Machine learning in the Chinese stock market

Markus Leippold, Qian Wang, Wenyu Zhou

Journal of Financial Economics, 2021 forthcoming

Long Zhen

Contents

- Introduction
  - Background
  - Motivation
  - Research problem
  - Contribution
- Research Design
  - Data
  - Method
- Empirical Results
- Conclusion

Introduction – Backgrounds

- The size and specificity of the Chinese stock market make it particular attractive.
- Three key features of the Chinese stock market:
  - Dominated by retail investors→ increased turnover
  - Centrally controlled, bank-dominated, and uniquely relationship-driven
    - → SOEs' special treatment
  - Limited history of short sales
    - 2010.3 allow to short sell 90 stocks
    - 2016.12 expanded to 950 firms
    - → long-only portfolios

Introduction – Motivation

- Chinese stock market allow us to deepen our understanding of emerging market and complement our knowledge of financial systems in other institutional settings.
- Since China has been experiencing a highly dynamic development, which means highly flexible methods are required.
- Gu et al.(2020) suggest machine learning improves the description of expected return. But whether this result still hold for the Chinese market?

Introduction – Contribution

- Build a unique and comprehensive set of factors
- Apply the work of Gu et al.(2020) to the Chinese market and analysis the results reasonably.

Introduction – Research Problem

- How do the specificity of Chinese stock market influence its predictability?
- Whether machine learning methods improve the predictability just like that in US market?

Research Design – Data

- Wind & CSMAR
- 2000.1 ~ 2020.6
- All A-share stocks listed on the SH and SZ exchange
    - More than 3,900 stocks
    - Cross-sectionally rank characteristics and map into [-1,1]
    - 22 updated monthly, 51 quarterly, 6 semi-annually, 15 annually
    - 11 macroeconomic predictors (8 based on Welch(2008))
- 1-year government bond in China as the risk-free rate

Research Design

- Variable construction:
  - 90 stock-level characteristics
    - China-specific factors: abnormal turnover ratio(atr)/ the trend factor(er_trend)/ the top-10 shareholders ownership (top10holderrate)
  - 11 macroeconomic predictors
  - 80 industry dummy variables
  - → 90*(11+1) + 80 = 1,160
- Methods:
  - OLS / OLS-3 (only size, BM, and momentum)
  - PLS / Lasso / Enet
  - GBRT / RF/ VASA (variable subsample aggregation)
  - NN1 ~ NN5

Research Design

- Model selection
    - Training sample: 2000-2008
    - Validation sample: 2009-2011
    - Testing sample: 2012-2020
    - Refit annually

Empirical Results

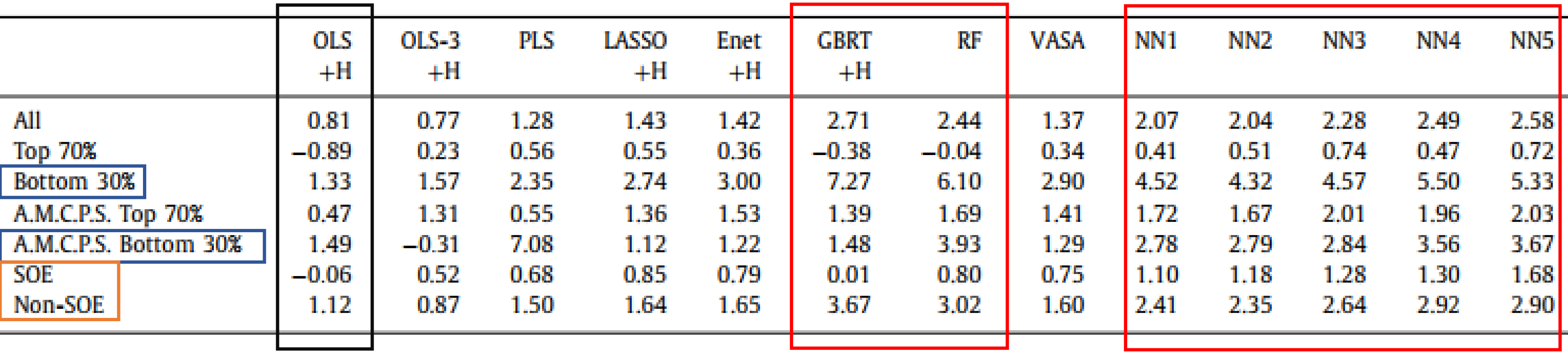
- Out-of-sample predictability

$$R^2_{\text{oos},S} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}} (r_{i,t} - \hat{r}^{(S)}_{i,t})^2}{\sum_{(i,t)\in\mathcal{T}} r^2_{i,t}}$$

  - Subsample
    - Size
    - A.M.C.P.S. = Market Cap/Number of Shareholders
    - SOE/non-SOE
- Which predictors matter?
- Portfolio analysis

- Full sample and subsample analysis

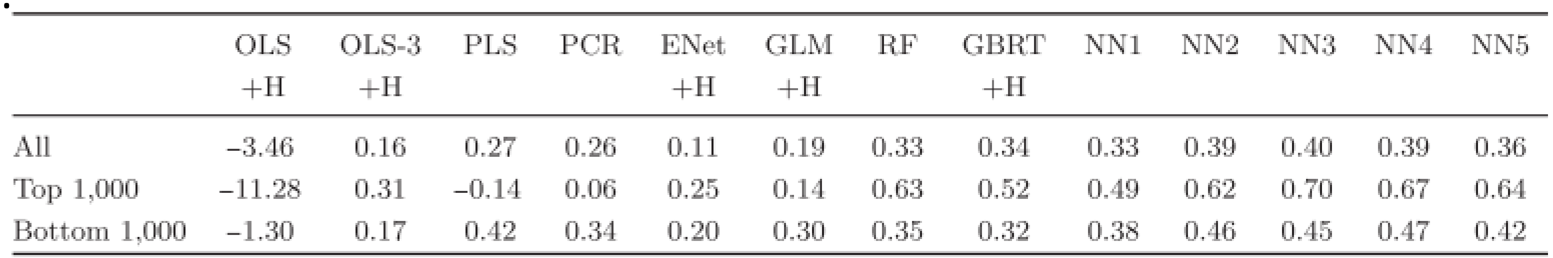
$$\hat{g}^{\text{VASA}}(\boldsymbol{z}_{i,t}) := \sum_{b=1}^{B} \omega_b \hat{g}^{\text{OLS}}(\tilde{\boldsymbol{z}}_{i,t,b}) = \sum_{b=1}^{B} \omega_b(\hat{\alpha}_b + \tilde{\boldsymbol{z}}'_{i,t,b}\hat{\boldsymbol{\beta}}_b)$$



	OLS +H	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
All	0.81	0.77	1.28	1.43	1.42	2.71	2.44	1.37	2.07	2.04	2.28	2.49	2.58
Top 70%	−0.89	0.23	0.56	0.55	0.36	−0.38	−0.04	0.34	0.41	0.51	0.74	0.47	0.72
Bottom 30%	1.33	1.57	2.35	2.74	3.00	7.27	6.10	2.90	4.52	4.32	4.57	5.50	5.33
A.M.C.P.S. Top 70%	0.47	1.31	0.55	1.36	1.53	1.39	1.69	1.41	1.72	1.67	2.01	1.96	2.03
A.M.C.P.S. Bottom 30%	1.49	−0.31	7.08	1.12	1.22	1.48	3.93	1.29	2.78	2.79	2.84	3.56	3.67
SOE	−0.06	0.52	0.68	0.85	0.79	0.01	0.80	0.75	1.10	1.18	1.28	1.30	1.68
Non-SOE	1.12	0.87	1.50	1.64	1.65	3.67	3.02	1.60	2.41	2.35	2.64	2.92	2.90

