Journal of Financial Economics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Machine learning in the Chinese stock market☆

Markus Leippold[a], Qian Wang[a], Wenyu Zhou[b,c,*]

[a] Department of Banking and Finance, University of Zurich, Plattenstrasse 14, Zurich 8032, Switzerland
[b] International Business School, Zhejiang University, Haining, Zhejiang 314400, China
[c] Academy of Financial Research, Zhejiang University, Hangzhou, Zhejiang 310058, China

ARTICLE INFO

Article history:
Received 7 April 2021
Revised 23 June 2021
Accepted 23 June 2021

JEL classification:
C52
C55
C58
G0
G1
G17

Keywords:
Chinese stock market
Factor investing
Machine learning
Model selection

ABSTRACT

We add to the emerging literature on empirical asset pricing in the Chinese stock market by building and analyzing a comprehensive set of return prediction factors using various machine learning algorithms. Contrasting previous studies for the US market, liquidity emerges as the most important predictor, leading us to closely examine the impact of transaction costs. The retail investors' dominating presence positively affects short-term predictability, particularly for small stocks. Another feature that distinguishes the Chinese market from the US market is the high predictability of large stocks and state-owned enterprises over longer horizons. The out-of-sample performance remains economically significant after transaction costs.

© 2021 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY license
(http://creativecommons.org/licenses/by/4.0/)

1. Introduction

As of October 2020, the total value of China's stock market has climbed to a record high of more than USD 10 trillion (RMB 67 trillion), as the country's accelerating economic recovery from the COVID-19 pandemic has surpassed the previous high reached during an equity bubble in 2015, making it the second-largest in the world, after the US at nearly USD 39 trillion.[1] However, it is not only the size but, equally important, the specificity of the Chinese stock market that makes this market particularly attractive for academic research and allows us to explore questions that contribute to our understanding of emerging markets and complement our knowledge of financial systems in other institutional settings. In particular, we identify at least three key features of the Chinese stock market.

First, unlike developed markets that are dominated by institutional investors, the Chinese stock market is dominated by retail investors. According to the 2019 yearbook of the Shanghai Stock Exchange, there are 214.5 million investors in China; 213.8 million are individual

☆ We thank Bill Schwert (the editor), Xuanjuan Chen, Honghai Yu, the seminar participants at the University of Zurich, the 2021 China Meeting of the Econometric Society, and an anonymous referee for helpful comment. We also thank Zhipeng Liao for sharing the code of the CSPA test and Zhe Wang for excellent research assistance.

* Corresponding author.
E-mail addresses: markus.leippold@bf.uzh.ch (M. Leippold), qian.wang@bf.uzh.ch (Q. Wang), wenyuzhou@intl.zju.edu.cn (W. Zhou).

[1] We adopt the market capitalization indexes from Bloomberg. These indices do not include ETFs and ADRs. They include only actively traded primary securities on the country's exchanges to avoid double counting.

https://doi.org/10.1016/j.jfineco.2021.08.017
0304-405X/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license
(http://creativecommons.org/licenses/by/4.0/)

Please cite this article as: M. Leippold, Q. Wang and W. Zhou, Machine learning in the Chinese stock market, Journal of Financial Economics, https://doi.org/10.1016/j.jfineco.2021.08.017

investors, and 0.7 million are institutional investors. Individual investors hold 99.8% of all accounts holding stocks. The speculative and short-term trading motives of many retail investors may lead to increased turnover. Consequently, the value of shares traded stood at 224% of market capitalization in 2019, compared to 108% for the US market.[2] This peculiarity creates heightened volatility that may disconnect share prices from the underlying economic conditions. Against this background, we ask whether, in such a market, technical indicators emerging from collectivistic investment behavior matter more for asset pricing than firm fundamentals.

Second, as Allen et al. (2005) argue in their seminal paper, a key characteristic of China's financial system from an institutional perspective is that it is centrally controlled, bank-dominated, and uniquely relationship-driven. For example, the process of IPOs and seasonal stock offerings is highly political, and companies cannot predict when the market value will be high. On the other hand, listed companies, especially state-owned enterprises (SOEs), are prevented from shares buy-backs when share prices fall below fundamental values. These automatic market correction mechanisms are therefore affected by government-oriented restrictions (Mei et al., 2009). The SOEs' prominent role in China's capital markets deserve a different treatment for their importance and uniqueness. Not only are they often criticized for the lack of information transparency, but the departure of the SOEs' political objectives from value maximization may harm their corporate performance. See, e.g., Bai et al. (2006), Gan et al. (2018), Jiang and Kim (2020). Therefore, we examine whether return predictability and portfolio performance are compromised for SOEs where government signaling plays such a prominent role.

Third, the Chinese market has a limited history of short sales. Before 2010, Chinese investors faced tight short-selling restrictions. These were partly relieved in March 2010, when the Chinese Security Regulatory Commission allowed a limited number of brokerage firms to short sell 90 stocks on a special list (Gao and Ding, 2019). After short-sale refinancing was officially allowed, the short-selling volume increased exponentially but decreased again after 2015, although the pilot program was expanded to 950 firms at the end of 2016. Although there is no broad consensus, many academics agree that short-selling helps price discovery, rendering markets more efficient (Saffi and Sigurdsson, 2011). While most of the studies on factor investing in US and European markets relies on long-short strategies, such a strategy is less realistic for the Chinese market. Hence, we also analyze long-only portfolios, which are more relevant from a practitioner's viewpoint.

Currently, there is no large database of factor returns available for the Chinese market. Therefore, we contribute to the research on empirical asset pricing in China by building a unique and comprehensive set of factors.[3] In total, we collect 1,160 signals for prediction, consisting of 90 stock-level characteristics, 11 macroeconomic variables, and a set of industry dummies. In a first step, we construct a set of factors in the same way as has been constructed for the US market. In a second step, we follow previous studies by adapting some of these US factors for the Chinese stock market. In a third step, we also include a set of China-specific factors. For instance, we add the abnormal turnover ratio (*atr*), introduced by Pan et al. (2015). The *atr* is designed to capture the impact of speculative trading in the stock market, which helps explain the Chinese A-shares' overpricing.

Given that China has been experiencing a highly dynamic development through a series of structural breaks, implementing various financial reforms, and expanding its capital markets' openness, we conjecture that highly flexible methods are required to account for the Chinese market's specificity. Therefore, we rely on different machine learning techniques for our analysis, whose application to finance and economics is rapidly emerging and has witnessed an explosion of research contributions, with encouraging results. A rapidly growing number of studies examine the cross-section and the time-series of stock returns with machine learning tools, predominantly focusing on the U.S. market.

In this study, we build on the work of Gu et al. (2020) who combine a broad repertoire of machine learning methods with modern empirical asset pricing research to understand the dynamics of market risk premia for stock returns.[4] Their results suggest that machine learning improves the description of expected return and, when applied to portfolio construction, performance improvements arise most prominently among the more sophisticated models and are due in large part to the allowance of non-linear predictor interactions that are missed by simpler methods. It is unclear whether these results also hold for the Chinese stock market. However, given its characteristics mentioned above, especially the large proportion of small investors with speculative short-term behavior, this market makes a highly attractive target for the application machine learning techniques.

Exploring the different machine learning methods' predictive ability, we find that neural networks robustly outperform other methods in terms of out-of-sample R^2. The out-of-sample R^2 are particularly large for the subsamples of small firms and non-state-owned firms. Hence, predictability is more significant for those subsamples of stocks in which retail traders play a much bigger role. Moreover, comparing the out-of-sample R^2 with studies in the US market, the Chinese market reveals substantially more predictability. As the out-of-sample R^2 has some

[2] See, World Development Indicators (2020). According to the 2018 yearbook of the Shanghai Stock Exchange, retail investors generated a turnover of 82% and a profit of 311 billion yuan (USD 47 billion in annual average exchange rate). At the same time, institutional investors generated a profit of 1,116 billion yuan (USD 168.6 billion in annual average exchange rate).

[3] The data can be obtained from the authors upon request.

[4] Their dataset includes 94 characteristics for each stock, each characteristic's interactions with eight aggregate time-series variables, and 74 industry sector dummy variables, totaling more than 900 baseline signals for prediction. Recently, numerous additional refinements of the basic algorithms surveyed in Gu et al. (2020) have been suggested. Examples include Bryzgalova et al. (2019), Chen et al. (2019a), Feng et al. (2019), De Nard et al. (2020), Gu et al. (2021).

limitations for model selection, we analyze the models' conditional predictive ability using a statistical test developed in Li et al. (2020), which allows us to compare the performance of machine learning methods in different macroeconomic environments. Again, the neural networks prove robust to this new statistical test and emerge as the best-performing method in terms of predictability.

In our empirical analysis, we make the following observations. The most relevant variables across all prediction models are stock characteristics that relate to market liquidity. The second important group of predictors, however, relate to fundamental factors like valuation ratios. This finding is in contrast to Gu et al. (2020)'s previous study for the US market, where classical trend indicators are the main drivers of predictability. However, we find notable differences across models. In particular, in addition to liquidity, neural networks tend to favor momentum and volatility factors over fundamentals. We also find that the predictability of SOEs in terms of out-of-sample predictive R^2 is weaker than for non-SOEs at a monthly prediction horizon, which confirms the SOE's reputation of being non-transparent (Piotroski et al., 2015). Lastly, given the short-selling constraints in China, we wonder how much value-added can be enjoyed in long-only mandates. Many of the results in previous studies relate to the performance of portfolios that include long and short positions. While such practices allow us to evaluate a signal's predictive power, not all stocks are available for shorting at all times, and the costs of shorting can be substantial. This is even more true for the Chinese market. Our results also indicate that a long-only portfolio can provide substantial and, even after including transaction costs, economically significant performance. Moreover, this strategy also performs well during the 2015 crash and remains unaffected by the COVID-19 pandemic in early 2020.

The remainder of the paper is structured as follows. In Section 2, we provide a description of our data and the methodologies used for prediction. We present our empirical analysis in Section 3. We look at the out-of-sample predictability, and discuss which predictors matter most. We also perform a model selection analysis using both the unconditional and conditional predictive ability tests. In Section 4, we explore whether predictability translates into portfolio gains. We conclude in Section 5. Detailed discussions of the methods used and additional results are in the Internet Appendix.

2. Data and methodology

For our analysis, we apply the empirical design of Gu et al. (2020) to the Chinese market. To this end, we obtain daily and monthly total stock returns for all A-share stocks listed on the Shanghai and Shenzhen stock exchanges from the Wind Database, the largest financial data provider in China. The corresponding quarterly financial statement data are downloaded from the China Stock Market and Accounting Research (CSMAR) database. Our data sample covers more than 3,900 A-share stocks traded from January 2000 to June 2020. Also, we obtain the yield rate for the one-year government bond in China from CSMAR to proxy for the risk-free rate, which is necessary for calculating individual excess returns.

