Machine-Learning in the Chinese Stock Market*

Markus Leippold† Qian Wang‡ Wenyu Zhou§¶

July 11, 2021

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 U.S. market, liquidity emerges as the most important predictor, leading us to examine the impact of transaction costs closely. The retail investors' dominating presence positively affects short-term predictability, particularly for small stocks. Another feature that distinguishes the Chinese from the U.S. 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.

JEL classification: C52, C55, C58, G0, G1, G17

Keywords: Chinese stock market, factor investing, machine learning, model selection.

*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 comments on an earlier version of the paper. Also, we thank Zhipeng Liao for sharing the code of the CSPA test and Zhe Wang for excellent research assistance. An earlier version of the paper circulated under the tile 'Machine-Learning in the Chinese Factor Zoo.'

†Department of Banking and Finance, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland, email: markus.leippold@bf.uzh.ch.

‡Department of Banking and Finance, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland, email: qian.wang@bf.uzh.ch.

§(Corresponding Author) International Business School, Zhejiang University, Haining, Zhejiang 314400, China, email: wenyuzhou@intl.zju.edu.cn.

¶Academy of Financial Research, Zhejiang University, Hangzhou, Zhejiang 310058, China

1 Introduction

As of October 2020, the total value of China's stock market has climbed to a record high of more than $10 trillion, as the country's accelerating economic recovery from the pandemic has surpassed the previous high reached during an equity bubble five years ago, making it the second-largest in the world, after the U.S. at nearly $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 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 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 U.S. 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-back when share prices fall below fundamental values. These automatic market correction mechanisms

[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.

[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. At the same time, institutional investors generated a profit of 1,116 billion yuan.

are therefore affected by government-oriented restrictions (Mei et al. 2009).[3] 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 ninety stocks on a special list.[4] 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 end of 2016. Although there is no broad consensus, many academics agree that short-selling helps price discovery, rendering markets more efficient.[5] While most of the literature on factor investing in U.S. and European markets rely 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.

As of today, 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.[6] 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 they have been constructed for the U.S. market. In a second step, we follow the literature by adapting some of these U.S. factors for the Chinese stock market.[7] 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'

[3]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 they are also suspected that 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).

[4]In February 2010, a pilot program was officially launched to further facilitate short selling by expanding the number of stocks that securities companies can lend to their clients. For more details, see Gao and Ding (2019).

[5]See, e.g., Saffi and Sigurdsson (2011).

[6]The data can be obtained by the authors on request.

[7]For instance, Liu et al. (2019) propose an adaptation of the classic Fama and French (1993) three-factor model to consider the special features of the Chinese stock market by adjusting the size and value factors. However, as they still fail to explain a large number of the anomalies observed, Liu et al. (2020) add a trend factor to account for the dominance of retail investors, who may be more susceptible to herding activities.

overpricing (Mei et al. 2009).

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.[8]

For our study, we build on the work of Gu et al. (2020b) which combines a broad repertoire of machine learning methods with modern empirical asset pricing research to understand the dynamics of market risk premia for stock returns.[9] 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.[10] 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, make this market a highly attractive target for applying modern 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. More-

[8]Recent contributions include, among many others, Giglio and Xiu (2017) and Kelly et al. (2019) using principal component analysis and its variants, Moritz and Zimmermann (2016) using tree-based models in the context of portfolio sorting, Freyberger et al. (2020) using the adaptive group LASSO to select important return predictors and providing evidence for substantial time variation in the predictive power of characteristics, Feng et al. (2019) using deep learning dynamic factor models.

[9]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. (2020b) have been suggested. Examples include Bryzgalova et al. (2019), Chen, Pelger and Zhu (2019), De Nard et al. (2020), Gu et al. (2020a), Feng et al. (2019).

[10]Although machine learning may substantially improve the pricing of assets, Arnott et al. (2019), Gu et al. (2020b), and Israel et al. (2020), among others, also point out that these methods cannot identify deep fundamental economic principles.

over, comparing the out-of-sample R^2 with studies in the U.S. market, the Chinese market reveals substantially more predictability. As the out-of-sample R^2 has some limitations for model selection, we analyze the models' conditional predictive ability using a statistical test recently developed in Li, Liao and Quaedvlieg (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 previous studies in the U.S. market (Gu et al. 2020b), where classical trend indicators are the main drivers of predictability. However, we find notable differences across models. In particular, in addition to liquidity, neural nets tend to favor momentum and volatility factors over fundamentals. We also find that 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.[11] Lastly, given the short-selling constraints in China, we wonder how much value-added can be enjoyed in long-only mandates. Many of the previous literature results 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 indicate that also 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. Section 3 presents our empirical analysis. We look at the out-of-sample predictability, and we discuss which predictors matter most. We also perform a model selection analysis using both the unconditional and conditional predictive ability tests. Section 4 explores whether predictability translates into portfolio gains. In Section 5, we conclude. Detailed discussions of the methods used and additional results are delegated to the Internet Appendix.

[11] For example, Piotroski et al. (2015) argue that SOEs have political incentives to suppress negative information.

2 Data and Methodology

For our analysis, we apply the empirical design of Gu et al. (2020b) 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 Wind Database, the largest financial data provider in China.[12] 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 that have been 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 literature listed in Green et al. (2017), and the literature on China-specific factors.[13] Our collection includes 94 characteristics in total, among which 86 have be documented in Green et al. (2017), four are valid China-specific factors identified in the previous literature, and four are binary variables which indicate ownership types for listed firms and are used for subsample analysis.[14] 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. (2020b), 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)

[12]There are two types of shares in the mainland China stock market, A-shares and B-shares. A-shares are quoted only in renminbi, while B-shares are quoted in foreign currencies, such as the U.S. dollar, and are available to foreign investors on a larger scale. Foreign investors may have difficulties accessing A-shares due to Chinese government regulations, and Chinese investors may have difficulties accessing B-shares, mainly for currency exchange reasons. Some companies choose to list their shares in both the A-share and B-share markets. According to public information, the Wind database serves more than 90% of Chinese financial institutions and 70% of Qualified Foreign Institutional Investors (QFII) operating in China.

[13]We closely adhere to the variable definitions in the original papers listed in Green et al. (2017), and the literature on China-specific factors. To avoid the forward-looking bias, we adopt the practice in Gu et al. (2020b). More precisely, we assume that monthly stock-level characteristics are delayed by at most one month, quarterly with at most four months lag, and annual with at most six months lag. Therefore, when predicting stock returns at $t + 1$, we use the most recent monthly characteristics at the end of month t, most recent quarterly data at the end of month $t - 4$, and most recent annual data at the end of month $t - 6$. If a characteristic is missing, we replace it with the cross-sectional median at each month for each stock, respectively.

[14]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. (2020b).

in 2012.[15] Table C.1 in Appendix C.1 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 website.[16] Eight of those variables are based on the variable definitions detailed 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 rest three include monthly turnover (*mtr*), M2 growth rate (*m2gr*), and international trade volume growth rate (*itgr*), which are identified by previous literature as effective macroeconomic predictors. Table C.5 in Appendix C.1 summarizes these macroeconomic variables.

Throughout our analysis, we adopt a general additive prediction error model to describe the relationship 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 eleven 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

[15] See http://www.csrc.gov.cn/pub/newsite/flb/flfg/bmgf/zh/gfxwjtj/201310/t20131016_236281.html.
[16] See http://www.stats.gov.cn/tjsj/.

set of dummy variables include the 80 industry dummies. Hence, the total number of covariates in $z_{i,t}$ is $90 \times (11+1) + 80 = 1160$.

In total, we consider eleven machine learning methods along with two simple linear models. In particular, we include ordinary least squares regression (OLS), 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. (2020b), we only focus on OLS, OLS-3, LASSO, Enet, and GBRT equipped with Huber loss function to avoid potential disturbance caused by extreme values in the data.[17]

We follow the standard approach in the literature for hyperparameters selection, model estimation, and performance evaluation.[18] In particular, we divide our data into three disjoint time 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 for optimizing the hyperparameters of our models. We select the hyperparameters that minimize the objective loss function[19] based on the observations in the validation sample. The testing sample contains the next twelve 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 similar as in Gu et al. (2020b) by refitting prediction models annually instead of monthly. When we refit a model, we increase the size of the training sample 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. Appendix A provides further details on hyperparameters training and prediction models.

[17] Given the tuning parameter M, the Huber loss function $H(x; M) = x^2$ if $|x| \leq M$, and $H(x; M) = 2M|x| - M^2$ if $|x| > M$. See Huber (2004) for more details.

[18] See, e.g., Gu et al. (2020b) and De Nard et al. (2020).

[19] The objective function is either l_2 loss or Huber loss depending on the model used for prediction.

3 Empirical analysis

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

3.1 Out-of-sample predictability

As in Gu et al. (2020b), we rely on the non-demeaned out-of-sample predictive R^2 to have a direct comparison with their results for the U.S. 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.[20] Therefore, to address potential behavioral stories, we also build two subsamples according to firm size with a 30% cutoff level. Table 1 summarizes the results for the different models and subsamples.

[Table 1 about here.]

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 OLS-3 is slightly lower than that for OLS (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

[20] Ng and Wu (2006) analyze data from Chinese brokerage accounts and find that retail investors prefer small company stocks, as they seem to be more attractive to investors with little capital and an inability to use leverage.

model performs much better in China's stock market than in the U.S. stock market. The R^2_{oos} for the latter has been reported to be negative (-3.46%) in Gu et al. (2020b). 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.[21]

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 a small advantage of LASSO (1.43%) and Enet (1.42%) relative to 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. (2020b) for the U.S. market. The R^2_{oos} for VASA is comparable to those for 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 has also been emphasized for the U.S. stock market in Gu et al. (2020b). The full-sample R^2_{oos} suggests that both GBRT and RF are competitive with neural networks. Unlike the U.S. 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, it strikes us that, in terms of monthly R^2_{oos}, machine learning techniques reveal much stronger predictability in the Chinese market than in the U.S. 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. (2020b) 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 doubled of the highest R^2_{oos} in the U.S. 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

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

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 state-owned enterprises, 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 big (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.[22] 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 naive 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 if retail investors play an important role in generating such a difference. To shed light on the connection between predictability and retail investors, we further 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

[22] We remark that these numbers are calculated by applying full estimated models (trained using all stocks) to these restricted subsamples, as we are interested in predictive heterogeneity across different types of stocks.