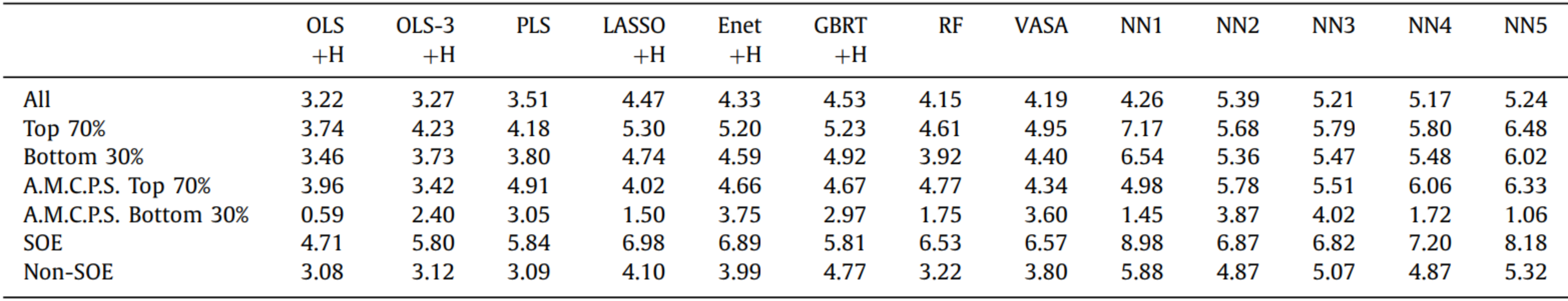
- NN and tree-based models still outperform
- Predictability is more significant for subsamples of stocks in which retail traders play a much bigger role
- Predictability of SOEs is weaker than for non-SOEs → non-transparent
- Same pattern for size and SOE/non-SOE subsamples → strongly correlated
- The Chinese market reveals substantially more predictability.

Gu et al.



	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	−3.46	0.16	0.27	0.26	0.11	0.19	0.33	0.34	0.33	0.39	0.40	0.39	0.36
Top 1,000	−11.28	0.31	−0.14	0.06	0.25	0.14	0.63	0.52	0.49	0.62	0.70	0.67	0.64
Bottom 1,000	−1.30	0.17	0.42	0.34	0.20	0.30	0.35	0.32	0.38	0.46	0.45	0.47	0.42

- At annual horizons



	OLS +H	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
All	3.22	3.27	3.51	4.47	4.33	4.53	4.15	4.19	4.26	5.39	5.21	5.17	5.24
Top 70%	3.74	4.23	4.18	5.30	5.20	5.23	4.61	4.95	7.17	5.68	5.79	5.80	6.48
Bottom 30%	3.46	3.73	3.80	4.74	4.59	4.92	3.92	4.40	6.54	5.36	5.47	5.48	6.02
A.M.C.P.S. Top 70%	3.96	3.42	4.91	4.02	4.66	4.67	4.77	4.34	4.98	5.78	5.51	6.06	6.33
A.M.C.P.S. Bottom 30%	0.59	2.40	3.05	1.50	3.75	2.97	1.75	3.60	1.45	3.87	4.02	1.72	1.06
SOE	4.71	5.80	5.84	6.98	6.89	5.81	6.53	6.57	8.98	6.87	6.82	7.20	8.18
Non-SOE	3.08	3.12	3.09	4.10	3.99	4.77	3.22	3.80	5.88	4.87	5.07	4.87	5.32

- Annual OOS R2 are higher than their monthly counterparts
- Pattern reverses for subsample results. They attribute this result to retail investors' short-term speculative behavior in the Chinese stock market.

- Monthly vs Annually / China vs US

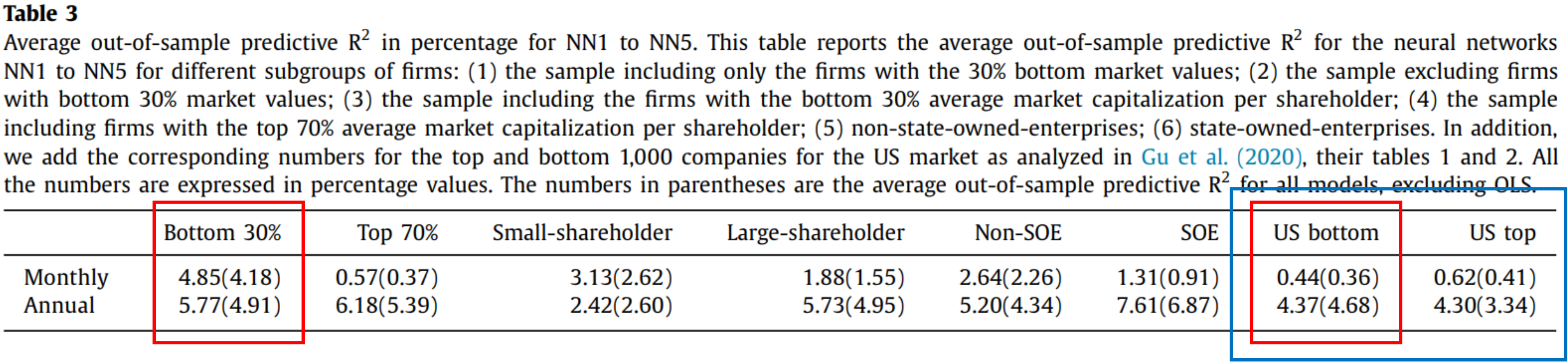


Table 3

Average out-of-sample predictive R^2 in percentage for NN1 to NN5. This table reports the average out-of-sample predictive R^2 for the neural networks NN1 to NN5 for different subgroups of firms: (1) the sample including only the firms with the 30% bottom market values; (2) the sample excluding firms with bottom 30% market values; (3) the sample including the firms with the bottom 30% average market capitalization per shareholder; (4) the sample including firms with the top 70% average market capitalization per shareholder; (5) non-state-owned-enterprises; (6) state-owned-enterprises. In addition, we add the corresponding numbers for the top and bottom 1,000 companies for the US market as analyzed in Gu et al. (2020), their tables 1 and 2. All the numbers are expressed in percentage values. The numbers in parentheses are the average out-of-sample predictive R^2 for all models, excluding OLS.

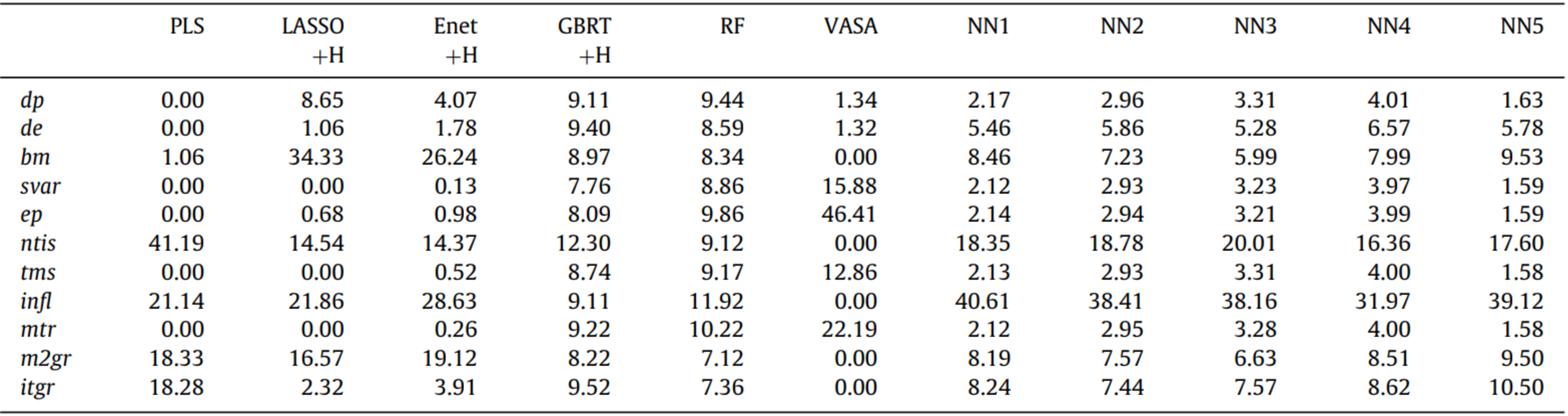
	Bottom 30%	Top 70%	Small-shareholder	Large-shareholder	Non-SOE	SOE	US bottom	US top
Monthly	4.85(4.18)	0.57(0.37)	3.13(2.62)	1.88(1.55)	2.64(2.26)	1.31(0.91)	0.44(0.36)	0.62(0.41)
Annual	5.77(4.91)	6.18(5.39)	2.42(2.60)	5.73(4.95)	5.20(4.34)	7.61(6.87)	4.37(4.68)	4.30(3.34)