With these data at hand, we build a large collection of stock-level predictive characteristics based on the variable definitions in the original papers listed in Green et al. (2017), and the papers on China-specific factors. Our collection includes 94 characteristics in total, among which 86 have been documented in Green et al. (2017), four are valid China-specific factors identified in previous studies, and four are binary variables that indicate ownership types for listed firms and are used for subsample analysis. To avoid outliers, we cross-sectionally rank all continuous stock-level characteristics period-by-period, and map them into the $[-1, 1]$ interval following Kelly et al. (2019) and Gu et al. (2020). In terms of data frequency, 22 stock-level characteristics are updated monthly, 51 are updated quarterly, six are updated semi-annually, and 15 are updated annually. It is noteworthy that our data frequency is higher than that in Gu et al. (2020), which may improve our prediction performance. Also, we include 80 industry dummies based on the Guidelines for Industry Classification of Listed Companies issued by the China Securities Regulatory Commission (CSRC) in 2012. Table C.1. in the Internet Appendix provides a summary of all stock-level characteristics.

In addition to the above characteristics, we also construct 11 macroeconomic predictors based on the data downloaded from CSMAR and the National Bureau of Statistics websites. Eight of those variables are based on the variable definitions in Welch (2008), including dividend price ratio (*dp*), dividend payout ratio (*de*), earnings price ratio (*ep*), book-to-market ratio (*bm*), net equity expansion (*nits*), stock variance (*svar*), term spread (*tms*), and inflation (*infl*). The remaining three include monthly turnover (*mtr*), M2 growth rate (*m2gr*), and international trade volume growth rate (*itgr*), which are identified in previous studies as effective macroeconomic predictors. In Table C.5 in the Internet Appendix, we summarize these macroeconomic variables.

Throughout our analysis, we adopt a general additive prediction error model to describe the relation between a stock's excess return and its corresponding predictors, i.e.,

$$r_{i,t+1} = \mathbb{E}_t[r_{i,t+1}] + \epsilon_{i,t+1}. \tag{1}$$

In addition, we further assume the conditional expectation of stock i's excess return $r_{i,t+1}$ given the information available at period t to be a constant function of a set of predictors:

$$\mathbb{E}_t[r_{i,t+1}] = g(z_{i,t}), \tag{2}$$

where $z_{i,t}$ is a P-dimensional vector of predictors, stocks are indexed by $i = 1, ..., N_t$, and months by $t = 1, ..., T$. The functional form of $g(\cdot)$ is left unspecified. Our target is to search for the prediction model from a set of candidates that gives the best prediction performance.

The vector of predictors, $z_{i,t}$, consists of stock i's characteristics, the interaction terms between stock-level characteristics and the 11 macroeconomic predictors, and

a set of dummy variables, which can be represented as:

$$z_{i,t} = \begin{pmatrix} c_{i,t} \\ x_t \otimes c_{i,t} \\ d_{i,t} \end{pmatrix}, \tag{3}$$

where $c_{i,t}$ is a 90×1 vector of stock-level characteristics, x_t is a 11×1 vector of macroeconomic predictors, $d_{i,t}$ is a 80×1 vector of dummy variables, and $\otimes$ denotes the Kronecker product. The set of dummy variables include the 80 industry dummies. Hence, the total number of covariates in $z_{i,t}$ is $90 \times (11 + 1) + 80 = 1{,}160$.

In total, we consider 11 machine learning methods, along with two simple linear models. In particular, we include ordinary least squares (OLS) regression, OLS using only size, book-to-market, and momentum as predictors (OLS-3), partial least squares (PLS), least absolute shrinkage and selection operator (LASSO), elastic net (Enet), gradient boosted regression trees (GBRT), random forest (RF), variable subsample aggregation (VASA), and neural networks with one to five layers (NN1-NN5). Similar to Gu et al. (2020), we only focus on OLS, OLS-3, LASSO, Enet, and GBRT equipped with a Huber loss function to avoid potential disturbance caused by extreme values in the data (Huber, 2004).

We follow the standard approach in the literature for hyperparameters selection, model estimation, and performance evaluation. In particular, we divide our data into three disjoint periods while maintaining the temporal ordering: the training sample (2000–2008), the validation sample (2009-2011), and the testing sample (2012–2020). We use the training sample to estimate the model parameters subject to some pre-specified hyperparameters for a specific machine learning model. The validation sample is used to optimize the hyperparameters of our models. We select the hyperparameters that minimize the objective loss function based on the observations in the validation sample. The testing sample contains the next 12 months of data right after the validation sample. These data, which never enter into model estimation or tuning, are used to test our models' prediction performance. Since machine learning models are computationally intensive, we adopt a sample splitting scheme as in Gu et al. (2020) by refitting prediction models annually instead of monthly. When we refit a model, we increase the training sample size by one year but maintain the same size for the validation sample. Meanwhile, both the validation sample and the one-year testing period are kept rolling forward to include the next twelve months. Table A.2 in the Internet Appendix provides further details on hyperparameters training and prediction models.

3. Empirical analysis

We start by exploring our models' prediction performance via out-of-sample predictive R^2 and discuss predictability across different subsamples.

3.1. Out-of-sample predictability

As in Gu et al. (2020), we rely on the non-demeaned out-of-sample predictive R^2 to have a direct comparison with their results for the US market. For a given model S, this measure is defined as:

$$R^2_{\text{oos},S} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}} (r_{i,t} - \hat{r}^{(S)}_{i,t})^2}{\sum_{(i,t)\in\mathcal{T}} r^2_{i,t}}, \tag{4}$$

where $\mathcal{T}$ denotes the set of predictions that are only assessed on the testing sample, and $\{\hat{r}_{i,t}\}_{(i,t)\in\mathcal{T}}$ are predicted monthly returns. As state-owned enterprises (SOEs) play an prominent in China's capital markets and are often criticized for information transparency, we explore the R^2_{oos} for both SOEs and non-SOEs. As Liu et al. (2019) argue, the smallest 30% of firms often serve as potential shells in reverse mergers that circumvent tight IPO constraints. At the same time, Chinese retail investors have a notorious preference for investing in small stocks, in particular growth and glamour stocks (Ng and Wu, 2006). Therefore, to address potential behavioral stories, we also build two subsamples according to firm size with a 30% cutoff level. The results for the different models and subsamples are summarized in Table 1.

3.1.1. Full sample analysis

When we include all companies, the OLS model achieves a positive R^2_{oos} of 0.81%, showing even the simplest model still has some predictive power. The R^2_{oos} for the OLS-3 model is slightly lower than that for the OLS model (0.77% v.s. 0.81%), indicating the three covariates alone (size, book-to-market, and momentum) are insufficient to account for all predictive power in linear models. It is noteworthy that the OLS model performs much better in China's stock market than in the US stock market. The R^2_{oos} for the latter is negative (−3.46%) in Gu et al. (2020). A possible explanation for such difference is that we set a relatively small value for the Huber loss function's tuning parameter, which leads to a high level of robustness to extreme values in the data.[5]

For regularized models including PLS, LASSO, and Enet, the improvement of the R^2_{oos} directly reflects the effectiveness of dimension reduction when we are faced with a large set of covariates. All three models raise the out-of-sample R^2 to above 1%, with LASSO (1.43%) and Enet (1.42%) having a small advantage over PLS (1.28%). This improvement of R^2_{oos} thus suggests that some stock characteristics are redundant for predicting monthly returns in China's stock market, which resonates well with the findings in Gu et al. (2020) for the US market. The R^2_{oos} for VASA is comparable to those of regularized linear models. This observation is most likely because we use VASA with linear submodels, which shares many similarities with PLS regarding forming a linear combination of predictors.

The tree models, GBRT and RF, and five neural network models improve R^2_{oos} even further to above 2% in all seven models. Such improvement demonstrates the superiority of machine learning methods in capturing complex interactions between predictors, which is emphasized for the US stock market in Gu et al. (2020). The full-sample

[5] In our study, we set the tuning parameter to $M = 1.35$, following the suggestion in Huber (2004), which can produce as much robustness as possible while remaining efficient for normally distributed data.

Table 1
Monthly out-of-sample predictive R^2 in percentage. This table reports monthly out-of-sample predictive R^2 of forecast models for different subgroups of firms: (1) the full sample; (2) the sample excluding firms with bottom 30% market values; (3) the sample including only the firms with the 30% bottom market values; (4) the sample including firms with top 70% average market capitalization per shareholder; (5) the sample including only the firms with the bottom 30% average market capitalization per shareholder; (6) state-owned-enterprises; and (7) non-state-owned-enterprises. The models considered include ordinary least squares (OLS) regression, OLS using only size, book-to-market and momentum (OLS-3), partial least squares regression (PLS), least absolute shrinkage and selection operator (LASSO), elastic net (Enet), gradient boosted regression trees (GBRT), random forest (RF), variable subsampling aggregation (VASA), and neural networks with 1 to 5 layers (NN1-NN5). "+H" indicates that the model is trained using Huber loss instead of l_2 loss. SOE and Non-SOE represent the subgroups of state-owned and non-state-owned enterprises, respectively. All the numbers are expressed as a percentage.

	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

R^2_{oos} suggests that both GBRT and RF are competitive with neural networks. Unlike the US stock market, we observe an increase in the R^2_{oos} when increasing hidden layers in neural networks, although such improvement seems to be marginal for models with more than four layers.

In addition, in terms of monthly R^2_{oos}, machine learning techniques reveal much stronger predictability in the Chinese market than in the US market. The highest R^2_{oos} in the Chinese market, produced by our GBRT (2.71%), is almost sevenfold of the highest R^2_{oos} reported in Gu et al. (2020) generated by their NN4 (0.40%). In addition, even the lowest R^2_{oos}, produced by OLS-3 based on all Chinese stocks (0.77%), is nearly double the highest R^2_{oos} in the US market.

Such significant gaps in R^2_{oos} further motivates us to consider the fundamental difference between these two markets, which we conjecture, can be attributed to two critical aspects. First, the Chinese stock market is characterized by a large fraction of retail investors and their preference for small-cap stocks. Second, the Chinese stock market is influenced by the prevalence of SOEs, which are less transparent than private firms. We next explore these two channels separately.

3.1.2. Small and large stocks

To investigate the potential heterogeneity in model predictability, we conduct subgroup analysis for small (the bottom 30% stocks by market equity each month) and large (the top 70% stocks each month) stocks. Table 1 reports the R^2_{oos} for the largest 70% stocks and smallest 30% stocks by monthly market equity. The results in Table 1 suggest that all models have a much better predictive performance for small stocks. The linear models, OLS and OLS-3, now raise their R^2_{oos} to above 1%, while the regularized linear models, including PLS, LASSO, and Enet, nearly double their performance.

The tree-based models and neural networks still keep an advantage over regression-based methods. GBRT seems to be especially successful, with the highest R^2_{oos} of 7.27%. While predictability improves drastically for the 30% smallest stocks, the predictability for the 70% largest stocks deteriorates. The out-of-sample R^2s reduce to below 1% for all models. Interestingly, OLS, RF, and even GBRT, now have negative R^2_{oos}, indicating they are easily dominated by a naïve forecast of zero returns for all stocks in all periods. However, the neural networks still show stable performance, except for some on par with regularized linear models (PLS and LASSO).