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/the Number of Shareholders, and classify all stocks into two groups based on the top 70% threshold.[23] At 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 very well for small-shareholder stocks in China. In brief, even though 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 relationship between model predictability and the role of retail investors.

3.1.4 SOEs and non-SOEs

When we turn our 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.[24] 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 state-owned enterprises in China tend to have a large market capitalization, as they usually represent the dominant companies in fundamental industries like

[23] 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, 0.6 quantiles, which generate the same pattern of model predictability. These results are not presented for the sake of simplicity but available upon request.

[24] 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 period of 2012-2020, respectively.

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.[25] However, our results contrast 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.[26]

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 has unveiled two important Chinese stock market features that differ from the U.S. market studied in Gu et al. (2020b). 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^2's 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

[25] We leave a further investigation of this observation as an exciting avenue for future research.

[26] See, e.g., Lee and Wang (2017).

smallest 30% stocks. The improved predictability of larger stocks could be caused by the improved predictability of SOEs, which substantially improves.[27] 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.

[Table 2 about here.]

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 state-owned enterprises. We attribute the short-term predictability, particularly for small stocks, to retail investors' prominent role in the Chinese stock market. As we will see later,[28] at shorter horizons, neural nets 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.

[Table 3 about here.]

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. (2020b) for the U.S. 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 U.S. market. Simultaneously, the out-of-sample R^2 for SOEs, which are usually large stocks, is more than double the value for large U.S. stocks. Strikingly, for small Chinese stocks, we observe an out-of-sample R^2 that is ten times higher than for the U.S. small stocks. For U.S. 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

[27] According to Jiang and Kim (2020), SOEs currently account for roughly one-third of firm numbers but two-thirds of market capitalization.

[28] See Section 3.4.

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 U.S. market, which we suspect are mainly due to retail investors' dominant effect on the short horizon and government initiatives, which can predominantly benefit SOEs.

Internet Appendix D explores the time variations in the out-of-sample R^2_{oos} of our models. For most models, we observe in Figure 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 United States, 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 case, the political risk related to a trade war between the U.S. 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. (2020b), 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 weights on *infl*, *m2gr* and *itgr*, showing these macroeconomic variables are also influential.

[Table 4 about here.]

As Table 4 suggests, 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 more on 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 these two methods can detect potentially complicated nonlinear interactions between macroeconomic variables and stock characteristics.

[Figure 1 about here.]

Figure 1 aggregates 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 to predict stock returns, and their importance may depend strongly on the prediction model. To get an overview, Figure 2 illustrates the overall importance of all characteristics based on the pooled full sample. We order characteristics along its 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).

[Figure 2 about here.]

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. (2020b) for the U.S. 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 literature that applies linear factor models to predict the Chinese stock market (see, e.g., Li et al. (2010) and Cakici et al. (2017)). Nevertheless, the prominent role of fundamental factors surprises us since, according to Gu et al. (2020b), these factors turn out to be of minor importance for the U.S. market. To be more specific, when we take the first three (ten) factors from figure 5 in Gu et al. (2020b), 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 3rd in 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) have demonstrated 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. (2020b), 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 weights 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.[29] 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 with only about 2/3 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).

[29] We only discuss their interpretation here and we plot only the relevant figure for the NN4 model, Figure D.2, in Internet Appendix D. All other model-specific variable importance plots across time can be obtained from the authors.

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. (2020b). 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 with 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^2's but very different performance in reality.[30] As an alternative model selection method, we first use the Unconditional Superior Predictive Ability (USPA) test of Hansen (2005). However, within our analysis, we have noticed 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. (2020b). To address this issue, we further look into the models' conditional predictive ability using the Conditional Superior Predictive Ability (CSPA) test recently developed in Li, Liao and Quaedvlieg (2020), which allows us to compare the performance of machine learning methods in different macroeconomic environments.[31]

[30]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, e.g., NN4's prediction performance measured by R^2_{oos} is, in fact, more robust than GBRT in most periods (see Figure D.1).

[31]For the reader's reference, we provide a detailed description of both tests in Internet Appendix B.

[Table 5 about here.]

Table 5 reports the number of rejections of a given model under the USPA and CSPA tests. The USPA test indicates that the naive OLS model and the modified OLS-3 model perform poorly, having the largest total number of rejections. The GBRT, RF, NN3, NN4, and NN5 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.[32] We observe that specially 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 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. 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.[33]

Figure 3, Panel A, plots 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

[32] 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.

[33] In the following discussion, we focus on small and large stocks. Similar arguments will hold for the differences between the other subcategories. These results can be obtained from the authors.

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 succesfully applied in the U.S. 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, performs well since individual investors contribute to most of the total trading volume. While all these three variables are equally important for large and small firms at a monthly horizon, Panel B of Figure 3 suggests that their influence within the two groups goes down at an annual horizon, which is entirely in line with intuition.

[Figure 3 about here.]

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, also volatility-related variables like *volatility*, *idiovol*, and *max_ret*, and the China-specific trend variable *er_trend* obtain more importance. We briefly 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,[34] which postulates that anomalies become stronger for high idiosyncratic risk stocks and hence, leading to increased overall predictability.[35]

Second, the fact that *max_ret* also plays a more prominent role confirms our conjecture that the influence of retail investors on the price dynamics of small stocks must be considerable. 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.

[34] See, e.g., Shleifer and Vishny (1997), Wurgler and Zhuravskaya (2002), Pontiff (2006).

[35] The differences inR^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-SOE 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*.

Moreover, they find that this effect is more prevalent for small stocks with extreme positive returns. Hence, their finding is nicely reflected in 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 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 Figure 3 shows that they become substantially less important on an annual horizon. Obviously, speculative effects tend wash out at longer horizons.

Panel A of Figure 3 reveals the general tendency that under NN4 fundamental variables have less impact on the predictability of smaller stocks. Nevertheless, the sales-to-price variable *sp* using in Barbee et al. (1996) stands out as it obtains more relevance for smaller stocks. [36] Interestingly, the importance of *sp* for the Chinese market has also been confirmed by Bin et al. (2017), where they show that top-performing stocks having smaller firm sizes tend to have significantly higher sales-to-price ratios relative to all other stocks.

[Figure 4 about here.]

Instead of focusing further on the importance of specific characteristics, we decided to collect different characteristics into representative categories to avoid analyzing potential outliers. Table C.4 groups all of our variables into ten different categories related to liquidity, momentum, ownership, size, volatility, earnings, beta, book-value ratios, growth, and leverage.[37] Figure 4, Panel A, 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 category for small stocks and the underweighting of market factors (*C_beta*) and fundamentals like (*C_growth* and *C_size*).[38]

[36] 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.

[37] This categorization serves as an intuitive aggregation. Of course, other alternative groupings can be justified.

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

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 shed further light on the relative differences, Figure 4, Panel C, shows that the relative importance differences for annual predictions levels 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.[39]

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 appears still substantial compared to, e.g., the U.S. 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.[40]

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. It will be interesting to analyze whether this predictability can indeed be exploited in portfolio strategies that account for short-selling constraints and other restrictions in the Chinese market. This is what we do next.

4.1 Portfolio sorts

We consider two types of machine learning portfolios. The first one is the long-short portfolio, which is constructed following the schemes in Gu et al. (2020b). 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

[39] Note that we find other differences between SOE and Non-SOE 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*.

[40] See, e.g., Nagel (2005) on the impact of short-selling constraints, also addressing the role of institutional investors.

value weights. Finally, a zero-net-investment portfolio is constructed 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.[41] 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.[42] As a comparison, 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.

[Table 6 about here.]

Figure 5 illustrates the evolution of the cumulative returns for the three portfolios constructed by different methods, along with the market index CSI 300 as a benchmark. The neural network models dominate their competitors in all three portfolio types.[43] VASA, despite its simplicity, proves to be the second-best method, following NN4 closely. It is intriguing to note that the long-short portfolio for these two methods performs very well during the stock market crash in 2015, indicated by the shaded area. Moreover, the most recent global shock due to the COVID-19 pandemic in early 2020 does not lead to a notable downturn in portfolio levels. Neural nets 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 as well as the market index.

[Figure 5 about here.]

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

[42] In addition to the value-weighted portfolios, we also consider the equally-weighted portfolios, whose performance is reported in Table E.7 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.

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

Our results in Figure 5 and Table 6 confirm the finding of Gu et al. (2020b) in that neural networks outperform all other models considered in their study. For the long-short portfolios, it is noteworthy that we obtain substantially higher Sharpe ratios in the Chinese stock market than those in the U.S. market found in Gu et al. (2020b). 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 before, 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 in the U.S. market. Given this high level, it will be crucial to assess the performance of the long-only portfolio under more realistic assumptions.

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, large stocks in general have higher levels of liquidity and lower price volatility, and thus are less affected by the daily price limits of 10% in China.

[Table 7 about here.]

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 perform 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 prediction performance in the Chinese stock market.

4.3 Performance of SOEs

Our results in Table 3 have revealed considerable return predictability for SOEs, particularly for complex models like neural networks. Political connections may boost the SOEs' performance through various channels.[44] At the same time, it is well known that the SOEs' 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.[45] Therefore, it is interesting to see 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.

[Table 8 about here.]

Given that SOEs are mostly large companies, we discuss the results in Table 8 in comparison to those in Table 7. First, we note that the long-short strategy's performance in terms of 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 successfully 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 successfully mitigate the previous literature's concern that SOEs generate a larger exposure to crash risk.

