(p_i^{crash})] \end{aligned} \quad (11)$$

- Fix N_{crash} and let $N_{plain}/N_{crash} \to \infty$, second term $\to$ zero

Intuition for EEC-AdaBoost

- "Easy Ensemble" (EEC) (Liu et al., 2008):
  - Randomly sample a subset of non-crash obs and pair them with the crash obs
  - Fit an estimator on this sample and save the parameters
  - Repeat 50 times $\rightarrow$ 50 bootstrapped and balanced samples
  - An Ensemble is built upon these results and arrives at a final estimate
- Adaptive Boosting (Freund and Schapire, 1997) (AdaBoost):
  - Each iteration dynamically adapts to the falsely classified instances of the last iteration

Forecasting Performance

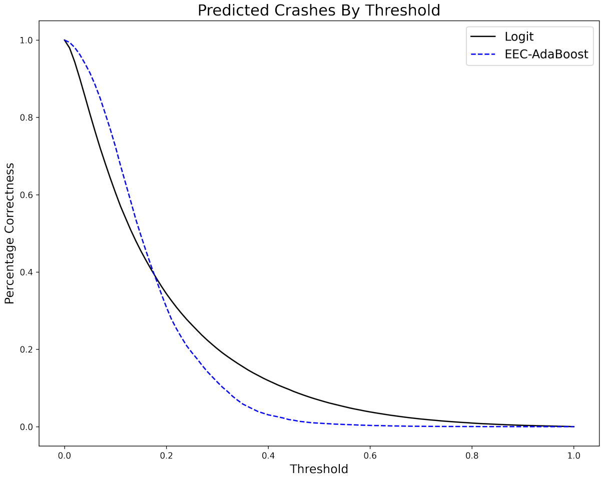


Figure: Logit versus Machine Learning

Variable Importance

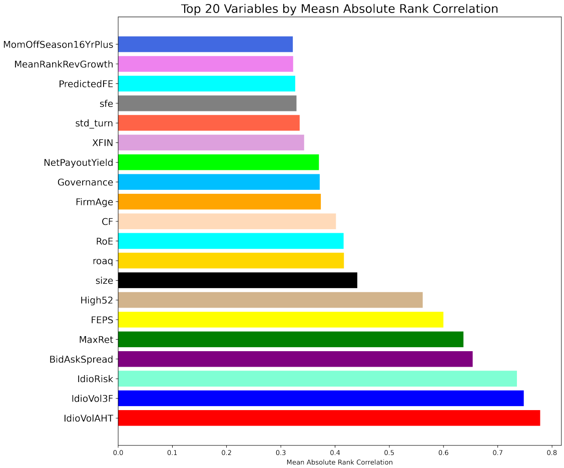


Figure: Top 20 Variables with Highest Absolute Rank Correlations with Crash Risk.

Pricing Tests for Alternative Thresholds

Table: Decile High-Minus-Low Alphas: Alternative Definitions

		Logit		EEC-AdaBoost	
Threshold	Weighting	Alpha	T-stat	Alpha	T-stat
$log(ret) < -10\%$	value	-0.405	-1.291	-1.164	-3.989
	equal	-1.637	-6.467	-1.783	-6.466
$log(ret) < -15\%$	value	-0.855	-2.920	-1.249	-4.059
	equal	-1.601	-6.704	-1.758	-6.615
$log(ret) < -25\%$	value	-0.825	-2.764	-1.157	-3.920
	equal	-1.475	-5.816	-1.716	-6.358
$log(ret) < -30\%$	value	-0.751	-2.444	-1.047	-3.714
	equal	-1.444	-5.544	-1.603	-6.120

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Realized Monthly Crashes

Table: Wallstreetbets Conversations on Realized Crashes

VARIABLES	(1) Crash10	(2) Crash15	(3) Crash20	(4) Crash25	(5) Crash30
Treated	0.015***	0.010***	0.008***	0.008***	0.011***
	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
Constant	0.170***	0.110***	0.075***	0.052***	0.035***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	215,770	215,770	215,770	215,770	215,770
R-squared	0.550	0.552	0.548	0.547	0.541
Cohort × Units FE	YES	YES	YES	YES	YES
Cohort × Month FE	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Alternative Settings for DiD

Table: Wallstreebets and Crash Risk: Alternative Settings

	(1)	(2)	(3)	(4)
	Dependent Var: Crash Risk			
	Setting 1		Setting 2	
VARIABLES	logit	EEC	logit	EEC
Treated	0.008***	0.004***	0.004***	0.004***
	(0.002)	(0.001)	(0.001)	(0.001)
Controls	YES	YES	YES	YES
Observations	51,842	51,842	211,984	211,984
R-squared	0.677	0.787	0.691	0.814
Firm & Time FE	YES	YES	YES	YES