3.1.3. Small and large shareholders

The above results indicate that machine learning methods can strongly predict the monthly returns of small stocks. However, it is still unclear whether retail investors play an important role in generating such a difference. To provide insight on the connection between predictability and retail investors, we conduct subgroup analysis based on the average market capitalization per shareholder. We collect numbers of shareholders of outstanding A-shares for all listed companies from CSMAR, which are reported quarterly, and the corresponding market capitalization. Then, we calculate the average market capitalization per shareholder, i.e., A.M.C.P.S. = Market Cap/Number of Shareholders, and classify all stocks into two groups based on the top 70% threshold.[6] And last, we investigate model predictability by looking into the out-of-sample R^2 for these two groups.

The fourth and fifth rows in Table 1 report the R^2_{oos} for firms with the top 70% and the bottom 30% average market cap per shareholder, respectively. Overall, these results show that machine learning methods, especially PLS, random forests, and neural networks, have better predictive performance in the sample of stocks with small shareholders, as their R^2_{oos} are substantially larger for stocks with small shareholders than large shareholders. At the same time, LASSO, Enet, and VASA perform similarly on both subsamples. Interestingly, OLS-3 generates much worse predictions in the sample of small-shareholder stocks than large-shareholder stocks, which implies that the conventional three-factor model might not work well for small-shareholder stocks in China. In brief, even though

[6] The main results in this subsection are not sensitive to the choice of classification threshold. In addition to the 0.7 quantile, we also investigate the 0.9, 0.8, and 0.6 quantiles, which generate the same pattern of model predictability. These results are not presented for the sake of simplicity but are available upon request.

it is infeasible to accurately identify the prevalence of retail investors for every stock due to the lack of data, we believe the average market capitalization per shareholder could still be a useful proxy, which helps to unveil the relation between model predictability and the role of retail investors.

3.1.4. SOEs and non-SOEs

When we focus on the stock returns of SOEs and non-SOEs, Table 1 suggests that neural networks produce robust and positive R^2_{oos} for both subsamples.[7] For tree-based models, the results are mixed. While they perform exceptionally well for non-SOE stocks, they fail to outperform regression-based models for SOE stocks. Overall, the pattern of R^2_{oos} for SOE and non-SOE stocks resembles the one from our analysis of 30% smallest and 70% largest companies. This similarity arises, in part, from the fact that SOEs in China tend to have a large market capitalization, as they usually represent the dominant companies in fundamental industries like banking, infrastructure, and military. Therefore, company size is strongly correlated with the notion of SOE and non-SOE stocks.

Nevertheless, comparing the level of predictability, we see that, when using neural networks, SOEs provide a much larger R^2_{oos} than the top 70% companies. For the former subgroup, the average R^2_{oos} for models NN1 to NN5 is 1.31, while for the latter, it is only 0.57. What also strikes us is that, for SOEs, neural networks are consistently better than all other models. For all other subgroups, we always find some models that are performing comparably with neural networks. This observation underlines the uniqueness of SOEs again. It seems that predicting SOEs' returns requires a highly flexible method that can account for nonlinear effects. This additional complexity may be required since SOEs are controlled by the state, having two primary objectives: to generate profit and to carry out state policies. However, our results contrast with earlier studies that argue that predicting stock returns for Chinese SOEs is not easy due to their financial opacity and low informativeness of share prices (e.g., Lee and Wang (2017)).

Based on the above subsample analysis, we conclude that machine learning techniques, especially tree models and neural networks, perform satisfactorily in the Chinese stock market in terms of out-of-sample R^2. Moreover, our analysis unveils two important Chinese stock market features that differ from the US market studied in Gu et al. (2020). First, monthly returns of small (non-SOE) stocks in the Chinese market can be much better predicted than large (SOE) stocks for almost all models. Second, neural networks can provide robust performance (in terms of R^2_{oos}) across different subsamples.

3.1.5. Predictability at annual horizon

Next, we investigate the prediction performance of our models at the annual horizon. Table 2 reports the annual out-of-sample predictive R^2 for different models and subsamples. We find that the annual out-of-sample R^2s are higher than their monthly counterparts, indicating machine learning methods can successfully isolate persistent risk premiums at longer horizons. Interestingly, with the given methods, we now obtain a better prediction performance for the largest 70% stocks than for the smallest 30% stocks. The improved predictability of larger stocks could be caused by the improved predictability of SOEs. According to Jiang and Kim (2020), SOEs currently account for roughly one-third of firm numbers but two-thirds of market capitalization. In addition, the same pattern also appears in subgroups with different levels of average market cap per shareholder, as all methods generate better predictions in the subsample of large-shareholder stock than in the sample of small-shareholder stock.

Our finding contrasts our previous observation made on a monthly level, where the small stocks, small-shareholder stocks, and the non-SOE firms exhibit considerably stronger predictability than their counterparts. The differences in predictability on an annual horizon are not as large and seem to level out, but they indicate some advantage for large firms, stocks with larger shareholders, and SOEs. We attribute the short-term predictability, particularly for small stocks, to retail investors' prominent role in the Chinese stock market. As shown in Section 3.4, neural networks put more weight on volatility and momentum-related variables for small stocks, which may reflect the short-term speculative behavior of retail investors, together with their well-known preference for trading small stocks.

In Table 3, we compare the average monthly and annual out-of-sample predictive R^2 for different subsamples, and we compare our results with those of Gu et al. (2020) for the US market. For firms with the top 70% market values, we find comparable predictability at the monthly level, as is the case for the top 1,000 companies in the US market. Simultaneously, the out-of-sample R^2 for SOEs, which are usually large stocks, is more than double the value for large US stocks. Strikingly, for small Chinese stocks, we observe an out-of-sample R^2 that is ten times higher than for the US small stocks. For US stocks, predictability seems to improve more for small stocks than for large stocks when moving from a monthly to an annual time horizon. The opposite is true for the Chinese market. Predictability for large stocks, stocks with larger stockholders, and SOEs, in particular, is much better than for small stocks, stocks with small stockholders, and non-SOEs. These observations reveal some striking differences between the Chinese market and the US market, which we suspect are mainly due to retail investors' dominant effect on the short horizon and government initiatives, which can predominantly benefit SOEs.

In the Internet Appendix D, we explore the time variations in the out-of-sample R^2_{oos} of our models. For most models, we observe in Fig. D.1 a significant drop in R^2_{oos} in 2018. We conjecture that the cause of this drop lies in the Chinese stock market's persistent fall caused by the severe trade conflicts between China and the US, pointing out a potential weakness for machine learning techniques when predicting stock returns: their performances can be vulnerable to unexpected systematic risk, such as, in this

[7] As our testing sample spans from 2012 to 2020, we report the fraction of SOEs year by year during this period. The fractions of SOEs are 40.62%, 39.95%, 38.79%, 37.03%, 34.88%, 31.53%, 30.19%, 29.59%, and 28.59% during the 2012–2020 period, respectively.

Table 2
Annual out-of-sample predictive R^2 in percentage. This table reports annual out-of-sample predictive R^2 of forecast models for different subgroups of firms: (1) the full sample; (2) the sample excluding firms with bottom 30% market values; (3) the sample including only the firms with the 30% bottom market values; (4) the sample including firms with top 70% average market capitalization per shareholder; (5) the sample including only the firms with the bottom 30% average market capitalization per shareholder; (6) state-owned-enterprises; and (7) non-state-owned-enterprises. The models considered include ordinary least squares (OLS) regression, OLS using only size, book-to-market and momentum (OLS-3), partial least squares regression (PLS), least absolute shrinkage and selection operator (LASSO), elastic net (Enet), gradient boosted regression trees (GBRT), random forest (RF), variable subsampling aggregation (VASA), and neural networks with 1 to 5 layers (NN1-NN5). "+H" indicates that the model is trained using Huber loss instead of l_2 loss. SOE and Non-SOE represent the subgroups of state-owned and non-state-owned enterprises, respectively. All the numbers are expressed as a percentage.

	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

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)

case, the political risk related to a trade war between the US and China.

3.2. Which predictors matter?

Given the large number of predictors, we next investigate whether certain predictors are more important than others. To this end, we differentiate between the macroeconomic variables and the stock characteristics.

3.2.1. Macroeconomic variables

We first explore the variable importance of 11 macroeconomic variables and 94 stock characteristics for all prediction models based on the Chinese stock market. The variable importance is defined similarly as in Gu et al. (2020), i.e., 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.

Table 4 reports the relative variable importance of our 11 macroeconomic variables. For PLS, *ntis*, which measures the level of issuance activity, has the largest variable importance. China has been adopting an approval-based IPO system ever since its stock market opened, and it is well-known that the China Securities Regulatory Commission often suspends or reduces the volume of IPOs when the market is down, making it reasonable for *ntis* to play an important role in predicting monthly returns. It is worth noting that *ntis* is also the most important macroeconomic variable for GBRT and the second important variable for neural networks. Moreover, PLS also puts substantial weight on *infl*, *m2gr*, and *itgr*, showing these macroeconomic variables are also influential.

The results in Table 4 suggest that penalized linear models, including LASSO and Enet, strongly favor the aggregate book-to-market ratio (*bm*), which is, however, less important for PLS and VASA. In addition, variables like *infl*, *ntis*, and *m2gr* also have high priority in LASSO and Enet. Differing from other models, VASA favors the aggregate earnings price ratio (*ep*), as well as variables that reflect market liquidity (*mtr*) and volatility (*svar*). The distribution of macroeconomic variable importance for tree models GBRT and RF is relatively more uniform than other regression-based methods, indicating that these two methods can detect potentially complicated nonlinear interactions between macroeconomic variables and stock characteristics.

In Fig. 1, we aggregate the variable importance across models for each of the macroeconomic variables. Overall, we find that *infl* and *ntis* are the two most influential macroeconomic variables for predicting monthly returns in China's stock market, especially for neural networks. On the other hand, the dividend price ratio (*dp*), market volatility (*svar*), aggregate earnings per share (*ep*), term spread (*tms*), and market liquidity (*mtr*) are less important, as they are overlooked by most models.