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 A-share transaction mainly

[44] Political connections may lead to easier access to bank loans, loose regulation, and lighter taxation.

[45] See, e.g., Xu et al. (2014) and Li et al. (2017) argue that the more severe agency problems for SOEs lead to more bad-news hoarding.

consists of three components: commission, stamp tax and slippage. Compared with commission and 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.[46] Here we consider two trading schemes to quantify the size of slippage. The first one relies on the time weighted average price (TWAP) with respect to 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 TWAP and open price. Similarly, the second 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 TWAP and VWAP from open price is around 10 bps after accounting for both buying and selling.[47] Hence, 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.[48] However, given the fact 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.

[Table 9 about here.]

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

[46] In the Chinese stock market, commission fee for institutional investors was around 5 bps in 2012 but decreased quickly after then. In recent years, commission fee is usually 2-3 bps for retail investors and even lower for institutional investors. The stamp tax is set to 10 bps since 2008 and is collected unilaterally from sellers.

[47] 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 fiercely right after 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.

[48] Take the NN4 model as an example: transaction cost = 2×2 bps (commission) + 10 bps (stamp tax) + 10 bps (slippage) = 24 bps for TWAP, and 25 bps for VWAP.

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 about here.]

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 Aug 2020), 5% on special treatment (ST) stocks, and 20% on stocks listed in Sci-Tech Innovation Board. For the Chinese market, Chen, Gao, He, Jiang and Xiong (2019) 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 as counterproductive.

[Table 11 about here.]

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 the 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

The Chinese stock market shares some remarkable features, such as the presence of SOEs and the dominance of retail investors with a strong preference for a collectivistic investment behavior. In such speculative markets, we expected technical indicators to matter more than fundamentals. 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 already today, 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 finding that different subsamples show different predictive patterns at different time horizons may motivate us to explore retail investors' roles further.

With our rich and unique dataset at hand, we employed several machine learning methods' predictive power. Our portfolio analysis shows that the high predictability at short horizons translates into high Sharpe ratios for long-short portfolios. In particular neural nets and VASA also provide a robust performance during the Chinese stock market crash in 2015. 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 presented a new way of performing an ex-ante model selection, which allows us to generate significant performance. Overall, we have shown that machine learning methods can be (even more) successfully applied to other markets with entirely different characteristics than the U.S. market. Finally, having identified several exciting avenues for research, we hope that providing our data to the research community simplifies these endeavors and further fosters our understanding of financial markets and the Chinese market in particular.

References

Allen, F., Qian, J. and Qian, M.: 2005, Law, finance, and economic growth in China, *Journal of Financial Economics* **77**(1), 57 – 116.

Arnott, R., Harvey, C. R. and Markowitz, H.: 2019, A backtesting protocol in the era of machine learning, *The Journal of Financial Data Science* **1**(1), 64–74.

Bai, C.-E., Lu, J. and Tao, Z.: 2006, The multitask theory of state enterprise reform: Empirical evidence from China, *American Economic Review* **96**(2), 353–357.

Baker, M. and Stein, J. C.: 2004, Market liquidity as a sentiment indicator, *Journal of Financial Markets* **7**(3), 271–299.

Bali, T. G., Cakici, N. and Whitelaw, R. F.: 2011, Maxing out: Stocks as lotteries and the cross-section of expected returns, *Journal of Financial Economics* **99**(2), 427–446.

Barbee, W., Mukherji, S. and Raines, G.: 1996, Do sales-price and debt-equity explain stock returns better than book-market and firm size?, *Financial Analysts Journal* **52(2)**, 56–60.

Bin, L., Chen, J., Puclik, M. and Su, Y.: 2017, Predicting extreme returns in Chinese stock market: An application of contextual fundamental analysis, *Journal of Accounting and Finance* **17**(3), 10.

Breiman, L.: 2001, Random forests, *Machine learning* **45**(1), 5–32.

Bryzgalova, S., Pelger, M. and Zhu, J.: 2019, Forest through the trees: Building cross-sections of stock returns, *Available at SSRN 3493458* .

Bühlmann, P. and Van De Geer, S.: 2011, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Cakici, N., Chan, K. and Topyan, K.: 2017, Cross-sectional stock return predictability in China, *The European Journal of Finance* **23**(7-9), 581–605.

Chen, L., Pelger, M. and Zhu, J.: 2019, Deep learning in asset pricing, *Available at SSRN 3350138* .

Chen, S.-S.: 2009, Predicting the bear stock market: Macroeconomic variables as leading indicators, *Journal of Banking & Finance* **33**(2), 211–223.

Chen, T., Gao, Z., He, J., Jiang, W. and Xiong, W.: 2019, Daily price limits and destructive market behavior, *Journal of econometrics* **208**(1), 249–264.

Chernozhukov, V., Lee, S. and Rosen, A. M.: 2013, Intersection bounds: estimation and inference, *Econometrica* **81**(2), 667–737.

De Nard, G., Hediger, S. and Leippold, M.: 2020, Subsampled factor models for asset pricing: The rise of vasa, *Available at SSRN* .

Deb, S. S., Kalev, P. S. and Marisetty, V. B.: 2010, Are price limits really bad for equity markets?, *Journal of Banking & Finance* **34**(10), 2462–2471.

Diebold, F. M. and Mariano, R.: 1995, Comparing predictive accuracy, *Journal of Business & Economic Statistics* **20**(1).

Fama, E. F. and French, K. R.: 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**(1), 3–56.

Feng, G., Polson, N. and Xu, J.: 2019, Deep learning in characteristics-sorted factor models, *Available at SSRN 3243683* .

Fisher, K. L.: 1984, Super stocks irwin professional publishing, *New York* .

Freyberger, J., Neuhierl, A. and Weber, M.: 2020, Dissecting characteristics nonparametrically, *The Review of Financial Studies* **33**(5), 2326–2377.

Gan, J., Guo, Y. and Xu, C.: 2018, Decentralized privatization and change of control rights in China, *The Review of Financial Studies* **31**(10), 3854–3894.

Gao, K. and Ding, M.: 2019, Short-sale refinancing and price adjustment speed to bad news: Evidence from a quasi-natural experiment in China, *China Journal of Accounting Research* **12**(4), 379–394.

Giglio, S. and Xiu, D.: 2017, Inference on risk premia in the presence of omitted factors, *Technical report*, National Bureau of Economic Research.

Green, J., Hand, J. and Zhang, F.: 2017, The characteristics that provide independent information about average US monthly stock returns, *The Review of Financial Studies* **30**, 4389–4436.

Gu, S., Kelly, B. and Xiu, D.: 2020a, Autoencoder asset pricing models, *Journal of Econometrics* **forthcoming**.

Gu, S., Kelly, B. and Xiu, D.: 2020b, Empirical asset pricing via machine learning, *The Review of Financial Studies* **33**(5), 2223–2273.

Hansen, P. R.: 2005, A test for superior predictive ability, *Journal of Business & Economics Statistics* **23**, 365–380.

Hornik, K., Stinchcombe, M., White, H. et al.: 1989, Multilayer feedforward networks are universal approximators., *Neural networks* **2**(5), 359–366.

Huber, P. J.: 2004, *Robust statistics*, Vol. 523, John Wiley & Sons.

Ioffe, S. and Szegedy, C.: 2015, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167* .

Israel, R., Kelly, B. T. and Moskowitz, T. J.: 2020, Can machines' learn'finance?, *Available at SSRN 3624052* .

James, G., Witten, D., Hastie, T. and Tibshirani, R.: 2013, *An introduction to statistical learning*, Vol. 112, Springer.

Jiang, F. and Kim, K. A.: 2020, Corporate Governance in China: A Survey, *Review of Finance* **24**(4), 733–772.

Kelly, B. T., Pruitt, S. and Su, Y.: 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* **134**(3), 501–524.

Kingma, D. P. and Ba, J.: 2014, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* .

Lee, W. and Wang, L.: 2017, Do political connections affect stock price crash risk? firm-level evidence from China, *Review of Quantitative Finance and Accounting* **48**(3), 643–676.

Li, B., Qiu, J. and Wu, Y.: 2010, Momentum and seasonality in Chinese stock markets, *Journal of Money, Investment and Banking* **17**(5), 24–36.

Li, J. and Liao, Z.: 2020, Uniform nonparametric inference for time series, *Journal of Econometrics* **219**(1), 38–51.

Li, J., Liao, Z. and Gao, M.: 2020, Uniform nonparametric inference for time series using stata, *Working paper* .

Li, J., Liao, Z. and Quaedvlieg, R.: 2020, Conditional superior predictive ability, *Available at SSRN 3536461* .

Li, X., Wang, S. S. and Wang, X.: 2017, Trust and stock price crash risk: Evidence from China, *Journal of Banking & Finance* **76**, 74–91.

Liu, D., Gu, H. and Xing, T.: 2016, The meltdown of the Chinese equity market in the summer of 2015, *International Review of Economics and Finance* **45**, 504 – 517.

Liu, J., Stambaugh, R. F. and Yuan, Y.: 2019, Size and value in China, *Journal of Financial Economics* **134**(1), 48–69.

Liu, Y., Zhou, G. and Zhu, Y.: 2020, Trend factor in China, *Available at SSRN 3402038* .

Medeiros, M. C. and Mendes, E. F.: 2016, ℓ1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors, *Journal of Econometrics* **191**(1), 255–271.

Mei, J., Scheinkman, J. A. and Xiong, W.: 2009, Speculative trading and stock prices: Evidence from Chinese ab share premia., *Annals of Economics & Finance* **10**(2).

Moritz, B. and Zimmermann, T.: 2016, Tree-based conditional portfolio sorts: The relation between past and future stock returns, *Available at SSRN 2740751* .

Nagel, S.: 2005, Short sales, institutional investors and the cross-section of stock returns, *Journal of Financial Economics* **78**(2), 277–309.

Ng, L. and Wu, F.: 2006, Revealed stock preferences of individual investors: Evidence from Chinese equity markets, *Pacific-Basin Finance Journal* **14**(2), 175–192.