- For small Chinese stocks, the OOS R2 is ten times higher than for the US small stocks
- Opposite pattern for Chinese and US market

Which predictor matter?

for a specific model, we calculate the reduction in predictive R_2 when setting all values of a given predictor to zero within each training sample, and average them into a single importance measure for each predictor

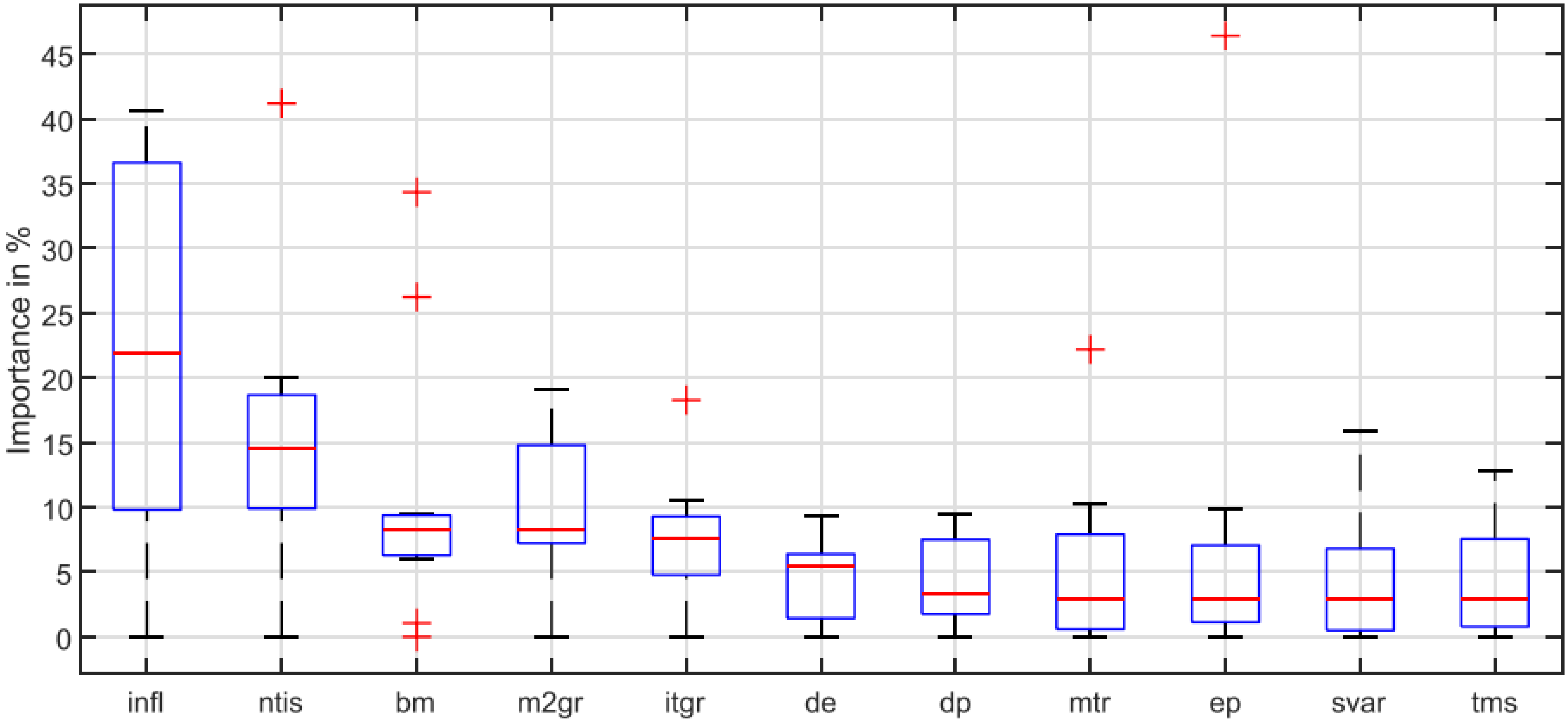
- Macroeconomic variables

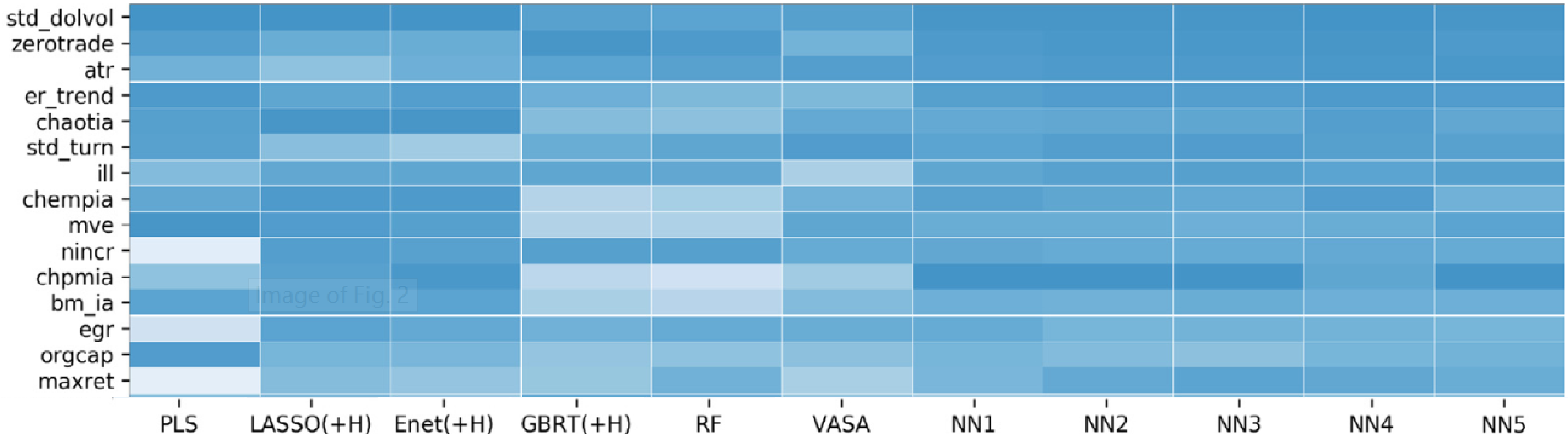


	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
dp	0.00	8.65	4.07	9.11	9.44	1.34	2.17	2.96	3.31	4.01	1.63
de	0.00	1.06	1.78	9.40	8.59	1.32	5.46	5.86	5.28	6.57	5.78
bm	1.06	34.33	26.24	8.97	8.34	0.00	8.46	7.23	5.99	7.99	9.53
svar	0.00	0.00	0.13	7.76	8.86	15.88	2.12	2.93	3.23	3.97	1.59
ep	0.00	0.68	0.98	8.09	9.86	46.41	2.14	2.94	3.21	3.99	1.59
ntis	41.19	14.54	14.37	12.30	9.12	0.00	18.35	18.78	20.01	16.36	17.60
tms	0.00	0.00	0.52	8.74	9.17	12.86	2.13	2.93	3.31	4.00	1.58
infl	21.14	21.86	28.63	9.11	11.92	0.00	40.61	38.41	38.16	31.97	39.12
mtr	0.00	0.00	0.26	9.22	10.22	22.19	2.12	2.95	3.28	4.00	1.58
m2gr	18.33	16.57	19.12	8.22	7.12	0.00	8.19	7.57	6.63	8.51	9.50
itgr	18.28	2.32	3.91	9.52	7.36	0.00	8.24	7.44	7.57	8.62	10.50