3.2.2. Stock characteristics

Not all of our stock characteristics are equally important in predicting stock returns, and their importance may depend strongly on the prediction model. To get an overview, Fig. 2 illustrates the overall importance of all characteristics based on the pooled full sample. We order

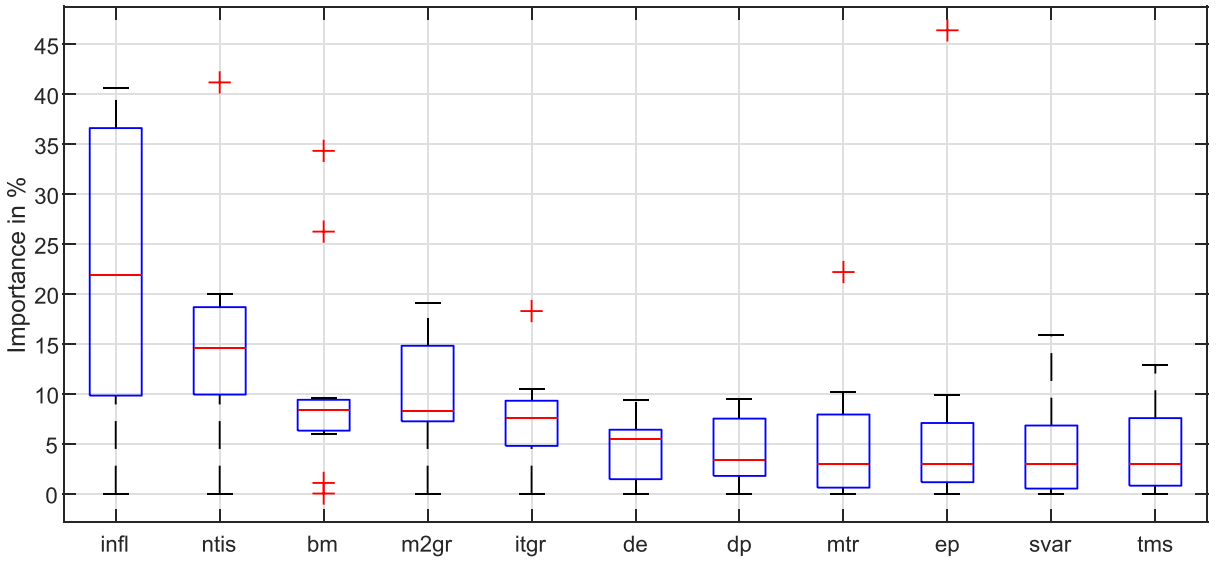

Fig. 1. Variable importance for eleven macroeconomic variables. This figure illustrates a box plot for the relative variable importance in Table 4 aggregated for each of the eleven macroeconomic variables.

Table 4
Relative variable importance of macroeconomic variables. This table reports the R^2-based variable importance for macroeconomic variables in each model. For a given model, the sum of variable importance is normalized to one. All values are in percentage.

	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

characteristics along the vertical axis by calculating the sum of the ranks of R^2-based variable importance for every predictor in each model and sorting them from the highest to the lowest. Such an ordering reflects the overall contribution of a characteristic to all models. Each column corresponds to a prediction model, where the color gradient indicates the model-specific importance from the highest to the lowest important (darkest to lightest).

With regards to the ordering of overall variable importance, we find that stock characteristics relating to market liquidity are most relevant when predicting the Chinese stock market, namely volatility of liquidity (*std_dolvol* and *std_turn*), zero trading days (*zerotrade*), and the illiquidity measure (*ill*) as the most salient predictors. The second influential group contains fundamental signals and valuation ratios, such as industry-adjusted change in asset turnover (*chaotia*), industry-adjusted change in employees (*chempia*), total market value (*mve*), number of recent earning increases (*nincr*), industry-adjusted change in profit margin (*chpmia*), and industry-adjusted book-to-market (*bm_ia*). The third group consists of risk measures, including idiosyncratic return volatility (*idiovol*), total return volatility (*volatility*), and market beta (*beta*). Our finding contrasts those in Gu et al. (2020) for the US market. They find that conventional price trend indicators are the most influential predictors, which turn out to be less important for the Chinese stock market except for recent maximum return (*maxret*). This observation resonates well with previous studies that apply linear factor models to predict the Chinese stock market (e.g., Li et al. (2010); Cakici et al. (2017)). Nevertheless, the prominent role of fundamental factors surprises us since, according to Gu et al. (2020), these factors turn out to be of minor importance for the US market. To be more specific, when we take the first three (ten) factors from Fig. 5 in Gu et al. (2020), their average rank in the Chinese market would be 41 (34). Hence, the two markets disagree substantially on the importance of the predictors.

Interestingly, the abnormal turnover ratio (*atr*), a China-specific factor initially introduced by Pan et al. (2015) to capture the impact of prevalent speculative trading, is also influential in machine learning models (ranked the third in

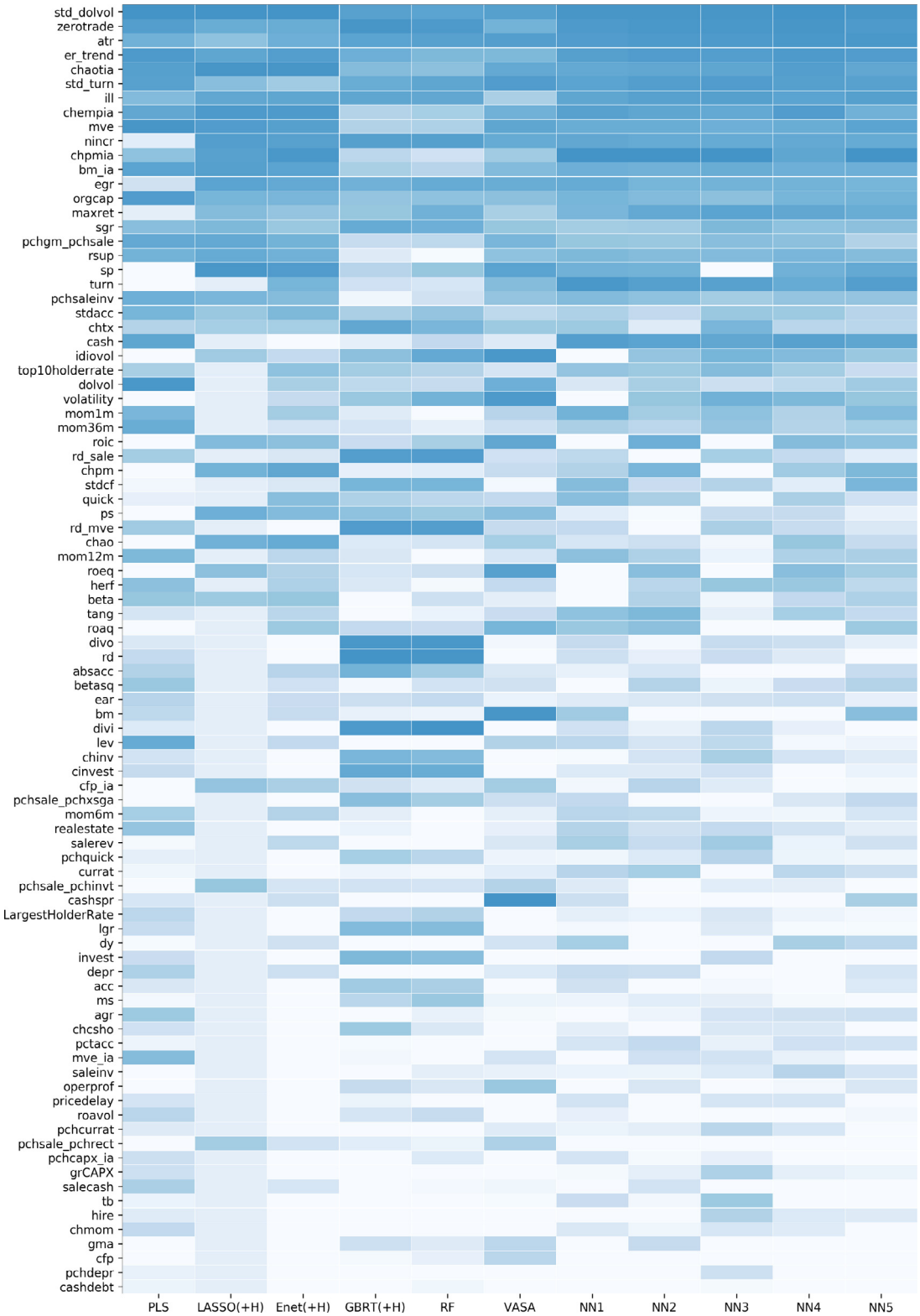

Fig. 2. Characteristic importance for all models. This figure shows the ordering of all stock-level characteristics ranked by their overall model contribution. Characteristics on the vertical axis are ordered based on the sum of their ranks over all models, with the most influential characteristics on the top and the least influential on the bottom. Columns correspond to the individual models, and the color gradients within each column indicate the most influential (dark blue) to the least influential (white) variables.

terms of overall variable importance). Also, the trend factor introduced by Liu et al. (2020) (*er_trend*) to account for the persistent trends in price and volume in the Chinese stock market has the fourth-largest overall variable importance. It is worth noting that the authors originally introduce both *atr* and *er_trend* to accommodate the influence of a large amount of active individual investors in the Chinese stock market on empirical asset pricing. Those individual investors are known to be more short-term oriented and trade speculatively, with a contribution of more than 80% of the total trading volume. Previous studies, such as Pan et al. (2015) and Liu et al. (2020), demonstrate the importance of including China-specific factors in factor models, while here we provide further evidence that these factors also have considerable explanatory power in more complicated machine learning models.

Similar to Gu et al. (2020), we also observe that neural network models (NN1-NN5), regularized linear models (PLS, LASSO, Enet), and VASA tend to emphasize a similar set of stock-level predictors. At the same time, the tree-based models, GBRT and RF, instead put more weight on a few predictors than others, such as *divo, rd*, and *divi*. We conjecture that such a difference is due to tree models' generic properties as they randomly choose a subset of stock characteristics when building decision trees. In this way, predictors like *divo, rd*, and *divi*, can become quite influential in some decision trees and thus become more relevant for the whole tree models, while they play a minor role in all other models.

From a practical and theoretical viewpoint, we are also interested in the time variation of the variable importance. We find that regularized linear models, including PLS, LASSO, and Enet, share a similar set of relevant predictors, with liquidity measures and fundamental signals being the two important groups of predictors. LASSO usually selects around 20 relevant predictors, and Enet selects around 35 predictors, indicating many characteristics are, in fact, redundant. There are only minor time variations in variable importance for PLS, compared to only about two-thirds of predictors selected by LASSO and Enet being stable across different periods. It is interesting to note that, particularly for LASSO, there seems to be a gap in variable importance between the periods before and after 2015, indicating a structural change in the stock market. As is well-known, the Chinese stock market went through a dramatic boom and a sudden crash in 2015, potentially explaining this finding (Liu et al., 2016).

The tree-based models, including GBRT and RF, tend to select a broader set of characteristics than alternative models, which has also been observed in Gu et al. (2020). Again, liquidity variables and fundamental signals are the two most important groups of predictors for GBRT and RF, but their orderings of variables slightly differ from other models. On the other hand, the time variations of variable importance for the tree models are relatively low. Here we also observe a gap in variable importance before and after 2015, especially for RF, such as *ill, idiovol*, and *maxret*. VASA's behavior in terms of variable importance is quite similar to PLS because VASA is built with linear submodels, except for a higher level of time variations in variable importance.

Lastly, neural network models (NN1 - NN5) favor liquidity variables, fundamental signals, valuation ratios, and China-specific factors including the abnormal turnover ratio (*atr*), the trend factor (*er_trend*), and the top-10 shareholders ownership (*top10holderrate*). Compared to other models, neural networks have substantially larger time variations in variable importance, indicating they can detect and account for the structural breaks in the forecasting ability of different predictors. We attribute this finding to the flexibility and adaptability of neural network models, especially when they are fine-tuned and well-trained with a sufficient amount of data.