Pan, L., Tang, Y. and Xu, J.: 2015, Speculative Trading and Stock Returns, *Review of Finance* **20**(5), 1835–1865.

Piotroski, J. D., Wong, T. and Zhang, T.: 2015, Political incentives to suppress negative information: Evidence from Chinese listed firms, *Journal of Accounting Research* **53**(2), 405–459.

Pontiff, J.: 2006, Costly arbitrage and the myth of idiosyncratic risk, *Journal of Accounting and Economics* **42**(1-2), 35–52.

Rapach, D. E., Strauss, J. K. and Zhou, G.: 2013, International stock return predictability: what is the role of the united states?, *The Journal of Finance* **68**(4), 1633–1662.

Saffi, P. A. C. and Sigurdsson, K.: 2011, Price efficiency and short selling, *The Review of Financial Studies* **24**(3), 821–852.

Shleifer, A. and Vishny, R. W.: 1997, The limits of arbitrage, *The Journal of Finance* **52**(1), 35–55.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: 2014, Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research* **15**(1), 1929–1958.

Welch, I.: 2008, The link between Fama-French time-series tests and Fama-Macbeth cross-sectional tests, *Technical report*, UCLA Anderson School of Management. Available at `https://ssrn.com/abstract=1271935`.

White, H. L.: 2000, A reality check for data snooping, *Econometrica* **68**(5), 1097–1126.

Wurgler, J. and Zhuravskaya, E.: 2002, Does arbitrage flatten demand curves for stocks?, *The Journal of Business* **75**(4), 583–608.

Xu, N., Li, X., Yuan, Q. and Chan, K. C.: 2014, Excess perks and stock price crash risk: Evidence from China, *Journal of Corporate Finance* **25**, 419–434.

Zou, H. and Hastie, T.: 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

Tables

Table 1. Monthly out-of-sample predictive R^2 in 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

Notes: This table reports monthly out-of-sample predictive R^2 of thirteen models for four 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; (7) non-state-owned-enterprises. The models considered include ordinary least squares regression (OLS), 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. (Non-)SOE represents the subgroup of (non-) state-owned enterprises. All the numbers are expressed in percentage values.

Table 2. Annual out-of-sample predictive R^2 in 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

Notes: This table reports annual out-of-sample predictive R^2 of thirteen models for four different subgroups of firms: (1) the full sample; (2) the sample excluding firms with the bottom 30% market values;(3) the sample including only the firms with the 30% bottom market values; (4) the sample including firms with the top 70% average market capitalization per shareholder; (5) the sample including the firms with the bottom 30% average market capitalization per shareholder; (6) state-owned-enterprises; (7) non-state-owned-enterprises. The models considered include ordinary least squares regression (OLS), 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. (Non-)SOE represents the subgroup of (non-) state-owned enterprises. All the numbers are expressed in percentage values.

Table 3. Average out-of-sample predictive R^2 in percentage for NN1 to NN5

	Bottom 30%	Top 70%	Small-shareholder	Large-shareholder	Non-SOE	SOE	U.S. bottom	U.S. 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)

Notes: 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 U.S. market as analyzed in Gu et al. (2020b), their tables 1 and 2. All the numbers are expressed in percentage values. The numbers in brackets represent the average out-of-sample predictive R^2 for all models, excluding OLS.

Table 4. Relative variable importance of eleven macroeconomic variables

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

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

Table 5. Comparison of (un)conditional superior predictive ability based on full sample

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

Notes: The first column reports the number of rejections of the one-versus-one USPA test for row models at 5% significance level based on the full sample. The next six columns report similar summary statistics of the conditional superior predictive ability tests (Li, Liao and Quaedvlieg (2020)) for different conditioning variables. For the CSPA tests, the entries report the number of rejections of the CSPA tests against the rest twelve 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 significance level of 5%.

Table 6. Performance of machine learning portfolios based on the full sample (value-weighted)

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

Notes: 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 (%).

Table 7. Performance of machine learning portfolios based on the top 70% sample (value-weighted)

	"1/N" Portfolio	OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
		Machine Learning Portfolios											
Long-Short													
Avg	–	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

Notes: 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 (%).

Table 8. Performance of machine learning portfolios based on SOEs (value-weighted)

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

Notes: 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 (%).

Table 9. Slippage of machine learning portfolios

	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

Notes: 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, market capacity are detailed in the first paragraph in Subsection 4.4.

Table 10. Portfolio performance including transaction costs (value-weighted)

	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

Notes: 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.

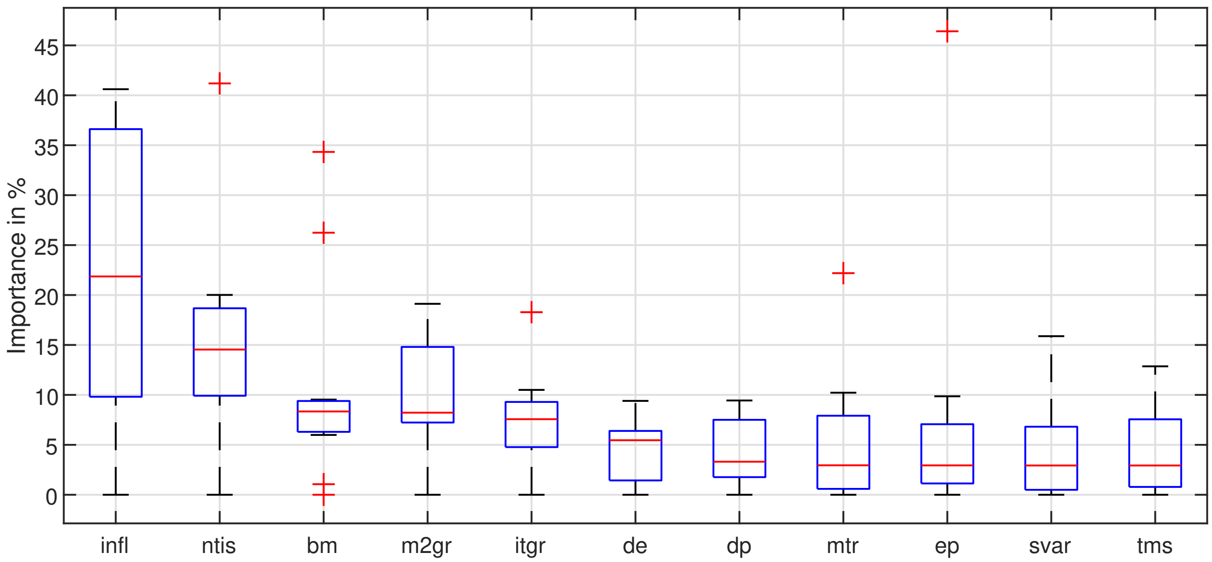
Table 11. Impacts of machine learning portfolios

	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

Notes: 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 (%).