- Vary across different models

- Stock characteristics

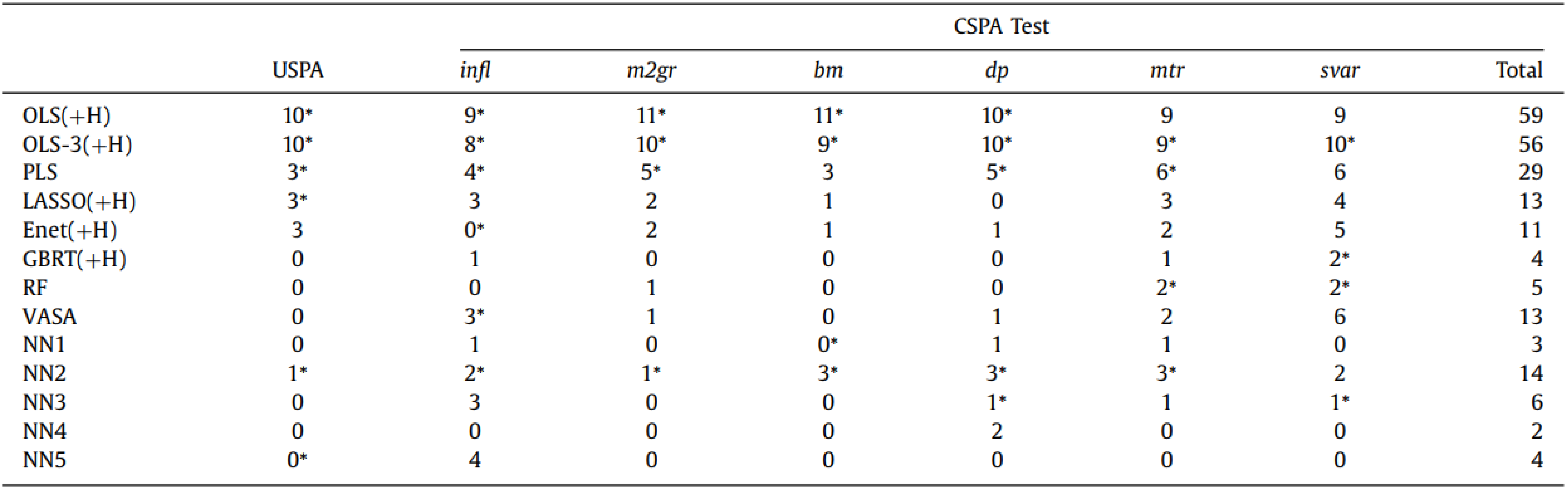




- The most relevant variables: Liquidity / fundamental factors……
  - Gu et al.: trend factors are important, and fundamentals are not.
- Notable differences across models
  - NNs tend to favor momentum and volatility factors over fundamentals
  - NNs, regularized linear models and VASA tend to emphasize a similar set of predictors
- There seem to be a gap in variable importance between the periods before and after 2015 → a structural change in the stock market

• Alternative model selection

- Performance of a given model under the USPA and CSPA tests.



	USPA	CSPA Test						
		infl	*m2gr*	*bm*	*dp*	*mtr*	*svar*	Total
OLS(+H)	10*	9*	11*	11*	10*	9	9	59
OLS-3(+H)	10*	8*	10*	9*	10*	9*	10*	56
PLS	3*	4*	5*	3	5*	6*	6	29
LASSO(+H)	3*	3	2	1	0	3	4	13
Enet(+H)	3	0*	2	1	1	2	5	11
GBRT(+H)	0	1	0	0	0	1	2*	4
RF	0	0	1	0	0	2*	2*	5
VASA	0	3*	1	0	1	2	6	13
NN1	0	1	0	0*	1	1	0	3
NN2	1*	2*	1*	3*	3*	3*	2	14
NN3	0	3	0	0	1*	1	1*	6
NN4	0	0	0	0	2	0	0	2
NN5	0*	4	0	0	0	0	0	4

$$Y_{j,t} \equiv L\left(F_t^{\dagger}, F_{j,t}\right) - L\left(F_t^{\dagger}, F_{0,t}\right).$$

$$H_0^{UEPA} : \mathbb{E}\left[Y_{j,t}\right] = 0, \quad 1 \leq j \leq J,$$

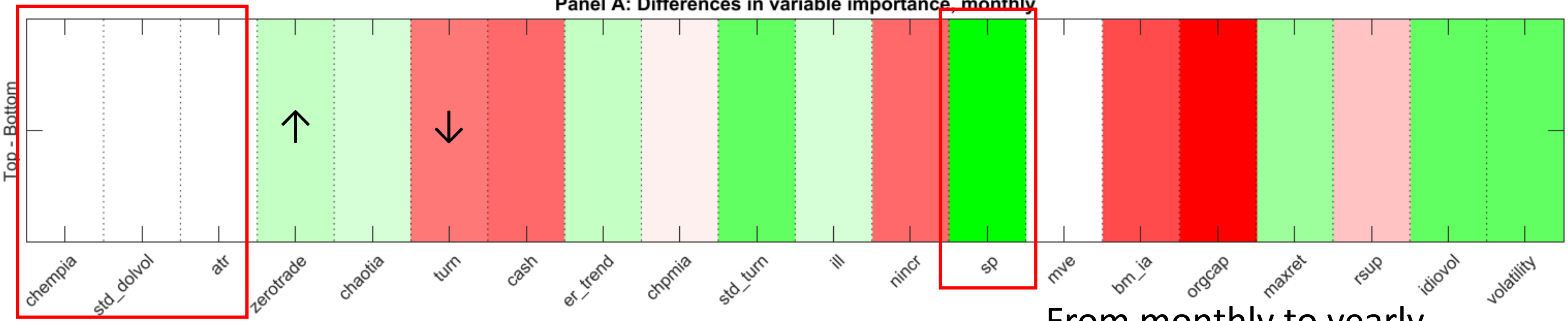
$$H_0^{CEPA} : \mathbb{E}\left[Y_{j,t} | X_t = x\right] = 0, \quad x \in \mathcal{X}, \quad 1 \leq j \leq J,$$

- NN1, NN4, and NN5 have the smallest total number of CSPA test rejections.

- Dissecting the predictability performance of NN4

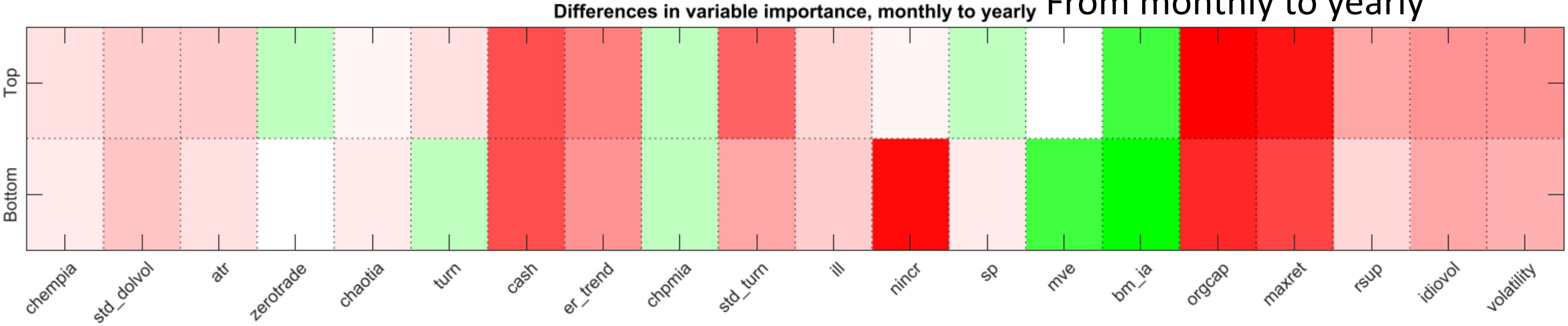
We focus on the striking differences in the monthly and annual OOSR2 for small and large stocks generated by the NN4 model