3.3. Alternative model selection

Using the out-of-sample R^2 for model selection may not work well in practice, as some predictive models can have close out-of-sample R^2s but very different performance in reality. For example, in Table 1, the GBRT model has a slightly larger overall out-of-sample R^2 than NN4. However, this overall performance is mainly driven by GBRT's performance in 2018, while, for example, NN4's prediction performance measured by R^2_{oos} is, in fact, more robust than GBRT in most periods (see Fig. D.1 in the Internet Appendix D). As an alternative model selection method, we first use the unconditional superior predictive ability (USPA) test of Hansen (2005). However, within our analysis, we notice that Hansen's (2005) test alone still fails to distinguish some prediction models' performance, which is also the case for the Diebold and Mariano (1995) test used in Gu et al. (2020). To address this issue, we further look into the models' conditional predictive ability using the conditional superior predictive ability (CSPA) test in Li et al. (2020), which allows us to compare the performance of machine learning methods in different macroeconomic environments. See Internet Appendix B for a detailed description of both tests.

Table 5 reports the number of rejections of a given model under the USPA and CSPA tests. The USPA test results indicate that the naïve OLS model and the modified OLS-3 model perform poorly, having the largest total number of rejections. The GBRT, RF, NN3, NN4, and NN5 models have uniformly better unconditional prediction performance than their alternatives, but the USPA test fails to differentiate their performance. Therefore, we also compare the CSPA test results.[8] We observe that NN1, NN4, and NN5 have the smallest total number of CSPA test rejections. Even though tree models, including RF and GBRT, also perform well, their one-versus-all comparisons get rejected when conditioning on the market-level stock variance, while NN4 and NN5 can survive the same comparison. Also, NN4 and NN5 perform remarkably well under most macroeconomic conditions. Hence, the CSPA

[8] In particular, we condition on six conditioning variables, which can be classified into three groups: (1) inflation (*infl*) and M2 growth rate (*m2gr*), which reflect the overall macroeconomic environment; (2) market-level book-to-market ratio (*bm*) and dividend price ratio (*dp*), which measure the valuation level; (3) monthly turnover (*mtr*) and stock variance (*svar*), which indicate market-level volatility and liquidity. All other CSPA tests can be obtained from the authors, together with the analysis of different subsamples confirming our main results.

Table 5
Comparison of (un)conditional superior predictive ability based on full sample. The first column reports the number of rejections of the one-versus-one USPA test for row models at the 5% significance level based on the full sample. The next six columns report similar summary statistics of the conditional superior predictive ability tests (Li et al. (2020)) for different conditioning variables. For the CSPA tests, the entries report the number of rejections of the CSPA tests against the rest 12 competing models for a specific pair of the row model and the column conditioning variable. The last column reports the total number of rejections of the CSPA tests. For each entry, an asterisk indicates the rejection of a one-versus-all test at the 5% significance level.

		CSPA Test						
	USPA	*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

test enables us to differentiate the prediction performance of VASA, NN2, and regularized linear models more comprehensively, providing statistical evidence that these models are less favorable than NN4 and NN5. The Internet Appendix E.1 shows how the CSPA could be used for an ex-ante selection of the prediction model when forming portfolio strategies.

3.4. Dissecting the predictability performance of NN4

The previous analysis demonstrates that neural networks seem to outperform other models in terms of predictability. An often mentioned drawback of these algorithms is their lack of interpretability. Nevertheless, as a sanity check and to provide some intuition about which variables are causing the considerable predictability, we dig deeper into the drivers of the prediction performance. To this end, we focus on the striking differences in the monthly and annual R^2_{oos}s for small and large stocks generated by the NN4 model, as we later will use this neural net for portfolio analysis. In the following discussion, we focus on small and large stocks. Similar arguments will hold for the differences between the other subcategories.

In Panel A of Fig. 3, we plot the differences in the 20 most important variables using NN4 to predict the top 70% and the bottom 30% stocks on a monthly horizon. The three most important variables do not change their ordering when we move from large to small stocks: (1) *chempia*, the industry-adjusted change in the number of employees, is a proxy for a firm's distress using the industry-adjusted change in employees, and has been successfully applied in the US market by (Asness et al., 2000); (2) *std_dolvol* measures the standard deviation of daily trading volume and serves as a proxy for liquidity; and (3) *atr* is a China-specific liquidity factor. As Pan et al. (2016) argue, *atr* isolates speculative trading from liquidity and other components in trading volume. Therefore, it performs well since individual investors contribute to most of the total trading volume. While all three variables are equally important for large and small firms at a monthly horizon, the results in Panel B of Fig. 3 suggest that their influence within the two groups goes down at an annual horizon, which is entirely in line with intuition.

While the first three variables are equally important, the relative importance for most of the other variables changes. In particular, we find that liquidity-related variables like *zerotrade* and *std_turnorver* obtain more weight for small stocks, while fundamental variables like *cash, nincr, bm_ia*, and *orgcap* obtain less weight. Besides the liquidity-related variables, volatility-related variables like *volatility, idiovol*, and *max_ret*, and the China-specific trend variable *er_trend* obtain more importance. We discuss these latter variables next. First, with *idiovol* being a more important predictor for small stocks, our results lend support to the theory of limited arbitrage (see, e.g., Shleifer and Vishny (1997); Wurgler and Zhuravskaya (2002); Pontiff (2006)), which postulates that anomalies become stronger for high idiosyncratic risk stocks, leading to increased overall predictability.[9]

Second, the fact that *max_ret* also plays a more prominent role confirms our conjecture that retail investors significantly influence the price dynamics of small stocks. As Bali et al. (2011) show, if there is a strong preference among investors for assets with lottery-like payoffs, extreme positive returns exhibit significant predictability in the cross-sectional pricing of stocks. Moreover, they find that this effect is more prevalent for small stocks with extreme positive returns. Hence, their finding nicely coincides with our finding of the importance that NN4 attaches to *max_ret*.

Lastly, Liu et al. (2020) show that their China-specific trend factor (*er_trend*) works well because it reflects the

[9] The differences in R^2_{oos}'s between large and small stocks seems to be the most substantial among all the three subgroups. However, we also analyzed the relative differences between small stocks and the non-SOEs and A.M.C.P.S. Bottom 30%. We find that compared with non-SOEs, the small stock category puts considerably more weight on *atc* and *zerotrade*. Compared to A.M.C.P.S. Bottom 30%, small stocks put more weight on *idiovol* and *volatility*.

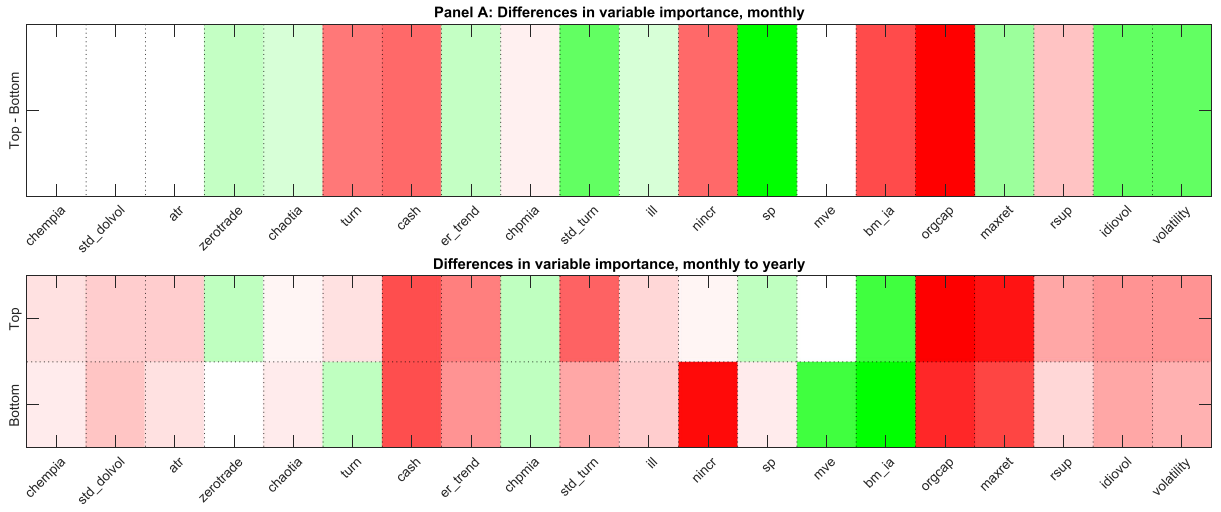

Fig. 3. Relative variable importance. This figure visualizes the changes in variable importance for the NN4 model. In Panel A, we plot the change in variable importance when moving from the top 70% to the bottom 30% stocks for the monthly strategy. In Panel B, we plot the changes with these two groups when moving from a monthly to a yearly strategy. The red color denotes a decrease, and the green color denotes an increase in importance. The ordering of the variables corresponds to their variable importance for the whole sample of stocks at the monthly prediction horizon.

market sentiment measured by the volatility of noise trader demand, and this effect is enforced by the dominance of retail investors in the Chinese market. Our NN4 model underscores the importance of this China-specific trend factor for monthly predictions for small stocks. While these latter variables are related to the influence of retail investors on monthly predictions, Panel B of Fig. 3 shows that they become substantially less important on an annual horizon. Obviously, speculative effects tend to wash out at longer horizons.

Panel A of Fig. 3 reveals the general tendency that under the NN4 model fundamental variables have less impact on the predictability of smaller stocks. Nevertheless, the sales-to-price variable *sp* used in Barbee et al. (1996) stands out as it obtains more relevance for smaller stocks.[10] Interestingly, the importance of *sp* for the Chinese market has also been noticed by Bin et al. (2017), where they show that smaller firms with top-performing stocks tend to have significantly higher sales-to-price ratios than all other stocks.

Instead of focusing further on the importance of specific characteristics, we place different characteristics into representative categories to avoid analyzing potential outliers. In Table C.4 in the Internet Appendix, we group all of our variables into ten different categories related to liquidity, momentum, ownership, size, volatility, earnings, beta, book-value ratios, growth, and leverage. Panel A in Fig. 4 shows that for both large and small stocks, liquidity measures turn out to be the most crucial driver of monthly predictability. However, what drives a wedge between the R^2_{oos}s is the overweighting of volatility and momentum categories for small stocks and the underweighting of market factors (*C_beta*) and fundamentals like (*C_growth* and *C_size*).[11]

Moving from a monthly to an annual forecast horizon, we find that liquidity and momentum lose their importance in favor of ownership, growth, and leverage. The size category seems to become more important for small firms. To provide additional insight on the relative differences, Panel C in Fig. 4 shows that the relative importance differences for annual predictions level off for small and large stocks. We identify only some differences in *C_bpr* and *C_size*. This finding resonates well with the small differences in the R^2 values of small and large stocks for annual predictions.[12]

Overall, the importance that the neural network NN4 gives to the different firm characteristics and their categories aligns well with our intuition. Moreover, it helps us to rationalize the differences between the predictability of small and large stocks. However, the overall predictability of the Chinese stock market still appears substantial compared to, for example, the US market. The overall predictability in the Chinese market might result from short-sale constraints, which are a universal feature of the Chinese market. Especially when retail investors dominate, these constraints might further enforce predictability and potential overpricing, compared to other markets.