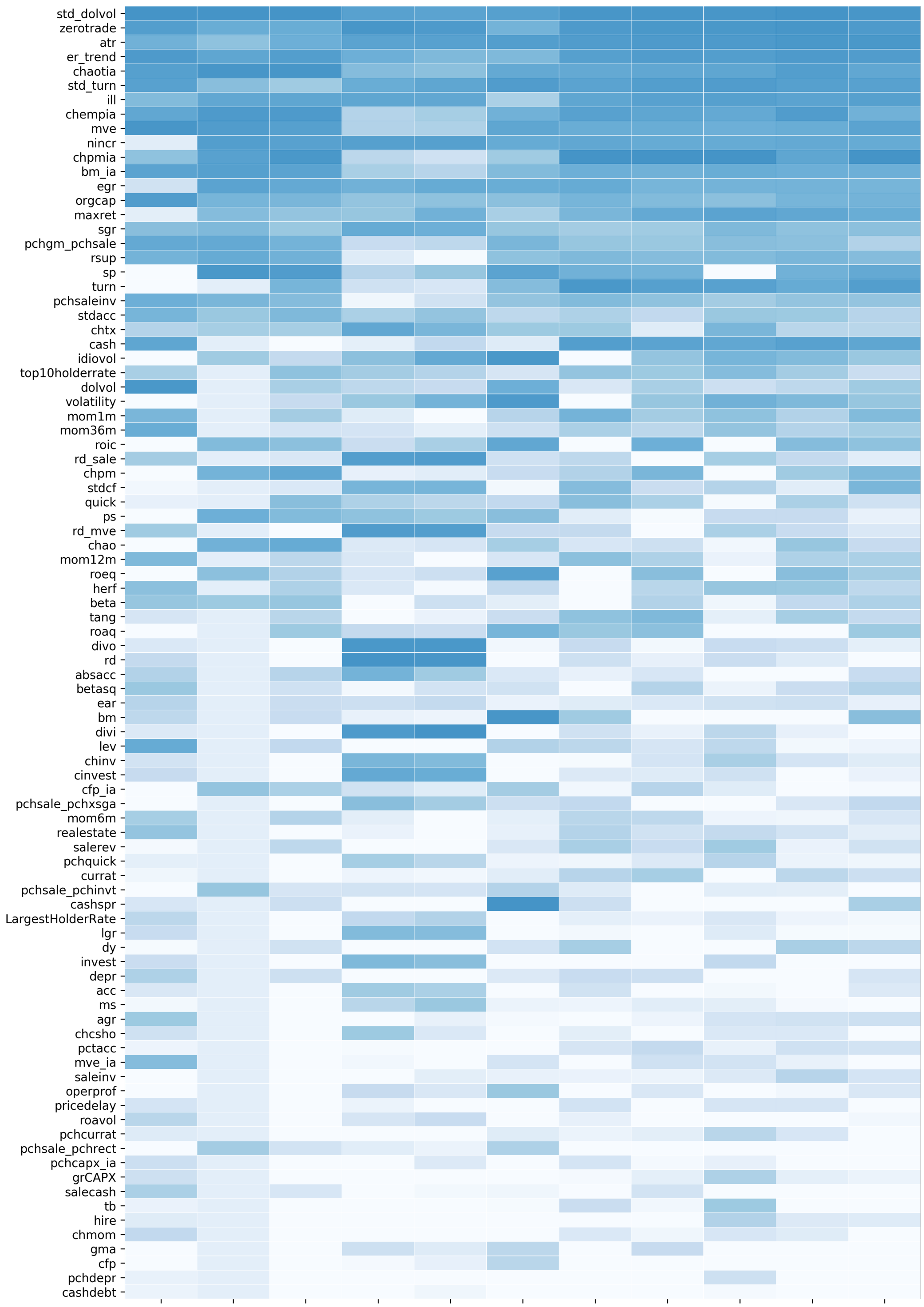
Figures

Figure 1. Variable importance for eleven macroeconomic variables



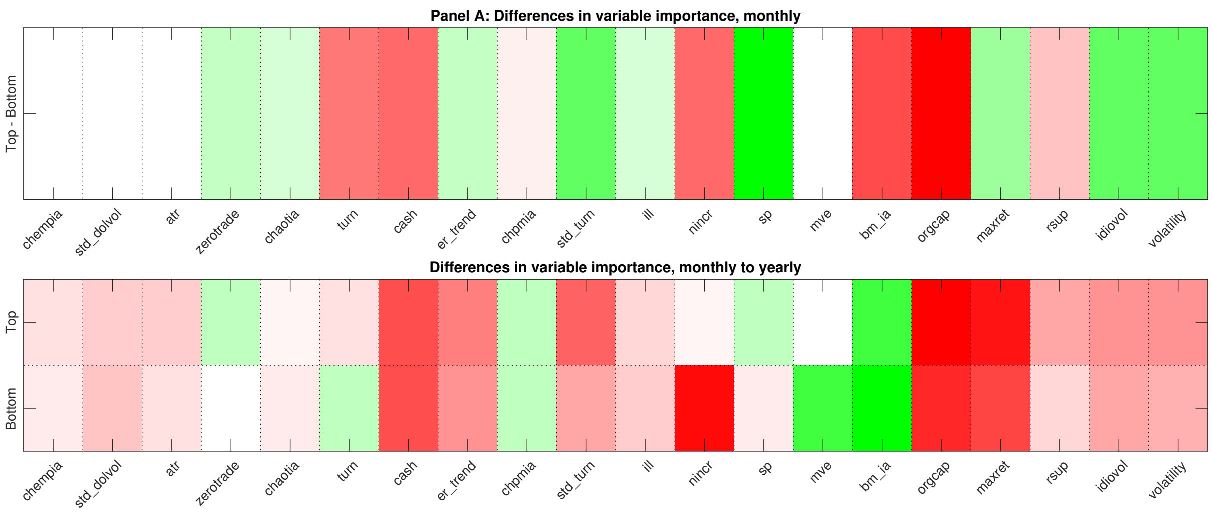
Notes: This figure illustrates a box plot for the relative variable importance in Table 4 aggregated for each of the eleven macroeconomic variables.

Figure 2. Characteristic importance for all models



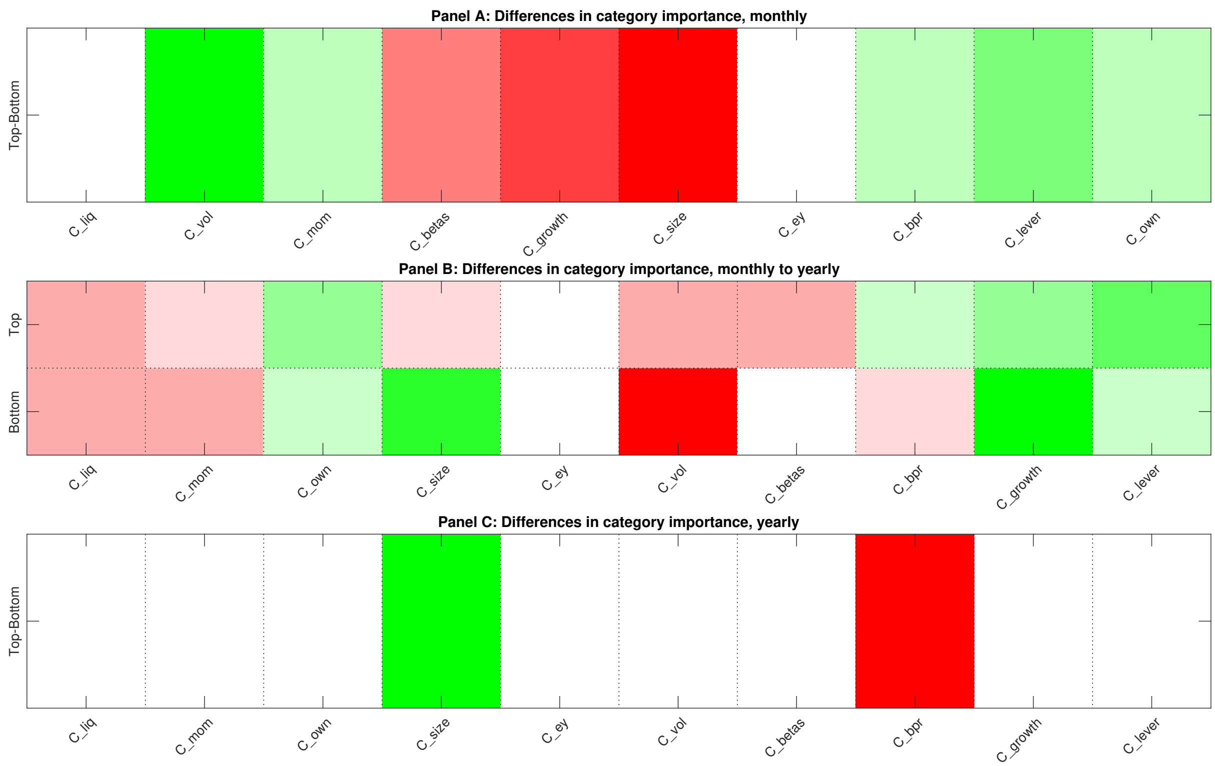
Notes: This figure shows the ordering of all stock-level characteristics ranked by their overall model contribution. Characteristics on the vertical axis are ordered based on the sum of their ranks over all models, with the most influential characteristics on the top and the least influential on the 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.

Figure 3. Relative variable importance



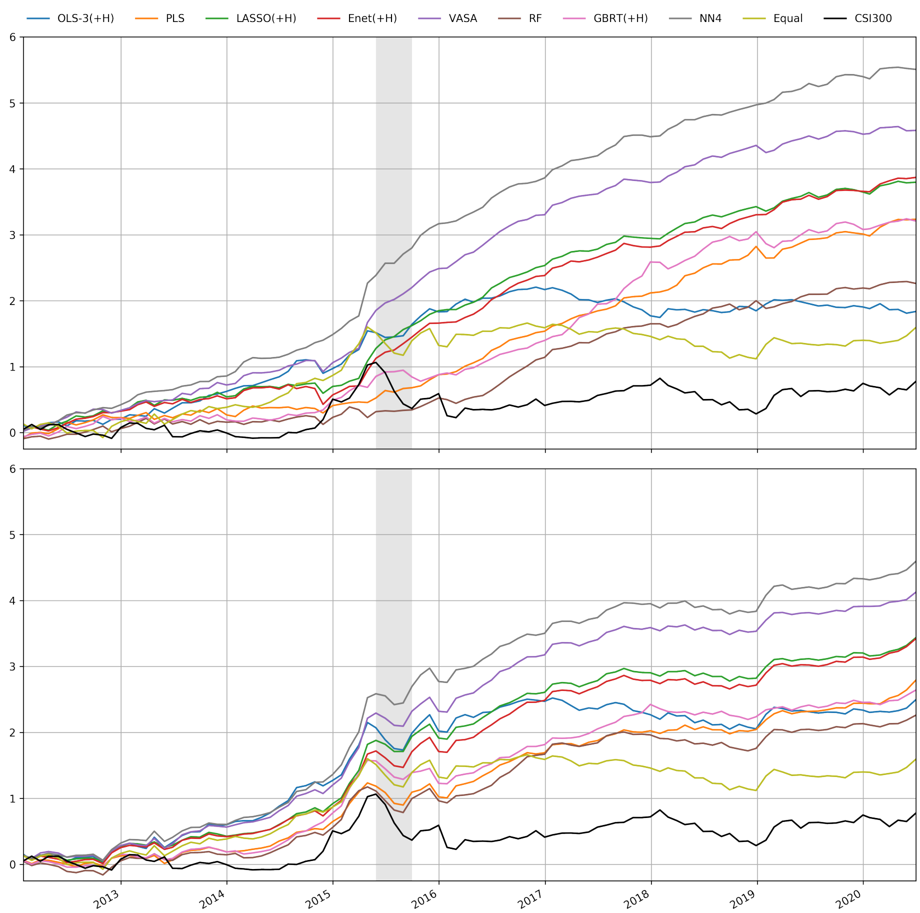
Notes: This figure visualizes the changes in variable importance for the NN4 model. In Panel A, we plot the change in variable importance when moving from the top 70% to the bottom 30% stocks for the monthly strategy. In Panel B, we plot the changes with these two groups when moving from a monthly to 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.

Figure 4. Relative importance of variable categories



Notes: 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. 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.

Figure 5. Cumulative log return of machine learning portfolios (full sample)



Notes: 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 the first panel, the portfolios are based on a long-short strategy. The second panel plots the long-only portfolios.

Internet Appendix

“Machine-Learning in the Chinese Stock Market”

A Methodology

A.1 Hyperparamters

Table A.1 summarizes the hyperparameters for all prediction models and the corresponding specifications.