From large stocks to small



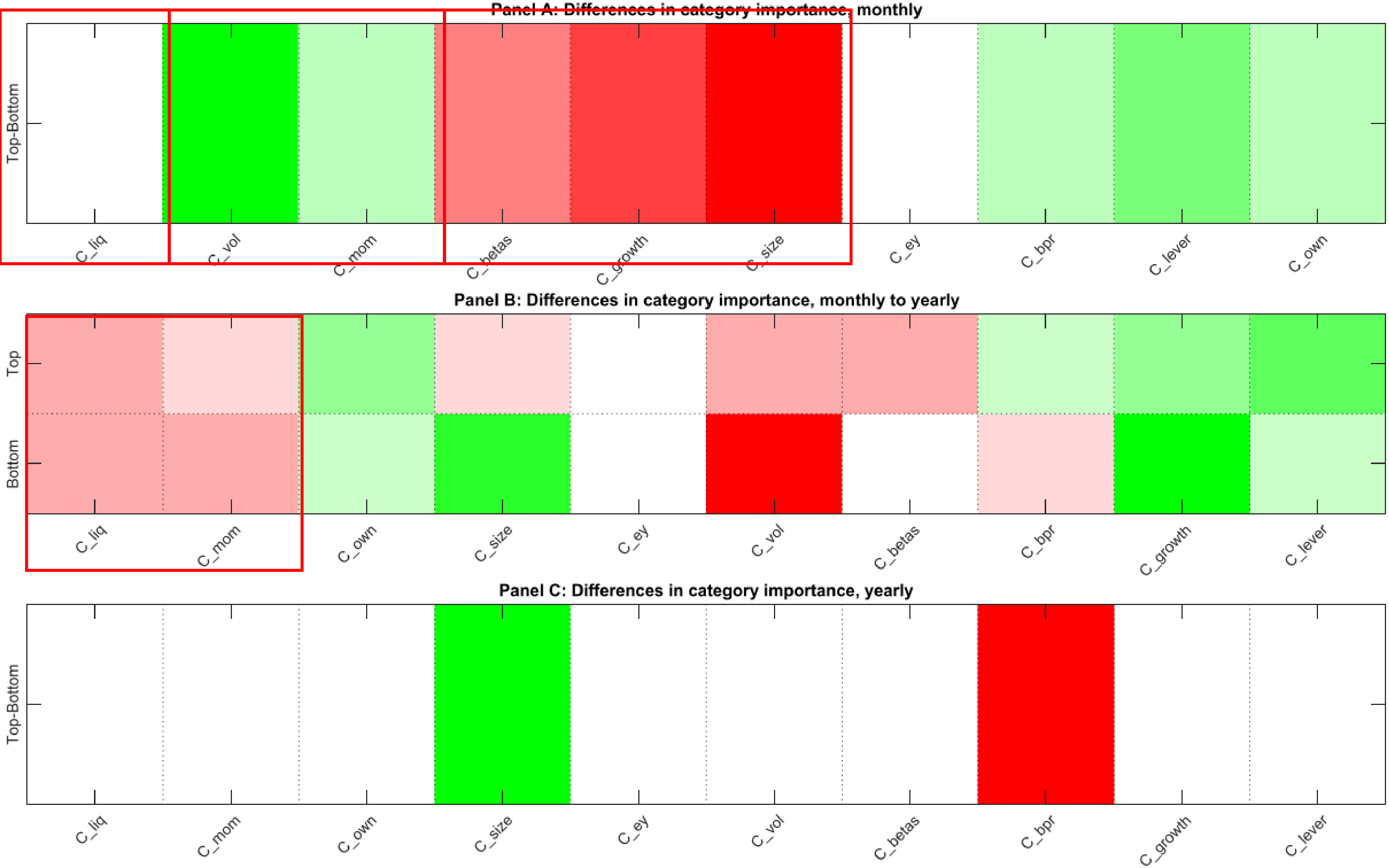
Panel A: Differences in variable importance, monthly

From monthly to yearly



Differences in variable importance, monthly to yearly

- Fundamentals have less impact of smaller stocks, but sales-to-price doesn’t.



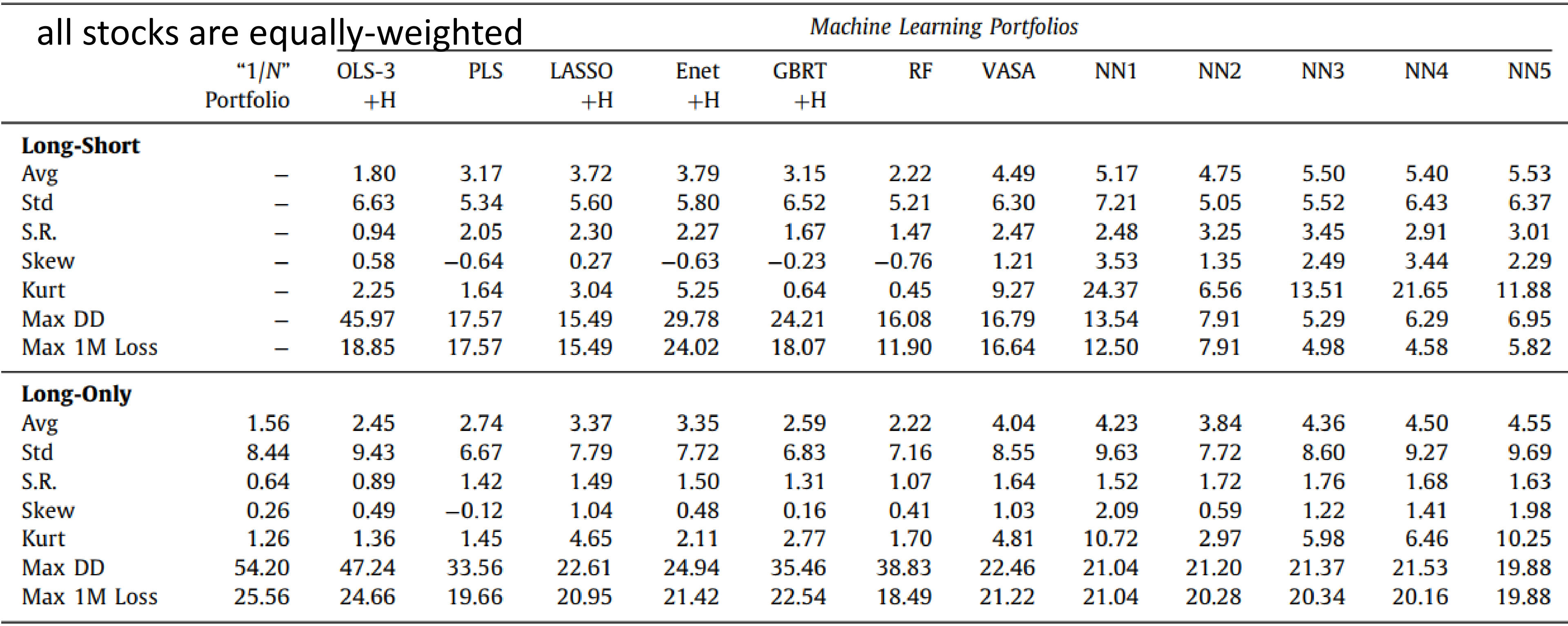
Panel A: Differences in category importance, monthly