[10] As Fisher (1984) argued, a high *sp* indicates that the stocks are popular with investors, providing buying opportunities. Fisher is an American billionaire investment analyst who ran Forbes' "Portfolio Strategy" column from 1984 to 2017, making him the longest continuously-running columnist in the magazine's history.

[11] The ranking of variables under NN4 (and other neural networks) is quite different to the average ranking across all prediction models, which puts more weight on the fundamental factors. In contrast, neural networks seem to favor momentum and volatility factors over fundamentals.

[12] Note that we find other differences between SOEs and Non-SOEs, and the A.M.C.P.S subgroups. For instance, SOEs put more emphasis on *C_size* and *C_growth*, and less on *C_bpr* and *C_ey* relative to non-SOEs. The top 70% in terms of A.M.C.P.S. put more weight on *C_own* and *C_vol* and much less on *C_beta*.

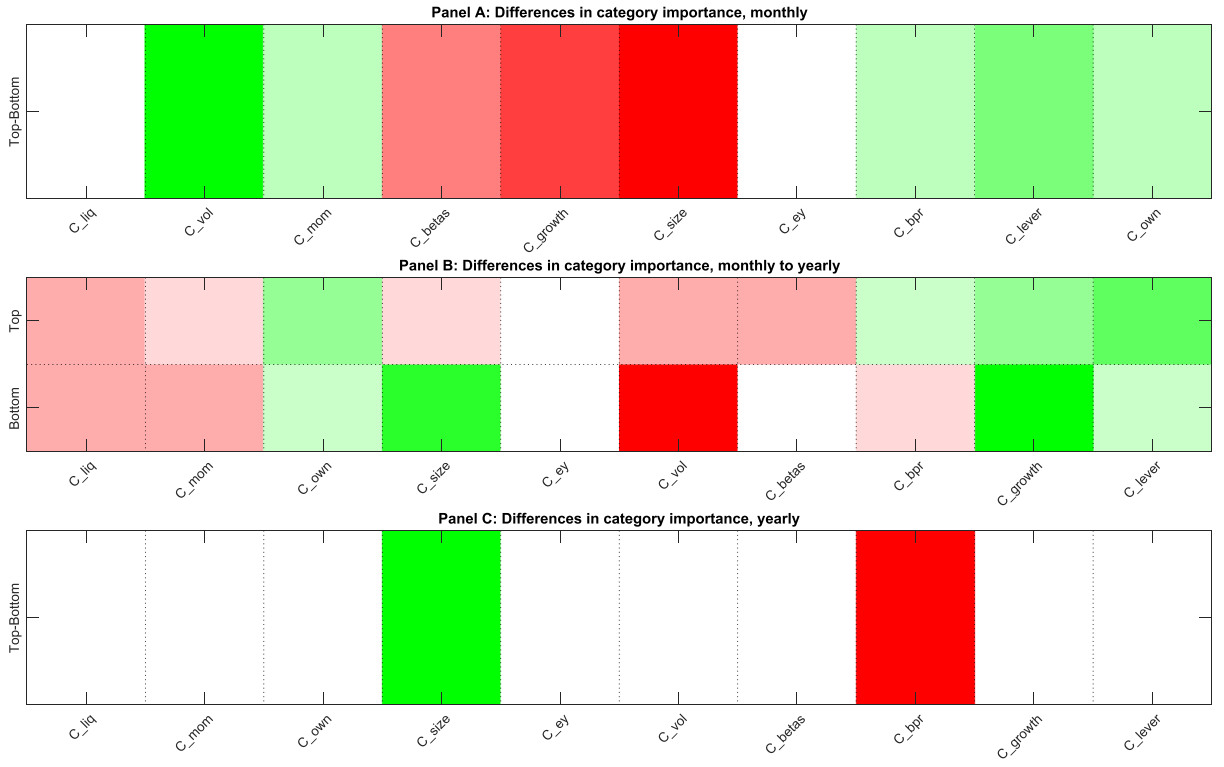

Fig. 4. Relative importance of variable categories. This figure visualizes the changes in aggregated variable importance for the NN4 model. We aggregate the variables into the categories defined in Table C.4 in the Internet Appendix. Panel A shows the differences between the top 70% and the bottom 30%, and Panel B shows the corresponding changes from monthly to yearly predictions. In Panel C, we show the same graph as Panel A but for yearly predictions. The red color denotes a decrease, and the green color denotes an increase in importance. The ordering of the variables in Panel A (Panels B and C) corresponds to the median rank of the categories' variable importance for the whole sample of stocks at the monthly (yearly) prediction horizon. Having defined these categories, we then sort them according to the median rank in monthly predictions for each category and all stocks. To analyze the differences, we look for each category at the two most important variables and how their average changes when we move from large to small stocks.

4. Portfolio analysis

So far, our assessment of prediction performance has been entirely statistical, relying on comparisons of out-of-sample predictive R^2 and two statistical tests. We next analyze whether this predictability can be exploited in portfolio strategies that account for short-selling constraints and other restrictions in the Chinese market.

4.1. Portfolio sorts

We consider two types of machine learning portfolios. The first one is the long-short portfolio, which we construct following the schemes in Gu et al. (2020). More precisely, at the end of each month, the one-month-ahead out-of-sample stock returns are generated for each method. We then sort stocks into deciles based on the predicted returns and reconstitute portfolios each month using value weights. Hence, a zero-net-investment portfolio we construct by buying the highest expected return stocks (decile 10) and selling the lowest (decile 1). Even though the long-short portfolio is a useful tool for evaluating machine learning methods' portfolio-level performance, it can hardly be implemented in the Chinese stock market due to strict short-selling restrictions.[13] We thus also include the long-only portfolio, which only holds stocks in the top decile.

Table 6 reports the out-of-sample performance for the value-weighted long-short and long-only portfolios.[14] For comparative purposes, we also report the performance of the 1/*N*-portfolio in which all stocks are equally-weighted. All machine learning portfolios dominate the OLS-3 portfolio and the 1/*N*-portfolio in terms of average expected monthly return, Sharpe ratio, and other measures. Overall, the results clearly demonstrate that machine learning techniques, especially neural network models, are advantageous for portfolio-level forecasts.

Figure 5 illustrates the evolution of the cumulative returns for the three portfolios constructed by different

[13] The China Securities Regulatory Commission (CSRC) introduced margin trading and short selling in March 2010. There were only 90 stocks available for short-selling initially but had increased to 800 as of July 2020. However, this number is still small relative to the total number of stocks in the Chinese market, which is over 4,000.

[14] In addition to the value-weighted portfolios, we also consider equally-weighted portfolios, whose performance is reported in Table E.6 in the Internet Appendix. The results are qualitatively similar to those of Table 6 except for slightly higher Sharpe ratios that are mostly driven by micro-cap stocks.

Table 6
Performance of machine learning portfolios based on the full sample (value-weighted). This table reports the out-of-sample performance measures for all machine learning models of the value-weighted long-short and long-only portfolios based on the full sample. All measures are based on 103 monthly out-of-sample returns from January 2012 to June 2020. "Avg": average predicted monthly return (%). "Std": the standard deviation of monthly predicted monthly returns (%). "S.R.": annualized Sharpe ratio. "Skew": skewness. "Kurt": kurtosis. "Max DD": the portfolio maximum drawdowns (%). "Max 1M Loss": the most extreme negative monthly return (%).

		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	–	1.80	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

methods, along with the market index CSI 300 as a benchmark. The neural network models dominate their competitors in all three portfolio types.[15] VASA, despite its simplicity, proves to be the second-best method, following NN4 closely. Note that the long-short portfolio for these two methods performs very well during the stock market crash in 2015, as indicated by the shaded area. Moreover, the recent global shock due to the COVID-19 pandemic in early 2020 does not lead to a notable downturn in portfolio levels. Neural networks and VASA are followed by penalized linear models, including LASSO and Enet, which have very similar performance as these two methods share much in common, while the performance of the tree models lags behind. However, all the machine learning portfolios outperform the 1/*N*-portfolio and the market index.

Our results in Fig. 5 and Table 6 confirm the finding of Gu et al. (2020) that neural networks outperform all other models considered in their study. For the long-short portfolios, we obtain substantially higher Sharpe ratios in the Chinese stock market than those for the US market found in Gu et al. (2020). For example, the highest Sharpe ratio (SR= 3.45) given by NN3 in the Chinese market is more than double their best Sharpe ratio (SR= 1.35) generated by NN4. As discussed above, the long-short strategy is nearly infeasible due to trading restrictions, so we are cautious in interpreting these results. At the same time, the highest Sharpe ratio for the long-only portfolio is 1.76, still higher than the long-short strategy for the US market. Given this high level, it is crucial to assess the performance of the long-only portfolio under more realistic assumptions.

[15] Here, we only include NN4 in the figure for the sake of simplicity as the performance of the other neural network models is very similar.

4.2. Excluding small stocks

As a robustness check, we repeat the previous portfolio analysis based on the top 70% subsample. There are three main reasons for such practice. First, small stocks are well-known for their high price volatility in the Chinese stock market, making it difficult for investors to find appropriate buying points. Second, the bottom 30% stocks often suffer the so-called shell-value problem caused by the IPO constraints in China, as documented in Liu et al. (2019). Third, in general, large stocks have higher levels of liquidity and lower price volatility and thus are less affected by the 10% daily price limits in China.

Table 7 reports the results. The performance of machine learning portfolios based on the top 70% large stocks are qualitatively similar to the full sample. However, all portfolios achieve lower average monthly returns, Sharpe ratios, standard deviations, and extreme negative monthly returns because small stocks are excluded. Nevertheless, machine learning methods still substantially dominate the simple OLS-3 model and the 1/*N* portfolio, with neural networks performing the best, followed by the regularized linear models and the tree models. Therefore, these results confirm that machine learning methods also have an outstanding portfolio-level predictive power in the Chinese stock market.