Table A.1. Hyperparameters For All Models

	OLS-3+H	PLS	LASSO+H	Enet+H	GBRT+H
Huber loss $M = 1.35$	$\surd$		$\surd$	$\surd$	$\surd$
Specification	bm, mve, mom1m	K	$\lambda \in (10^{-4}, 10^{-1})$	$\rho = 0.5$ $\lambda \in (10^{-4}, 10^{-1})$	#Depth $L = 1 \sim 3$ #Trees $B = 1 \sim 1000$ Learning Rate $LR \in \{0.01, 0.1\}$
	RF	VASA	NN1-NN5		
Specification	#Depth $L = 1 \sim 7$ #Trees $B = 100 \sim 300$ #Features $f = 3 \sim 50$	#Subsamples $B = 1 \sim 300$ #Components $K = 1 \sim 50$	L1 penalty $\lambda \in (10^{-5}, 10^{-2})$ Learning Rate $LR \in (10^{-4}, 10^{-2})$ Batch Size $B \in \{64, 512, 2048, 10000\}$ Epochs $= 100$ Patience $= 5$ Adam Para.= Default Ensemble $= 10$		

The table describes the hyperparameters for all machine learning models.

A.2 Simple linear regression

We take linear regression equipped with the Huber loss function as the reference model, because it is arguably the simplest tool that has been widely used for prediction. Since the linear regression model does not have any hyperparameters, we thus merge the validation sample into the training sample. The model imposes a linear structure on the conditional expectation of stock i's excess return $g(z_{it})$, i.e.,

$$r_{i,t+1} = g(z_{i,t};\theta) + \epsilon_{i,t} = z_{i,t}'\theta + \epsilon_{i,t}, \tag{5}$$

where $g(\cdot)$ and z_{it} are defined in Section 2, θ is the vector of coefficients which includes the intercept term. This model can be estimated handily, with a feasible closed-form solution when l_2 loss is adopted for the objective function. However, linear regression model estimated using the l_2 loss is vulnerable to outliers in the data, which can leads to very bad prediction performance.[49] Given the fact that financial returns and stock-level characteristics are often heavy-tailed, we instead minimize the Huber robust objective function following Gu et al. (2020b), which is given by

$$L_H(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H(r_{i,t+1} - g(z_{i,t};\theta); M), \tag{6}$$

where

$$H(x;M) = \begin{cases} x^2, & \text{if } |x| \leq M, \\ 2M|x| - M^2, & \text{if } |x| > M. \end{cases}$$

The Huber loss function can be understood as a "trimmed" l_2 loss function, in which the threshold is determined by the tuning parameter M. In addition, the Huber loss function can be embedded into other machine learning methods as well, such as regularized linear models, and tree models. In our empirical analysis, we estimate OLS-3, LASSO, Enet, PLS, and GBRT based on the Huber robust objective function following the practice in Gu et al. (2020b).

A.3 Regularized linear regression

One big problem with simple linear regression is that the estimation results often becomes unreliable when there exist a large number of covariates. On one hand, in the high-dimensional setting, linear regression model estimated without regularization will be inconsistent.[50] On the other hand,

[49]We also consider the linear regression model equipped with the common l_2 loss function in our baseline analysis, which achieves very bad prediction performance as the corresponding R^2_{oos} is negative and its scale is more than an order of magnitude larger than other models. Therefore, we only report the results for linear regression model equipped with the Huber loss (OLS+H), and the results for OLS with l_2 loss is available upon request.

[50]Theoretically speaking, the high-dimensional setting refers to the scenario in which the number of covariates P grows with the number of observations NT, or $P > NT$. However, as pointed out in Gu et al. (2020b), it is reasonable to compare P with T, instead of with NT, due to the strong cross-sectional correlation between stock returns. Therefore, the prediction problem in this paper can also be understood as a quasi high-dimensional problem.

some covariates can be highly correlated, or even redundant, resulting in the multicollinearity problem and efficiency loss.

Fortunately, there is a well-established machine learning literature on regularized linear models which can well address these concerns. Popular methods include LASSO, the elastic net, Ridge regression, etc. In our empirical analysis, we include LASSO and the elastic net (Enet) as prediction models. The statistical properties of LASSO have been heavily studied in the literature for both *i.i.d.* and time series data.[51] The elastic net is a convex combination of LASSO and Ridge regression, and thus includes both as specially cases. These two methods share the same model specification with simple linear regression in Equation 5, while the main difference is that they shrink the coefficients of irrelevant covariates towards zero by imposing an extra penalty term to the original loss function. The objective function for LASSO is given by

$$L_H^{\text{LASSO}}(\theta) = L_H(\theta) + \lambda \sum_{j=1}^{P} |\theta_j|, \tag{7}$$

where λ is a hyperparameter which controls the size of the penalty. Furthermore, the objective function for Enet takes the form

$$L_H^{\text{Enet}}(\theta) = L_H(\theta) + (1-\rho)\lambda \sum_{j=1}^{P} |\theta_j| + \frac{1}{2}\rho\lambda \sum_{j=1}^{P} \theta_j^2, \tag{8}$$

where λ plays the same role as in Equation 7, and ρ determines the relative weight between l_1 and l_2 penalties. Following the convention, we do not penalize the intercept term θ_0 in both models. It is clearly that the objective function 8 degenerates to the one for LASSO when $\rho = 0$, and to the Ridge regression when $\rho = 1$. The tuning of λ and ρ are specified in Section A.1. For more details on the elastic net, we refer readers to the original paper Zou and Hastie (2005).

A.4 Partial least squares

Partial least squares (PLS) is a classic dimensional reduction technique which can effectively extract signals among a large number of covariates. This method often outperforms regularized linear models when covariates are highly correlated, which is a common attribute of stock-level

[51] See Bühlmann and Van De Geer (2011) for LASSO with *i.i.d.* data, and Medeiros and Mendes (2016) for time series data.

characteristics. Unlike LASSO and Enet, which directly penalize all covariates, PLS exploits the covariation between the predicted target and predictors by utilizing a model-averaging approach. We next briefly discuss the main context of the PLS regression. The PLS regression can be represented in its matrix form as follows

$$R = (ZW_K)\theta_K + \tilde{E}, \tag{9}$$

where R is the $NT \times 1$ vector of stock returns $r_{i,t+1}$, Z is the corresponding $NT \times P$ vector of stock-level characteristics, W_K is a $P \times K$ transformation matrix, θ_K is the $K \times 1$ vecor of model parameters, and $\tilde{E}$ is the $NT \times 1$ vector of residuals.[52] For PLS regression, K is the only tuning parameter, which is determined via the validation procedure. The transformation matrix W_K plays the crucial role, as it projects the original covariate matrix Z onto a K-dimensional linear space. With a proper transformation matrix, this method can reserve the useful information and rule out the noise as we have $K < P$ in general.

The main idea of PLS regression is to search for the transformation matrix which maximizes the correlation between the forecast target and the transformed covariates. In this sense, the columns of W_K, denoted by $w_1, ..., w_K$, solve a sequence of optimization problems, i.e.,

$$w_j = \arg\max_{w} \ \mathrm{Cov}^2(R, Zw), \quad \text{s.t.} \quad w^{'}w = 1, \quad \mathrm{Cov}(Zw, Zw_l) = 0, \quad l = 1, 2, ..., j-1, \tag{10}$$

for all $j = 1, ..., K$. In our analysis, we adopt the build-in algorithm in the *sklearn* package of Python for the calculation of W_K. Lastly, given a solution for W_K, θ_K can be easily estimated by regressing R on ZW_K.

A.5 Tree models

Tree models, including random forests (RF), gradient boosted regression trees (GBRT), and other variants, are very important machine learning techniques. They are fully nonparametric, and flexible enough to handle both classification and regression problems. These two methods, however, both build on a number of basic trees. Therefore, they can also be understood as different ensemble methods depending on the specific procedures taken.

[52] For the sake of notational simplicity, we assume the panel data is balanced when introducing our prediction models.

A basic tree is a set of decision rules which cluster observations into one of the multiple subgroups (partitions), usually named as "leaves". More precisely, the structure of a tree is determined by multiple decision nodes and the corresponding splitting variables. At each node, the tree chooses a splitting variable to generate two disjoint branches based on a splitpoint. The basic tree "grows" by sequentially developing branches until reaching the "leaves" (terminal nodes). Mathematically, a basic tree with K "leaves", and depth L, can be represented as

$$g(z_{i,t}; \theta, K, L) = \sum_{k=1}^{K} \theta_k \mathbf{1}_{\{z_{i,t} \in C_k(L)\}}, \tag{11}$$

where $C_k(L)$ is the k-th partition of the data, and the depth L is the largest number of nodes in a complete branch (from the top node to any terminal node). Suppose stock i with characteristics $z_{i,t}$ is clustered into the k-th leaf by the tree, then the basic tree will return θ_k as the predicted stock return. It is noteworthy that θ_k here is defined to be the sample average of outcomes within the k-th leaf based on the training data. There is an ample literature on selecting splitting variables and splitpoints for the basic tree model, and we refer readers to James et al. (2013) for an excellent description. Even though basic trees are fairly flexible, they are vulnerable to the overfitting problem, which often severely impairs their performance in practice. To address this problem, multiple regularization methods have been proposed in the literature, among which the most popular ones are gradient boosted regression trees (GBRT) and random forests (RF). We next briefly describe these two methods.

A.5.1 Gradient boosted regression trees

The main idea of GBRT is to produce a "strong learner" by recursively combining the prediction results from many "weak learners".[53] The GBRT model proceeds as follows. We start with a GBRT model with only two simple trees. First, we build a simple tree of depth L to fit stock returns based on stock-level characteristics. Next, a second simple tree of the same depth L is built to fit the residuals from the first tree. The forecast from the first tree plus the forecast from the second tree multiplied by the learning rate $v \in (0, 1)$ is the ensemble prediction of this basic GBRT model. To

[53] Simple tree models are usually thought as "weak learners", because they are prone to overfit the data and thus perform badly.

build a GBRT model with B trees, we simply build another $B - 2$ trees in the similar manner. At each step for the b-th tree, we grow a new tree to fit the residuals from the GBRT model with $b - 1$ trees, and add the product of its forecast and the learning rate to the previous final output to form the final output of the GBRT model with b trees. Repeat this step until B trees are grown.