Panel B: Differences in category importance, monthly to yearly

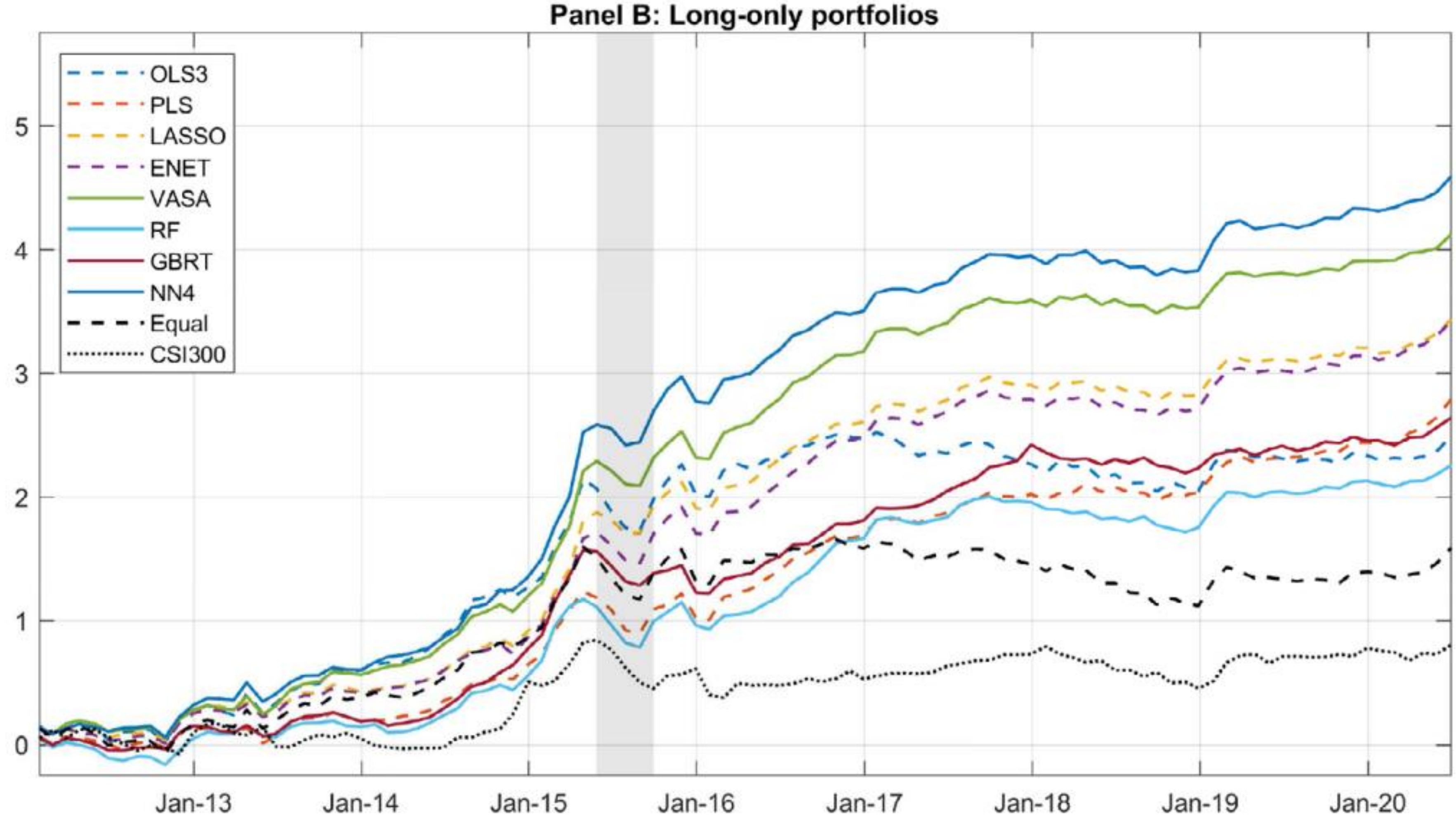
Panel C: Differences in category importance, yearly

- Portfolio analysis

all stocks are equally-weighted

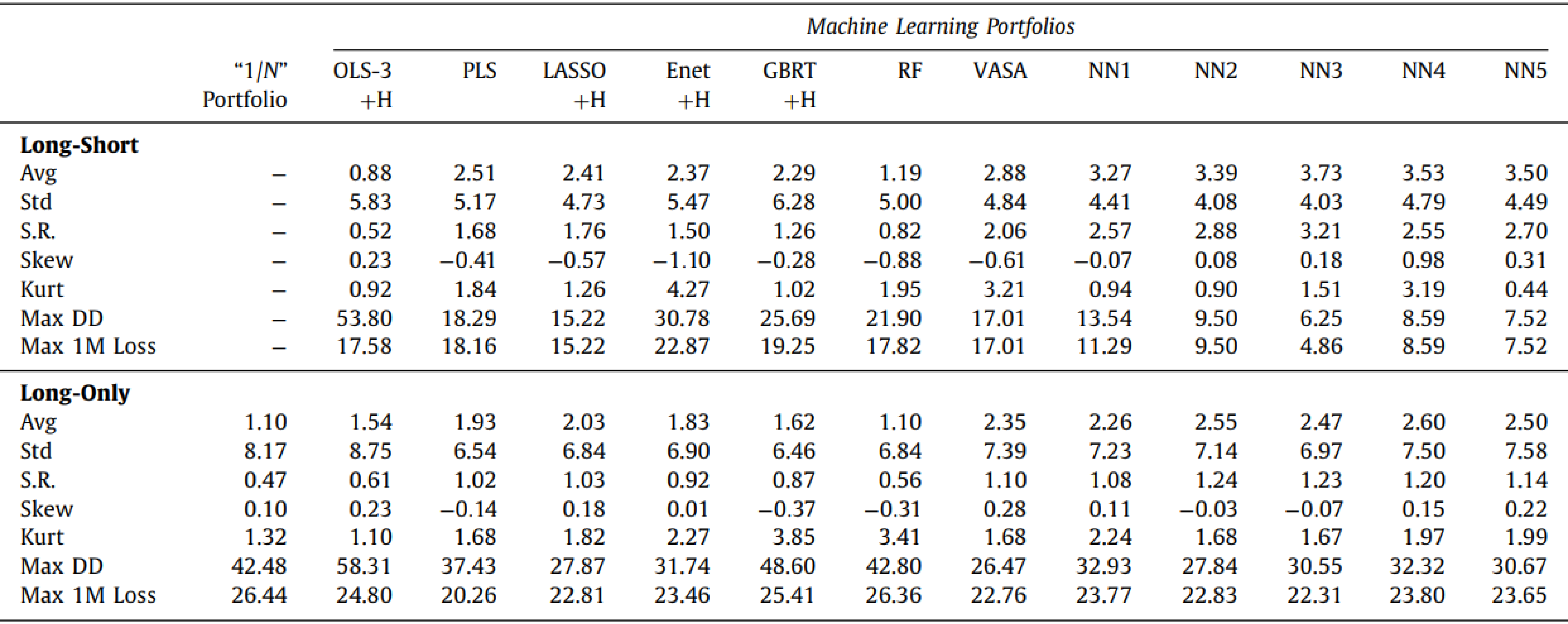


	"1/*N*" Portfolio	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
		Machine Learning Portfolios											
Long-Short													
Avg	–	1.80	3.17	3.72	3.79	3.15	2.22	4.49	5.17	4.75	5.50	5.40	5.53
Std	–	6.63	5.34	5.60	5.80	6.52	5.21	6.30	7.21	5.05	5.52	6.43	6.37
S.R.	–	0.94	2.05	2.30	2.27	1.67	1.47	2.47	2.48	3.25	3.45	2.91	3.01
Skew	–	0.58	−0.64	0.27	−0.63	−0.23	−0.76	1.21	3.53	1.35	2.49	3.44	2.29
Kurt	–	2.25	1.64	3.04	5.25	0.64	0.45	9.27	24.37	6.56	13.51	21.65	11.88
Max DD	–	45.97	17.57	15.49	29.78	24.21	16.08	16.79	13.54	7.91	5.29	6.29	6.95
Max 1M Loss	–	18.85	17.57	15.49	24.02	18.07	11.90	16.64	12.50	7.91	4.98	4.58	5.82
Long-Only													
Avg	1.56	2.45	2.74	3.37	3.35	2.59	2.22	4.04	4.23	3.84	4.36	4.50	4.55
Std	8.44	9.43	6.67	7.79	7.72	6.83	7.16	8.55	9.63	7.72	8.60	9.27	9.69
S.R.	0.64	0.89	1.42	1.49	1.50	1.31	1.07	1.64	1.52	1.72	1.76	1.68	1.63
Skew	0.26	0.49	−0.12	1.04	0.48	0.16	0.41	1.03	2.09	0.59	1.22	1.41	1.98
Kurt	1.26	1.36	1.45	4.65	2.11	2.77	1.70	4.81	10.72	2.97	5.98	6.46	10.25
Max DD	54.20	47.24	33.56	22.61	24.94	35.46	38.83	22.46	21.04	21.20	21.37	21.53	19.88
Max 1M Loss	25.56	24.66	19.66	20.95	21.42	22.54	18.49	21.22	21.04	20.28	20.34	20.16	19.88