4.3. Performance of SOEs

The results in Table 3 reveal considerable return predictability for SOEs, particularly for complex models like neural networks. Political connections may boost the SOEs' performance through various channels such as, e.g., easier access to bank loans, loose regulations, and lighter taxation. At the same time, it is well known that the SOEs'

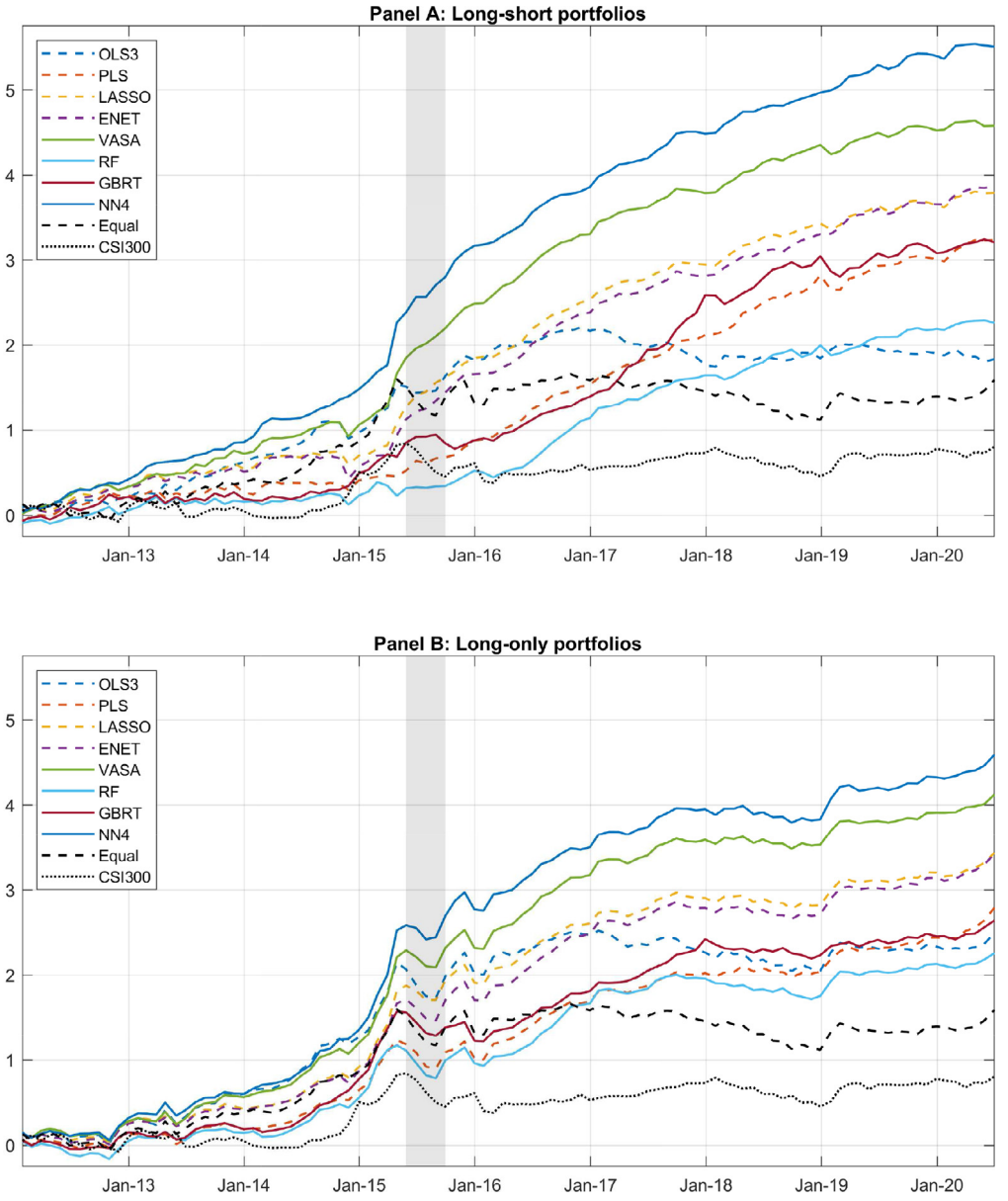

Fig. 5. Cumulative log return of machine learning portfolios (full sample). This figure shows the cumulative log returns of all portfolios and the CSI 300 market index. The shaded period corresponds to the 2015 stock market crash in China. All portfolios are constructed based on the full sample and are value weighted. In Panel A, the portfolios are based on a long-short strategy. Panel B plots the long-only portfolios.

highly concentrated state ownership, their financial opacity and low informative share prices, and their lack of corporate governance mechanisms could potentially exacerbate the crash risk for these firms. Therefore, it is interesting to examine how the SOEs' predictability manifests in different portfolio strategies' performance. In Table 8, we report the results for the long-short and long-only strategies.

Given that SOEs are mostly large companies, we compare the results in Table 8 those in Table 7. First, 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. For NN5, we get a Sharpe ratio of 4.12 compared to a Sharpe ratio of 2.70 for the top 70% stocks. For the long-only portfolio, we note that the $1/N$ portfolio indeed indicates a larger drawdown risk for SOE stocks than for the top 70% stocks (which also include SOEs). However, exploiting the predictability of SOE returns, we can reduce the maximum drawdown for the long-only strategy to levels that are considerably below the levels for the largest 70% stocks. At the same time, the Sharpe ratios are also higher for the long-only SOE portfolio. Therefore, using an appropriate prediction algorithm, we can mitigate the concerns of previous studies that SOEs generate a larger exposure to crash risk.

Table 7
Performance of machine learning portfolios based on the top 70% sample (value-weighted). This table reports the out-of-sample performance measures for all machine learning models of the value-weighted long-short and long-only portfolios based on the Top 70% sample. All measures are based on 103 monthly out-of-sample returns from January 2012 to June 2020. "Avg": average predicted monthly return (%). "Std": the standard deviation of monthly predicted monthly returns (%). "S.R.": annualized Sharpe ratio. "Skew": skewness. "Kurt": kurtosis. "Max DD": the portfolio maximum drawdowns (%). "Max 1M Loss": the most extreme negative monthly return (%).

		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	–	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

Table 8
Performance of machine learning portfolios based on SOEs (value-weighted). This table reports the out-of-sample performance measures for all machine learning models of the value-weighted long-short and long-only portfolios based on SOEs. All measures are based on 103 monthly out-of-sample returns from January 2012 to June 2020. "Avg": average predicted monthly return (%). "Std": the standard deviation of monthly predicted monthly returns (%). "S.R.": annualized Sharpe ratio. "Skew": skewness. "Kurt": kurtosis. "Max DD": the portfolio maximum drawdowns (%). "Max 1M Loss": the most extreme negative monthly return (%).

		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

4.4. Transaction costs

To assess the economic significance of the portfolios' performance, we ultimately have to include transaction costs in our analysis. For the Chinese market, the cost of an A-share transaction mainly consists of three components: commission, stamp tax, and slippage. Compared to commissions and the stamp tax, slippage requires a more careful investigation as it is often difficult to execute all transactions at the pre-specified price without affecting market price due to the liquidity issue. In the Chinese stock market, the commission fee for institutional investors was around 5 bps in 2012, then quickly decreased. In recent years, the commission fee is usually 2-3 bps for retail investors and even lower for institutional investors. The stamp tax has been set to 10 bps since 2008 and is collected unilaterally from sellers.

We consider two trading schemes to quantify the size of slippage. The first one relies on the time-weighted average price (TWAP) for the first 30 minutes in the first trading day of a given month, as we assume orders are split equally and implemented at the beginning of every minute. The slippage is thus the relative difference between the TWAP and the open price. Similarly, the second

Table 9
Slippage of machine learning portfolios. This table reports relevant summary statistics (average, standard error, skewness, kurtosis, first quantile, third quantile) of slippage for machine learning portfolios in the testing sample, including the time-weighted average price (in bps), the volume-weighted average price (in bps), and the conservative trading volume (in billion). The definitions of TWAP, VWAP, and market capacity are detailed in the first paragraph in Section 4.4.

	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
TWAP (buy)												
Avg	2.65	2.84	1.71	2.16	2.52	4.45	1.44	2.56	3.48	3.34	3.01	3.49
Std	59.32	47.62	48.64	50.13	49.49	50.73	51.73	50.24	49.04	49.47	51.15	50.60
Skew	−3.26	−3.57	−3.34	−3.35	−3.94	−3.33	−3.47	−3.62	−3.59	−3.34	−3.27	−3.13
Kurt	22.13	24.46	23.06	23.08	27.50	21.33	23.63	24.86	25.00	22.65	21.56	20.25
$q_{0.25}$	−12.95	−8.10	−11.63	−10.74	−9.87	−7.36	−9.28	−7.10	−6.97	−7.66	−7.03	−9.86
$q_{0.75}$	28.60	22.73	23.00	23.33	26.61	28.89	24.36	23.31	23.89	23.22	25.70	25.32
TWAP (sell)												
Avg	−5.62	−7.32	−7.14	−7.72	−8.40	−8.67	−7.60	−6.96	−8.30	−7.53	−8.02	−8.05
Std	35.90	30.81	29.80	31.86	31.40	34.09	32.01	32.29	31.59	30.56	32.54	32.53
Skew	1.89	1.52	1.39	1.73	1.88	1.50	1.63	1.44	1.60	1.17	1.04	1.33
Kurt	16.55	15.20	14.84	16.57	17.39	16.22	15.56	15.97	16.34	13.82	13.13	13.51
$q_{0.25}$	−21.98	−18.24	−19.02	−20.80	−19.81	−22.39	−21.02	−19.48	−21.34	−19.79	−20.76	−21.13
$q_{0.75}$	7.64	3.13	4.95	3.38	3.24	1.26	4.47	3.73	1.74	4.69	4.18	2.54
VWAP (buy)												
Avg	3.08	3.07	1.38	1.85	3.15	5.06	0.97	2.40	3.60	3.13	2.75	3.15
Std	61.48	50.01	51.50	52.81	50.93	52.98	54.60	53.49	51.99	52.06	53.59	53.55
Skew	−3.74	−3.98	−3.78	−3.80	−4.12	−3.74	−3.88	−4.08	−4.01	−3.78	−3.67	−3.52
Kurt	26.42	28.98	27.29	27.51	29.78	25.53	27.90	29.65	29.43	26.95	25.77	24.07
$q_{0.25}$	−11.71	−7.92	−12.29	−11.18	−9.40	−6.86	−10.38	−8.83	−6.58	−7.55	−8.53	−9.30
$q_{0.75}$	29.53	23.00	24.01	25.46	28.42	30.69	23.82	24.94	27.02	23.98	25.92	26.52
VWAP (sell)												
Avg	−5.11	−7.04	−6.69	−7.21	−8.08	−8.51	−7.04	−6.53	−7.60	−6.89	−7.49	−7.69
Std	37.16	31.31	30.90	32.90	31.90	35.05	32.72	33.34	32.81	31.38	33.40	33.29
Skew	3.10	2.63	2.66	3.04	3.08	2.89	2.84	2.79	3.00	2.42	2.25	2.49
Kurt	23.42	20.96	21.37	24.19	24.67	23.95	22.51	22.61	24.26	19.74	18.60	19.75
$q_{0.25}$	−22.70	−19.36	−18.35	−20.50	−20.28	−23.03	−20.82	−19.53	−20.38	−19.77	−21.43	−20.40
$q_{0.75}$	8.07	3.35	4.53	4.04	3.01	2.17	3.71	3.24	2.35	3.52	3.83	2.71
Market Capacity												
Avg	2.04	3.44	2.65	3.14	5.56	4.65	2.71	3.49	3.41	3.20	3.44	3.58
Std	1.96	3.65	3.35	4.01	4.57	4.80	3.37	4.20	3.63	3.49	3.86	3.64
Skew	4.79	4.24	6.58	6.16	2.29	4.31	6.26	4.09	5.03	5.85	5.57	5.16
Kurt	36.76	26.72	56.69	50.92	20.90	27.82	53.31	23.61	39.08	48.83	45.18	41.30
$q_{0.25}$	1.03	1.62	1.12	1.36	2.61	2.14	1.98	1.53	1.41	1.19	1.56	1.54
$q_{0.75}$	2.53	3.77	3.13	3.62	6.65	5.28	3.42	4.03	4.45	3.98	4.54	4.84

one estimates the volume-weighted average price (VWAP), where we impute trading volumes for each minute interval by taking the 20-day moving average and execute orders proportionally to the predicted trading volumes. In addition, we provide rough estimates of market capacities by calculating 5% of the trading volumes of the stocks traded.