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Option SKEW, Daily Returns, and Retail Trading

Table: Daily Returns, Retail Trading, and Crash Risk (*SKEW*)

	(1)	(2)	(3)	(4)
Panel A: Daily Stock Returns and Crash Risk (*SKEW*)				
VARIABLES	FMB		Panel	
Lag Option SKEW	-0.001***	-0.002***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Controls	NO	YES	NO	YES
Observations	2,071,209	2,010,815	2,071,209	2,010,815
R-squared	0.003	0.072	0.199	0.201
Panel B: Robinhood User Trading and Crash Risk (*SKEW*)				
VARIABLES	Change in Log(Robinhood Users)		% Change in Robinhood Users	
Option SKEW	0.001**		0.001**	
	(0.000)		(0.001)	
Controls	YES		YES	
Observations	703,614		862,423	
R-squared	0.011		0.003	

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

References I

Yigit Atilgan, Turan G Bali, K Ozgur Demirtas, and A Doruk Gunaydin. Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. *Journal of Financial Economics*, 135(3):725–753, 2020.

Turan G Bali, David Hirshleifer, Lin Peng, and Yi Tang. Attention, social interaction, and investor attraction to lottery stocks. Technical report, National Bureau of Economic Research, 2021.

Brad M Barber and Terrance Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies*, 21(2):785–818, 2008.

References II

Brad M Barber, Xing Huang, Terrance Odean, and Christopher Schwarz. Attention induced trading and returns: Evidence from robinhood users. *Journal of Finance, forthcoming*, 2021.

Nicholas Barberis and Ming Huang. Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, 98(5):2066–2100, 2008.

Tyler Beason and David Schreindorfer. Dissecting the equity premium. *Journal of Political Economy*, 130(8):2203–2222, 2022.

Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089, 2021.

References III

Fischer Black. Noise. *The journal of finance*, 41(3):528–543, 1986.

Nicolas PB Bollen and Robert E Whaley. Does net buying pressure affect the shape of implied volatility functions? *The Journal of Finance*, 59(2):711–753, 2004.

Markus K Brunnermeier, Christian Gollier, and Jonathan A Parker. Optimal beliefs, asset prices, and the preference for skewed returns. *American Economic Review*, 97(2): 159–165, 2007.

Doruk Cengiz, Arindrajit Dube, Attila Lindner, and Ben Zipperer. The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454, 2019.

References IV

Andrew Y Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*, 2021.

Jennifer Conrad, Nishad Kapadia, and Yuhang Xing. Death and jackpot: Why do individual investors hold overpriced stocks? *Journal of Financial Economics*, 113(3):455–475, 2014.

Gregory W Eaton, T Clifton Green, Brian S Roseman, and Yanbin Wu. Retail trader sophistication and stock market quality: Evidence from brokerage outages. *Journal of Financial Economics*, 146(2):502–528, 2022.

References V

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

Thierry Foucault, David Sraer, and David J Thesmar. Individual investors and volatility. *The Journal of Finance*, 66(4):1369–1406, 2011.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139, 1997.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

References VI

Bing Han, David Hirshleifer, and Johan Walden. Social transmission bias and investor behavior. *Journal of Financial and Quantitative Analysis*, 57(1):390–412, 2022.

Danqi Hu, Charles M Jones, Valerie Zhang, and Xiaoyan Zhang. The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *Available at SSRN 3807655*, 2021.

Jeewon Jang and Jangkoo Kang. Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. *Journal of Financial Economics*, 132(1):222–247, 2019.

References VII

Nishad Kapadia and Morad Zekhnini. Do idiosyncratic jumps matter? *Journal of Financial Economics*, 131(3):666–692, 2019.

Jeong-Bon Kim and Liandong Zhang. Financial reporting opacity and expected crash risk: Evidence from implied volatility smirks. *Contemporary Accounting Research*, 31(3): 851–875, 2014.

Jeong-Bon Kim, Leye Li, Louise Yi Lu, and Yangxin Yu. Financial statement comparability and expected crash risk. *Journal of Accounting and Economics*, 61(2-3):294–312, 2016.

References VIII

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 135(2): 271–292, 2020.

Albert S Kyle. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335, 1985.

Kai Li, Feng Mai, Rui Shen, and Xinyan Yan. Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315, 2021.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

References IX

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Andrew Van Buskirk. Volatility skew, earnings announcements, and the predictability of crashes. *Earnings Announcements, and the Predictability of Crashes (April 28, 2011)*, 2011.

Yuhang Xing, Xiaoyan Zhang, and Rui Zhao. What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis*, pages 641–662, 2010.