- the COVID-19 pandemic in early 2020 does not lead to a notable downturn in portfolio levels

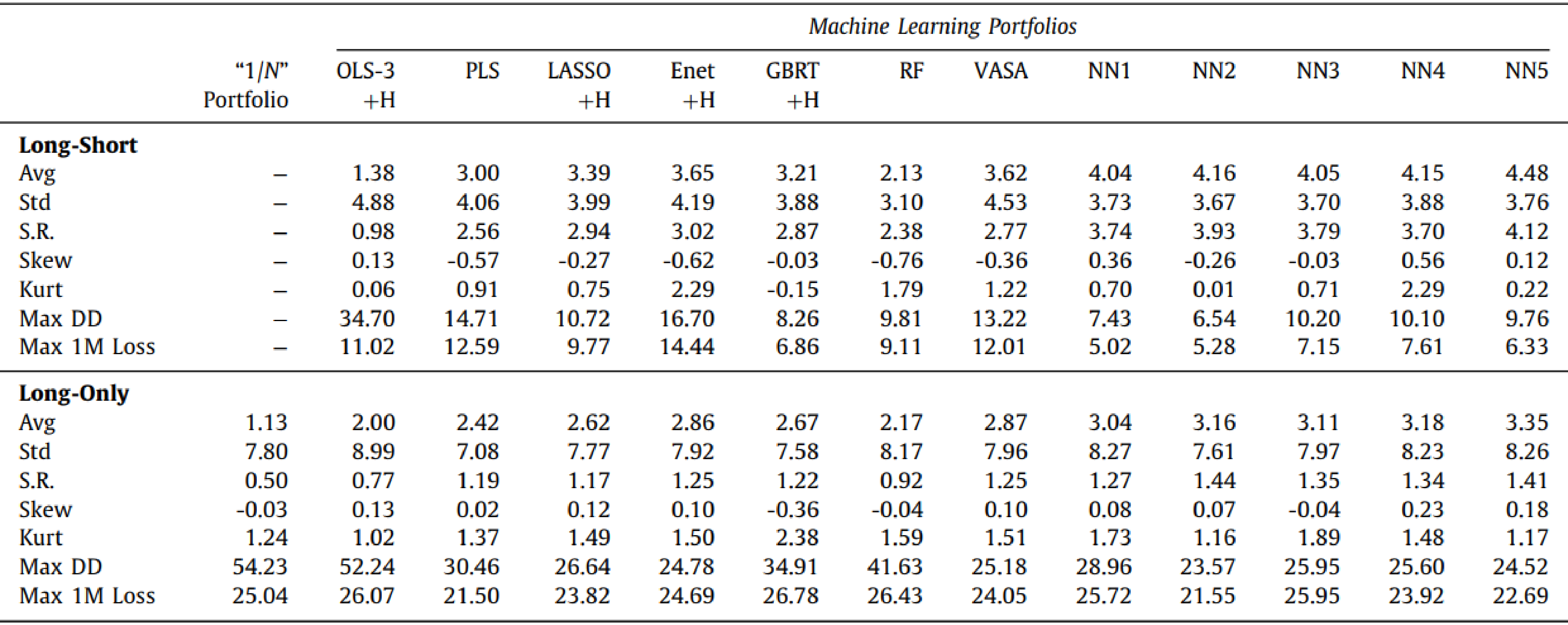
- Performance of machine learning portfolios based on the top 70% sample (value-weighted)



		Machine Learning Portfolios											
	“1/*N*”	OLS-3	PLS	LASSO	Enet	GBRT	RF	VASA	NN1	NN2	NN3	NN4	NN5
	Portfolio	+H		+H	+H	+H							
Long-Short													
Avg	–	0.88	2.51	2.41	2.37	2.29	1.19	2.88	3.27	3.39	3.73	3.53	3.50
Std	–	5.83	5.17	4.73	5.47	6.28	5.00	4.84	4.41	4.08	4.03	4.79	4.49
S.R.	–	0.52	1.68	1.76	1.50	1.26	0.82	2.06	2.57	2.88	3.21	2.55	2.70
Skew	–	0.23	−0.41	−0.57	−1.10	−0.28	−0.88	−0.61	−0.07	0.08	0.18	0.98	0.31
Kurt	–	0.92	1.84	1.26	4.27	1.02	1.95	3.21	0.94	0.90	1.51	3.19	0.44
Max DD	–	53.80	18.29	15.22	30.78	25.69	21.90	17.01	13.54	9.50	6.25	8.59	7.52
Max 1M Loss	–	17.58	18.16	15.22	22.87	19.25	17.82	17.01	11.29	9.50	4.86	8.59	7.52
Long-Only													
Avg	1.10	1.54	1.93	2.03	1.83	1.62	1.10	2.35	2.26	2.55	2.47	2.60	2.50
Std	8.17	8.75	6.54	6.84	6.90	6.46	6.84	7.39	7.23	7.14	6.97	7.50	7.58
S.R.	0.47	0.61	1.02	1.03	0.92	0.87	0.56	1.10	1.08	1.24	1.23	1.20	1.14
Skew	0.10	0.23	−0.14	0.18	0.01	−0.37	−0.31	0.28	0.11	−0.03	−0.07	0.15	0.22
Kurt	1.32	1.10	1.68	1.82	2.27	3.85	3.41	1.68	2.24	1.68	1.67	1.97	1.99
Max DD	42.48	58.31	37.43	27.87	31.74	48.60	42.80	26.47	32.93	27.84	30.55	32.32	30.67
Max 1M Loss	26.44	24.80	20.26	22.81	23.46	25.41	26.36	22.76	23.77	22.83	22.31	23.80	23.65

- All portfolios achieve lower average monthly returns, Sharpe ratios, standard deviations, and extreme negative monthly returns because small stocks are excluded

- Performance of machine learning portfolios based on SOEs (value-weighted)



		Machine Learning Portfolios											
	"1/*N*" Portfolio	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
Long-Short													
Avg	–	1.38	3.00	3.39	3.65	3.21	2.13	3.62	4.04	4.16	4.05	4.15	4.48
Std	–	4.88	4.06	3.99	4.19	3.88	3.10	4.53	3.73	3.67	3.70	3.88	3.76
S.R.	–	0.98	2.56	2.94	3.02	2.87	2.38	2.77	3.74	3.93	3.79	3.70	4.12
Skew	–	0.13	-0.57	-0.27	-0.62	-0.03	-0.76	-0.36	0.36	-0.26	-0.03	0.56	0.12
Kurt	–	0.06	0.91	0.75	2.29	-0.15	1.79	1.22	0.70	0.01	0.71	2.29	0.22
Max DD	–	34.70	14.71	10.72	16.70	8.26	9.81	13.22	7.43	6.54	10.20	10.10	9.76
Max 1M Loss	–	11.02	12.59	9.77	14.44	6.86	9.11	12.01	5.02	5.28	7.15	7.61	6.33
Long-Only													
Avg	1.13	2.00	2.42	2.62	2.86	2.67	2.17	2.87	3.04	3.16	3.11	3.18	3.35
Std	7.80	8.99	7.08	7.77	7.92	7.58	8.17	7.96	8.27	7.61	7.97	8.23	8.26
S.R.	0.50	0.77	1.19	1.17	1.25	1.22	0.92	1.25	1.27	1.44	1.35	1.34	1.41
Skew	-0.03	0.13	0.02	0.12	0.10	-0.36	-0.04	0.10	0.08	0.07	-0.04	0.23	0.18
Kurt	1.24	1.02	1.37	1.49	1.50	2.38	1.59	1.51	1.73	1.16	1.89	1.48	1.17
Max DD	54.23	52.24	30.46	26.64	24.78	34.91	41.63	25.18	28.96	23.57	25.95	25.60	24.52
Max 1M Loss	25.04	26.07	21.50	23.82	24.69	26.78	26.43	24.05	25.72	21.55	25.95	23.92	22.69

- The long-short strategy's performance in terms of the Sharpe ratio is considerably higher for SOEs than for the top 70% stocks, especially for neural networks.