Table 9 reports some relevant summary statistics for TWAP, VWAP, and market capacities. On average, the total deviation of the TWAP and VWAP from the open price is around 10 bps after accounting for both buying and selling. In some rare cases, such as the 2015 Chinese stock market turbulence, the scale of slippage can be quite large as the stock market goes up or down rapidly right after the stock market opening. However, in such cases, the signs of buying and selling slippage are likely the same, which could partly reduce the actual slippage that investors face. A back-of-the-envelope calculation indicates that 25 bps might be a reasonable estimate of transaction cost in the Chinese stock market during normal times. However, given that slippage can be higher than 10 bps under some extreme circumstances, we take a conservative approach by considering trading costs of 20, 40, 60, and 80 bps to account for the effect of transaction costs on portfolio performance.

In Table 10, we report the monthly returns and the Sharpe ratios when we include different levels of transaction costs. It turns out that, due to the low frequency of our strategies, the portfolios still provide a considerable and economically significant performance. For our benchmark strategy, the NN4, the Sharpe ratio in the long-short setting decreases from 2.91 to 2.34 in the extreme case when we assume a round trip cost of 80 bps. Using a more realistic assumption of 20 bps, the Sharpe ratio decreases only to 2.76. A similar observation can be made for the long-only strategy, which is more relevant from a practitioner's viewpoint. For the long-only strategy, the Sharpe ratio's decrease is from 1.68 to 1.46 under the assumption of 80 bps. Therefore, our transaction cost analysis shows that the different strategies' performance remains economically significant even under conservative assumptions about the magnitude of transaction costs.

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

Table 11
Impacts of machine learning portfolios. This table reports the out-of-sample performance measures for all machine learning models of the equally-weighted long-only and long-only portfolios with tradable stocks, i.e., excluding stocks at price limits. All measures are based on 103 monthly out-of-sample returns from January 2012 to June 2020. "Avg": average predicted monthly return (%). "S.R.": annualized Sharpe ratio. "Nontradable": fraction of stocks that are not tradable (%).

	Machine Learning Portfolios											
	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
Long-only												
Avg	2.24	3.67	4.05	4.20	3.83	3.48	4.38	4.50	4.45	4.74	4.91	4.85
S.R.	0.85	1.64	1.54	1.58	1.58	1.42	1.66	1.63	1.77	1.77	1.78	1.73
Tradable												
Avg	2.23	3.45	3.76	3.91	3.52	3.21	4.08	4.19	4.19	4.42	4.59	4.53
S.R.	0.84	1.55	1.47	1.50	1.48	1.31	1.57	1.55	1.68	1.68	1.70	1.65
Nontradable	0.1	0.5	0.6	0.6	0.7	0.7	0.6	0.7	0.7	0.5	0.7	0.8

4.5. Daily price limits

Daily price limit rules are widely used in stock exchanges around the world, especially in emerging markets, in the hope that they will serve as a market stabilization mechanism (Deb et al., 2010). China's market imposes daily price limits of 10% on regular stocks listed in Main Board and Second Board (20% on stocks listed in Second Board since August 2020), 5% on special treatment (ST) stocks, and 20% on stocks listed in Sci-Tech Innovation Board. For the Chinese market, Chen et al. (2019b) find that price limits incentivize large investors to pursue a destructive strategy of pushing up stock prices to the upper price limit and then selling on the next day. Hence, they argue that this unintended effect renders daily price limits counterproductive.

Given that our predicting horizon is the one-month forward return rather than daily returns, we conjecture that our main results will only be mildly affected by price limit rules. To explore the effect on portfolio performance, we proceed as follows. On each rebalancing date, we exclude stocks that are closed at the upper price limits for buying targets and postpone the selling targets to the date when the prices are not at the lower price limits.

Table 11 reports the results for the long-only portfolio. Indeed, we find that both the returns and the Sharpe ratios remain high. For instance, for NN4, the Sharpe ratio declines from 1.78 to 1.70. Hence, overall, our results remain robust to the inclusion of the price limit rule.

5. Conclusion

We investigate several machine learning method's predictive power in the Chinese stock market. We find that the most critical factors are liquidity-based trading signals. What surprises us is that signals based on price momentum only play a minor role. It takes many years for a stock market to develop the qualities that allow and encourage fundamental investing. The Chinese stock market is moving in that direction, but our results indicate that fundamental factors are the second most crucial factor category. We also find that the short-termism of retail investors generates substantial predictability at short investment horizons, particularly for small stocks. Simultaneously, since governmental signaling plays such an essential role in the Chinese market, we observe a substantial increase in SOEs' predictability at longer horizons.

Our portfolio analysis shows that the high predictability at short horizons translates into high Sharpe ratios for long-short portfolios. In particular, neural networks and VASA also provide a robust performance during the Chinese stock market crash in 2015. However, shorting stocks in the Chinese market is not practical. Therefore, we also analyze the long-only portfolio and find that the performance remains economically significant. We also present a new way of performing an ex-ante model selection, which generates significant performance. Overall, we show that machine learning methods can be (even more) successfully applied to markets that have entirely different characteristics than the US market.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jfineco.2021.08.017.

References

Allen, F., Qian, J., Qian, M., 2005. Law, finance, and economic growth in China. J. Financ. Econ. 77 (1), 57–116.
Asness, C.S., Porter, R.B., Stevens, R.L., 2000. Predicting Stock Returns Using Industry-Relative Firm Characteristics. Available at SSRN 213872.
Bai, C.-E., Lu, J., Tao, Z., 2006. The multitask theory of state enterprise reform: empirical evidence from China. Am. Econ. Rev. 96 (2), 353–357.
Bali, T.G., Cakici, N., Whitelaw, R.F., 2011. Maxing out: stocks as lotteries and the cross-section of expected returns. J. Financ. Econ. 99 (2), 427–446.
Barbee, W., Mukherji, S., Raines, G., 1996. Do sales-price and debt-equity explain stock returns better than book-market and firm size? Financ. Anal. J. 52 (2), 56–60.
Bin, L., Chen, J., Puclik, M., Su, Y., 2017. Predicting extreme returns in Chinese stock market: an application of contextual fundamental analysis. J. Account. Finance 17 (3), 10.
Bryzgalova, S., Pelger, M., Zhu, J., 2019. Forest Through the Trees: Building Cross-Sections of Stock Returns. Available at SSRN 3493458.
Cakici, N., Chan, K., Topyan, K., 2017. Cross-sectional stock return predictability in China. Eur. J. Finance 23 (7-9), 581–605.
Chen, L., Pelger, M., Zhu, J., 2019. Deep Learning in Asset Pricing. Available at SSRN 3350138.
Chen, T., Gao, Z., He, J., Jiang, W., Xiong, W., 2019. Daily price limits and destructive market behavior. J. Econom. 208 (1), 249–264.
De Nard, G., Hediger, S., Leippold, M., 2020. Subsampled Factor Models for Asset Pricing: The Rise of Vasa. Available at SSRN 3557957.
Deb, S.S., Kalev, P.S., Marisetty, V.B., 2010. Are price limits really bad for equity markets? J. Bank. Finance 34 (10), 2462–2471.
Diebold, F.M., Mariano, R., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 20 (1).
Feng, G., Polson, N., Xu, J., 2019. Deep Learning in Characteristics-Sorted Factor Models. Available at SSRN 3243683.
Fisher, K.L., 1984. Super Stocks. Irwin Professional Publishing.
Gan, J., Guo, Y., Xu, C., 2018. Decentralized privatization and change of control rights in China. Rev. Financ. Stud. 31 (10), 3854–3894.
Gao, K., Ding, M., 2019. Short-sale refinancing and price adjustment speed to bad news: evidence from a quasi-natural experiment in China. China J. Account. Res. 12 (4), 379–394.
Green, J., Hand, J., Zhang, F., 2017. The characteristics that provide independent information about average US monthly stock returns. Rev. Financ. Stud. 30, 4389–4436.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
Gu, S., Kelly, B., Xiu, D., 2021. Autoencoder asset pricing models. J. Econom. 222 (1), 429–450.
Hansen, P.R., 2005. A test for superior predictive ability. J. Bus. Econ. Stat. 23, 365–380.
Huber, P.J., 2004. Robust Statistics, vol. 523. John Wiley & Sons.
Jiang, F., Kim, K.A., 2020. Corporate governance in China: a survey. Rev. Finance 24 (4), 733–772.
Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: a unified model of risk and return. J. Financ. Econ. 134 (3), 501–524.
Lee, W., Wang, L., 2017. Do political connections affect stock price crash risk? Firm-level evidence from China. Rev. Quant. Finance Account. 48 (3), 643–676.
Li, B., Qiu, J., Wu, Y., 2010. Momentum and seasonality in Chinese stock markets. J. Money Invest. Bank. 17 (5), 24–36.
Li, J., Liao, Z., Quaedvlieg, R., 2020. Conditional superior predictive ability. Rev. Econ. Stud.. Forthcoming
Liu, D., Gu, H., Xing, T., 2016. The meltdown of the Chinese equity market in the summer of 2015. Int. Rev. Econ. Finance 45, 504–517.
Liu, J., Stambaugh, R.F., Yuan, Y., 2019. Size and value in China. J. Financ. Econ. 134 (1), 48–69.
Liu, Y., Zhou, G., Zhu, Y., 2020. Trend Factor in China. Available at SSRN 3402038.
Mei, J., Scheinkman, J.A., Xiong, W., 2009. Speculative trading and stock prices: evidence from Chinese AB share premia. Ann. Econ. Finance 10 (2).
Ng, L., Wu, F., 2006. Revealed stock preferences of individual investors: evidence from Chinese equity markets. Pacific-Basin Finance J. 14 (2), 175–192.
Pan, L., Tang, Y., Xu, J., 2015. Speculative trading and stock returns. Rev. Finance 20 (5), 1835–1865.
Pan, L., Tang, Y., Xu, J., 2016. Speculative trading and stock returns. Rev. Finance 20 (5), 1835–1865.
Piotroski, J.D., Wong, T.J., Zhang, T., 2015. Political incentives to suppress negative information: evidence from Chinese listed firms. J. Account. Res. 53 (2), 405–459.
Pontiff, J., 2006. Costly arbitrage and the myth of idiosyncratic risk. J. Account. Econ. 42 (1-2), 35–52.
Saffi, P.A.C., Sigurdsson, K., 2011. Price efficiency and short selling. Rev. Financ. Stud. 24 (3), 821–852.
Shleifer, A., Vishny, R.W., 1997. The limits of arbitrage. J. Finance 52 (1), 35–55.
Welch, I., 2008. The Link Between Fama-French Time-Series Tests and Fama-Macbeth Cross-Sectional Tests. Technical Report. UCLA Anderson School of Management.
Wurgler, J., Zhuravskaya, E., 2002. Does arbitrage flatten demand curves for stocks? J. Bus. 75 (4), 583–608.