It is noteworthy that there are three hyperparameters for the GBRT model, including the depth of trees L, the number of trees B, and the learning rate v. These parameters are determined based on the validation procedure specified in Section A.1.

A.5.2 Random forests

Random forest is another important ensemble method which also utilizes forecasts from many underlying simple trees. Unlike GBRT, random forests rely on a more general technique known as "bagging" (Breiman 2001) or bootstrap aggregation. The main idea of the method is to build B separate trees, and average over their forecasts to reduce prediction variation. More precisely, each of those B trees is trained on a bootstrapped sample of the original data, and uses only a randomly drawn subset of covariates for developing branches. Since each tree generated in this way is identically distributed, the expectation of the final output is the same as the expectation of a single tree. However, taking the average can significant benefit from reducing the variance while keeping the bias at a minimal level. Lastly, we note that there are also three tuning parameters for random forests, which include the depth of trees L, the number of trees B and the number of covariates for building simple trees. See Section A.1 for more details.

A.6 Variable subsample aggregation (VASA)

De Nard et al. (2020) introduce a new subsampling procedure, which they call VASA. VASA reduces the dimension by a subset selection of the predictors, similar as for LASSO or subset selection methods. However, it does not suffer as much from high variability and model-selection bias, as it averages over multiple subsampled (factor) model predictions. The conditional expected return is still assumed to be a linear function and is obtained by averaging over B OLS predictions, each trained on a (pseudo) random subset of the P predictors, i.e.,

$$g_i^{\text{VASA}}(z_{i,t}) \overset{\text{def}}{=} \sum_{b=1}^{B} \omega_b g_{i,b}(\tilde{z}_{i,t,b}) \overset{\text{def}}{=} \sum_{b=1}^{B} \omega_b(\alpha_{i,b} + \tilde{z}'_{i,t,b}\tilde{\beta}_{i,b}) \ , \tag{12}$$

where $\omega_b \in [0,1]$ is the weight of the b-th OLS prediction with $\sum_{b=1}^{B} \omega_b = 1$. Additionally, $\tilde{z}_{i,t,b}$ and $\tilde{\beta}_{i,b}$ are κ_b-dimensional vectors where κ_b represents the dimension (subsample size) of submodel b. We assume that the optimal subsampling size only depends on i and t but is constant across the submodels $\kappa \equiv \kappa_b$. The number of submodels B and their dimension κ are tuning parameters. Then, for each submodel $b = 1, \ldots, B$ the estimation problem is given by:

$$\underset{a_{i,b}, \tilde{b}_{i,b}}{\text{argmin}} \sum_{t=1}^{T} (r_{i,t+1} - a_{i,b} - \tilde{z}'_{i,t,b}\tilde{b}_{i,b})^2. \tag{13}$$

To compute the subsampling probabilities, we follow De Nard et al. (2020) and use the in-sample $R^2_{i,p}$ as variable importance measure. For more details, we refer to the original paper.

A.7 Neural networks

The final prediction method we introduce here is the neutral network, which is arguably the most popular machine learning technique in recent years. They have been widely applied for complex machine learning problems, such as computer vision, automated driving, and natural language processing. It is well-known that neural networks can approximate any smooth functions sufficiently well, which is ensured by the universal approximation theorem (Hornik et al. 1989). However, there is still much to learn about these models given the fact that they are among the least transparent and least interpretable machine learning methods. In fact, the name "neural networks" derives from the history that they were first developed as models for the human brain, which also remains highly mysterious given its complex structure.

For our analysis, we focus on the feed-forward multi-layer neural networks due to their excellent prediction performance for empirical asset pricing (Gu et al. 2020b). Similar to human's decision process, such models consist of an "input layer" for stock-level characteristics, one or more "hidden layers" that process the interactions between those predictors, and an "output layer" that generates a linear output. We consider neural networks with up to five hidden layers. Each layer consists of a

Table A.2. Number of Neurons and Parameters for All Neutral Network Models

Model	Hidden Layers	Number of Parameters
NN1	32	32(P+1) + 33
NN2	32, 16	32(P+1) + 545
NN3	32, 16, 8	32(P+1) + 673
NN4	32, 16, 8, 4	32(P+1) + 705
NN5	32, 16, 8, 4, 2	32(P+1) + 713

certain number of neurons, which are built with the commonly used rectified linear unit (ReLU), i.e., $\sigma(x) = \max(0, x)$.[54] As an illustration, the predicted stock return of the NN3 model can be written as

$$r_{i,t+1} = \alpha_1 + W_1\sigma(\alpha_2 + W_2\sigma(\alpha_3 + W_3\sigma(\alpha_4 + W_4 z_{i,t}))) + \epsilon_{i,t+1}, \tag{14}$$

where the activation function, $\sigma(\cdot)$, is applied elementwise, and $\{\alpha_1, ..., \alpha_4, W_1, ...W_4\}$ is the set of biases and weight matrices to be estimated. The network architectures and the corresponding numbers of parameters of NN1 - NN5 are summarized in Table A.2.

In our analysis, all neural networks are trained using TensorFlow, a powerful machine learning system. Similar as in De Nard et al. (2020), we adopt Adam optimization algorithm (Kingma and Ba 2014), early stopping, batch normalization (Ioffe and Szegedy 2015), ensembles, and dropout (Srivastava et al. 2014) when training our models.

B Details on statistical inference methods

We briefly introduce the statistical inference methods used in the main text for comparing the performance of different prediction models. Here we mainly focus on implementing two statistical tests: the unconditional superior predictability test (Hansen 2005) and the conditional superior predictability test (Li, Liao and Quaedvlieg 2020). For readers who are interested in the technical details, we refer them to the original papers. To facilitate the discussion, we first introduce some notations following the convention in the forecast evaluation literature. Let $\{F_t^\dagger\}_{t\geq 1}$ denote the

[54]ReLU is often preferred to other alternative activation functions, such as sigmoid, hyperbolic, and softmax, because it overcomes the vanishing gradient problem and its derivative is easy to calculate.

variable of interest to be predicted, which in our case is the monthly stock price. We want to compare the performance of J competing models relative to a benchmark model in terms of the forecast $\{F_t^\dagger\}_{t\geq 1}$. Let $\{F_{0,t}\}_{t=1}^n$ and $\{F_{j,t}\}_{t=1}^n$ for $1 \leq j \leq J$ denote the predicted values of the variable of interest in n periods. For a given loss function $L(\cdot,\cdot)$, we let $Y_{j,t}$ denote the performance of the j-th competing model relative to the benchmark model in period t, i.e.,

$$Y_{j,t} = L(F_t^\dagger, F_{j,t}) - L(F_t^\dagger, F_{0,t}). \tag{15}$$

It is noteworthy that the benchmark model has a better prediction performance in period t if $Y_{j,t} \geq 0$.

B.1 Unconditional superior predictive ability test

The unconditional superior predictive ability (USPA) test, which is first studied in White (2000) and later refined by Hansen (2005), has been widely used in finance research. In this paper, we adopt the refined version in Hansen (2005) instead of the original reality check (RC) for data snooping in White (2000), because the former usually has larger statistical power. The USPA test directly compares the unconditional (average) performance of different competing prediction models relative to the benchmark model, with the null hypothesis given by

$$H_0^{USPA} : \mathbb{E}[Y_{j,t}] \geq 0, \quad \text{for all } 1 \leq j \leq J.$$

It is noteworthy that both the USPA test and CSPA test build on the probabilistic properties of $Y_{j,t}$ directly instead of the underlying loss function and prediction models. Let $\mathbf{Y}_t = (Y_{1,t}, ..., Y_{J,t})'$, $\boldsymbol{\mu} = E[\mathbf{Y}_t]$, $\bar{\mathbf{Y}} = n^{-1}\sum_{t=1}^n \mathbf{Y}_t$ and $\bar{Y}_j = n^{-1}\sum_{t=1}^n Y_{j,t}$ for $j = 1, ..., J$. Under some regularity conditions, one can show that

$$n^{1/2}(\bar{\mathbf{Y}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Omega}), \tag{16}$$

where $\boldsymbol{\Omega}$ is the asymptotic covariance matrix of $n^{1/2}(\bar{\mathbf{Y}} - \boldsymbol{\mu})$. Based on this result, Hansen (2005) proposed to use the following test statistic

$$T_n^{USPA} = \min\Big[\min_{j=1,\ldots,J} \frac{n^{1/2}\bar{Y}_j}{\hat{\omega}_j},\ 0\Big],$$

in which the estimator of the asymptotic variance of $\hat{\omega}_j^2$ is given by

$$\hat{\omega}_j^2 = \hat{\gamma}_{0,j} + 2\sum_{i=1}^{n-1} K(n,i)\hat{\gamma}_{i,j},$$

where $\hat{\gamma}_{i,j} = n^{-1}\sum_{l=1}^{n-i}(Y_{j,l} - \bar{Y}_j)((Y_{j,l+i} - \bar{Y}_j))$, $i = 0, 1, ..., n-1$ and $K(n,i)$ is the kernel weight for stationary bootstrap (see Hansen (2005) for more details). To invoke a null distribution that is based on Equation 16, one also needs to select a consistent estimator of $\boldsymbol{\mu}$, for which Hansen (2005) recommends to use $\hat{\mu}_j = \bar{Y}_j 1\{n^{1/2}\bar{Y}_j \geq \sqrt{2\log\log n}\}$. Intuitively, the null hypothesis will be rejected if T_n^{USPA} is smaller than the critical value at the given significance level. However, the critical value can not be calculated analytically in general and the bootstrap method is recommended. For reader's convenience, the implementation steps of the test is presented as follows.