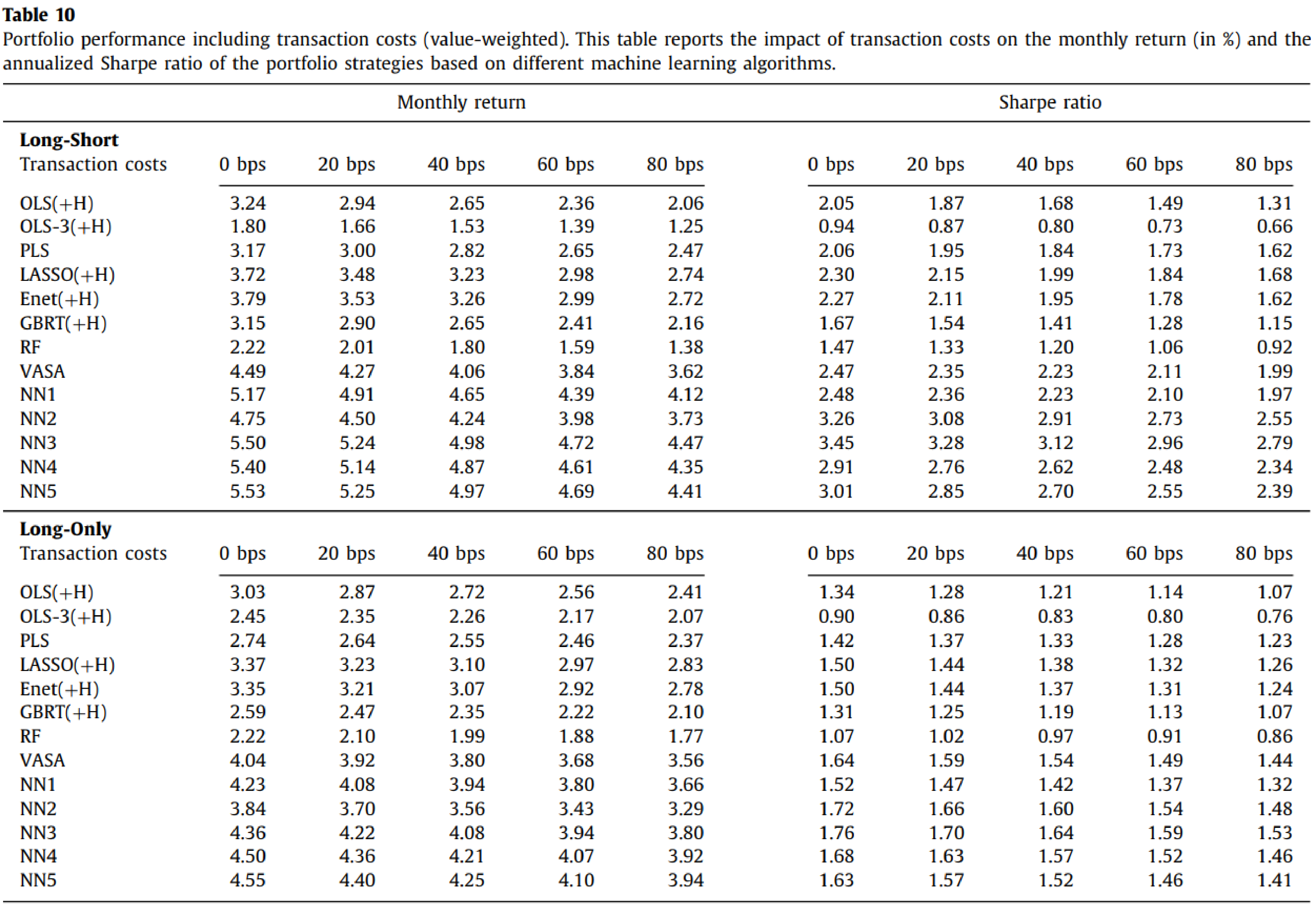


Table 10

Portfolio performance including transaction costs (value-weighted). This table reports the impact of transaction costs on the monthly return (in %) and the annualized Sharpe ratio of the portfolio strategies based on different machine learning algorithms.

	Monthly return					Sharpe ratio				
Long-Short										
Transaction costs	0 bps	20 bps	40 bps	60 bps	80 bps	0 bps	20 bps	40 bps	60 bps	80 bps
OLS(+H)	3.24	2.94	2.65	2.36	2.06	2.05	1.87	1.68	1.49	1.31
OLS-3(+H)	1.80	1.66	1.53	1.39	1.25	0.94	0.87	0.80	0.73	0.66
PLS	3.17	3.00	2.82	2.65	2.47	2.06	1.95	1.84	1.73	1.62
LASSO(+H)	3.72	3.48	3.23	2.98	2.74	2.30	2.15	1.99	1.84	1.68
Enet(+H)	3.79	3.53	3.26	2.99	2.72	2.27	2.11	1.95	1.78	1.62
GBRT(+H)	3.15	2.90	2.65	2.41	2.16	1.67	1.54	1.41	1.28	1.15
RF	2.22	2.01	1.80	1.59	1.38	1.47	1.33	1.20	1.06	0.92
VASA	4.49	4.27	4.06	3.84	3.62	2.47	2.35	2.23	2.11	1.99
NN1	5.17	4.91	4.65	4.39	4.12	2.48	2.36	2.23	2.10	1.97
NN2	4.75	4.50	4.24	3.98	3.73	3.26	3.08	2.91	2.73	2.55
NN3	5.50	5.24	4.98	4.72	4.47	3.45	3.28	3.12	2.96	2.79
NN4	5.40	5.14	4.87	4.61	4.35	2.91	2.76	2.62	2.48	2.34
NN5	5.53	5.25	4.97	4.69	4.41	3.01	2.85	2.70	2.55	2.39
Long-Only										
Transaction costs	0 bps	20 bps	40 bps	60 bps	80 bps	0 bps	20 bps	40 bps	60 bps	80 bps
OLS(+H)	3.03	2.87	2.72	2.56	2.41	1.34	1.28	1.21	1.14	1.07
OLS-3(+H)	2.45	2.35	2.26	2.17	2.07	0.90	0.86	0.83	0.80	0.76
PLS	2.74	2.64	2.55	2.46	2.37	1.42	1.37	1.33	1.28	1.23
LASSO(+H)	3.37	3.23	3.10	2.97	2.83	1.50	1.44	1.38	1.32	1.26
Enet(+H)	3.35	3.21	3.07	2.92	2.78	1.50	1.44	1.37	1.31	1.24
GBRT(+H)	2.59	2.47	2.35	2.22	2.10	1.31	1.25	1.19	1.13	1.07
RF	2.22	2.10	1.99	1.88	1.77	1.07	1.02	0.97	0.91	0.86
VASA	4.04	3.92	3.80	3.68	3.56	1.64	1.59	1.54	1.49	1.44
NN1	4.23	4.08	3.94	3.80	3.66	1.52	1.47	1.42	1.37	1.32
NN2	3.84	3.70	3.56	3.43	3.29	1.72	1.66	1.60	1.54	1.48
NN3	4.36	4.22	4.08	3.94	3.80	1.76	1.70	1.64	1.59	1.53
NN4	4.50	4.36	4.21	4.07	3.92	1.68	1.63	1.57	1.52	1.46
NN5	4.55	4.40	4.25	4.10	3.94	1.63	1.57	1.52	1.46	1.41

Conclusion

- Liquidity-based trading signals and fundamental factors are the most crucial factor categories, while price momentum only play a minor role.
- Short-termism of retail investors generates substantial predictability at short investment horizons, particular for small stocks.
- Substantial increase in SOEs' predictability at longer horizons.
- Portfolio analysis shows the high predictability translates into high Sharpe ratios. And machine learning can be more successful in other markets than the US.

Comparison

- This paper is organized from the perspective of specificity of the Chinese stock market. So they focus on the comparison between the Chinese market and the US market. It’s not the prevailing models, but the analysis and the underlying economic intuition that matter for this paper.
- Tell Chinese stories like a native speaker.