Algorithm for the USPA test (Hansen 2005)

Step 1. Generate B resembles $\{\mathbf{Y}_t^{(b)} : t = 1, ..., n\}_{b=1}^{B}$ from $\{\mathbf{Y}_t : t = 1, ..., n\}$ using the stationary bootstrap method in Hansen (2005).[55]

Step 2. For each resample, calculate $T_{(b),n}^{USPA} = \min[\min_{j=1,...,J}(n^{1/2}\bar{Y}_j/\hat{\omega}_j),\ 0]$.

Step 3. Calculate the bootstrap p-value: $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} 1\{T_{(b),n}^{USPA} < T_n^{USPA}\}$. Reject the USPA null hypothesis at significance level α if $\hat{p} < \alpha$.

B.2 Conditional superior predictive ability test

One potential limitation of the USPA test is that two prediction models may have an identical performance on average but can behave very differently under different macroeconomic (market-level) conditions. The USPA test cannot detect such differences because it only focuses on the average (unconditional) relative performance of prediction models. Motivated by this observation, Li, Liao and Quaedvlieg (2020) developed an innovative conditional superior predictive ability (CSPA) test based on the recent development of the series estimation literature (Chernozhukov et al. 2013, Li, Liao and Quaedvlieg 2020). The null hypothesis of the CSPA test is given by

$$H_0^{CSPA} : \mathbb{E}[Y_{j,t}|X_t = x] \geq 0, \quad \text{for all } x \in \mathcal{X}, 1 \leq j \leq J,$$

[55]In this paper, we set $B = 10,000$.

where X_t is the variable relates to the macroeconomic (market-level) conditions, such as the GDP growth rate, the macroeconomic uncertainty and the volatility index (VIX), and $\mathcal{X}$ is the support of X_t. It is worth noting that $\mathcal{X}$ should be compact, while the compactness condition can be trivially satisfied by implementing any one-to-one transformations in Li, Liao and Gao (2020). The main idea of the CSPA test is to directly estimate the conditional expectation functions $\mathbb{E}[Y_{j,t}|X_t = x]$ using the series estimation method and conduct uniform inference with the help of the strong approximation theory for time series data. Following the notation in Li, Liao and Quaedvlieg (2020), we let $h_{j,n} = \mathbb{E}[Y_{j,t}|X_t = x]$ for all $1 \leq j \leq J$ and $P(x) = (p_1(x), ..., p_{m_n})^\top$ be an $m_n \times 1$ vector of basis functions, such as Legendre polynomial, power series, spline, etc. Here m_n is the number of basis terms used to estimate $h_{j,n}$, which is a divergent number depends on the sample size n. The conditional expectation functions can then be estimated by a OLS-type regression

$$\hat{h}_{j,n}(x) = P(x)^\top \hat{b}_{j,n}, \tag{17}$$

where

$$\hat{b}_{j,n} = \Big(n^{-1}\sum_{t=1}^{n} P(X_t)P(X_t)^\top\Big)^{-1}\Big(n^{-1}\sum_{t=1}^{n} P(X_t)Y_{j,t}\Big).$$

To conduct statistical inference on $\hat{h}_{j,n}$, one need to derive its asymptotic distribution, which lies in the asymptotic properties of $\hat{b}_{j,n}$. However, as pointed out in Li, Liao and Quaedvlieg (2020), this is not a standard problem since the dimension of $\hat{b}_{j,n}$ divergences with the sample size n, making the classical central limit theorem invalid. The solution to such inference problem is to utilize the strong approximation theory for growing-dimensional statistics, which is first developed in Chernozhukov et al. (2013) and later generalized to time series data in Li and Liao (2020). Let $u_t = (u_{1,t}, ..., u_{J,t})^\top$ denote the vector of nonparametric regression error term, where $u_{j,t} = Y_{j,t} - h_j(X_t)$. Li and Liao (2020) and Li, Liao and Quaedvlieg (2020) have managed to show that $n^{-1/2}\sum_{t=1}^{n} u_t \otimes P(X_t)$,[56] which is the random part of $\hat{H}_n(x) = (\hat{h}_{1,n}(x), ..., \hat{h}_{J,n}(x))^\top$ conditional on X_t, can be approximated sufficiently well by some Jm_n-dimensional Gaussian vector $\tilde{N} \sim \mathcal{N}(0, A_n)$, where the covariance matrix A_n is given by

$$A_n = \mathrm{Var}\Big(n^{-1/2}\sum_{t=1}^{n} u_t \otimes P(X_t)\Big). \tag{18}$$

[56] $\otimes$ denotes the Kronecker product.

For the estimation of A_n, Li, Liao and Quaedvlieg (2020) recommend to use the pre-whitened HAC estimator and we follow their suggestion when implementing the test. Given the estimated covariance matrix $\hat{A}_n$, the $Jm_n \times Jm_n$ covariance matrix of the estimators $(n^{1/2}(\hat{b}_{1,n} - b^*_{1,n})^\top, ..., n^{1/2}(\hat{b}_{J,n} - b^*_{J,n})^\top)^\top$[57] is given by

$$\hat{\Omega}_n = \Big(I_J \otimes \hat{Q}_n\Big)^{-1} \hat{A}_n \Big(I_J \otimes \hat{Q}_n\Big)^{-1}, \tag{19}$$

where I_J is the $J \times J$ identity matrix and $\hat{Q}_n = (n^{-1} \sum_{t=1}^{n} P(X_t)P(X_t)^\top)^{-1}$. It follows that the standard deviation function of $n^{1/2}(\hat{h}_{j,n}(x) - h_j(x))$ can be estimated by

$$\hat{\sigma}_{j,n}(x) = \Big(P(x)^\top \hat{\Omega}(j,j) P(x)\Big)^{1/2}, \tag{20}$$

where $\hat{\Omega}(j,j)$ is the diagonal submatrix of $\hat{\Omega}_n$ which corresponds to $n^{1/2}(\hat{b}_{j,n} - b^*_{j,n})^\top$. Then we can conduct statistical inference based on $\hat{h}_{j,n}(x)$ for all $1 \le j \le J$ and $x \in \mathcal{X}$. For reader's convenience, we borrow the following implementation algorithm directly from Li, Liao and Quaedvlieg (2020).

Algorithm for the CSPA test (Li, Liao and Quaedvlieg 2020)

Step 1. Simulate a Jm_n-dimensional Gaussian vectors $\xi = (\xi_1^\top, ..., \xi_J^\top)^\top \sim \mathcal{N}(0, \hat{\Omega}_n)$, where $\xi_j^\top$ is m_n-dimensional and $\hat{\Omega}$ is defined in Equation 19. Repeat this step for B times and store $\{\xi^{(1)}, ..., \xi^{(B)}\}$.

Step 2. For each $\xi^{(b)}$, $1 \le b \le B$, calculate $\hat{t}^{(b)}_{j,n}(x) = P(x)^\top \xi_j^{(b)}$ for all $1 \le j \le J$. Set $\tilde{\gamma}_n = 1 - 0.1/\log(n)$. Let $\hat{K}_n$ be the $\tilde{\gamma}_n$-quantile of $\max_{1\le j\le J} \sup_{x\in\mathcal{X}} \hat{t}_{j,n}(x)$ based on the sample $\{\hat{t}^{(b)}_{j,n}(x)\}_{1\le b\le B, 1\le j\le J}$.

Step 3. Set

$$\hat{\mathcal{V}}_n = \Big\{(j,x) : \hat{h}_{j,n}(x) \le \min_{1\le j\le J} \inf_{x\in\mathcal{X}} \Big(\hat{h}_{j,n}(x) + n^{-1/2}\hat{K}_n\hat{\sigma}_{j,n}(x)\Big) + 2n^{-1/2}\hat{K}_n\hat{\sigma}_{j,n}(x)\Big\},$$

where $\hat{\sigma}_{j,n}(x)$ is defined in Equation 20.

Step 4. Set $\hat{k}_{n,1-\alpha}$ to be the $(1-\alpha)$-quantile of $\sup_{(j,x)\in\hat{\mathcal{V}}_n} \hat{t}_{j,n}$ and calculate

$$\hat{\eta}_{n,1-\alpha} = \min_{1\le j\le J} \inf_{x\in\mathcal{X}} \Big(\hat{h}_{j,n}(x) + n^{-1/2}\hat{k}_{n,1-\alpha}\hat{\sigma}_{j,n}(x)\Big).$$

Reject the CSPA null hypothesis at significance level α if $\hat{\eta}_{n,1-\alpha} < 0$.

[57] Here $b^*_{j,n}$ is some constant vector such that $P(x)b^*_{j,n}$ can uniformly approximate $h_j(x)$ sufficiently well on the support of x. For more details, see Li, Liao and Quaedvlieg (2020).

When implementing the CSPA test in the main text, we set $B = 10000$ and apply the Akaike Information Criterion (AIC) to select the basis functions following the procedure in Li, Liao and Quaedvlieg (2020). One beneficial byproduct of the CSPA test is that it also allows us to learn exactly what are the macroeconomic (market-level) conditions under which the benchmark model outperforms (or does not outperform) the alternatives by plotting the conditional expectation functions $h_{j,n}(x)$ for diagnosis. In this paper, we also conduct these standard exercises when comparing the performance of different prediction models.

C Stock characteristics

C.1 Variable list

Table C.1. Details on stock characteristics

No.	Acronym	Stock Characteristics	Author(s)	Year, Journal	Data Source	Frequency
1	*absacc*	Absolute accruals	Bandyopadhyay, Huang & Wirjant	2010, WP	CSMAR	Semi-annual
2	*acc*	Working capital accruals	Sloan	1996, TAR	CSMAR	Semi-annual
3	*agr*	Asset growth	Cooper, Gulen & Schill	2008, JF	CSMAR	Annual
4	*beta*	Beta	Fama & MacBeth	1973, JPE	WIND	Monthly
5	*betasq*	Beta squared	Fama & MacBeth	1973, JPE	CSMAR	Monthly
6	*bm*	Book-to-market	Rosenberg, Reid & Lanstein	1985, JPM	CSMAR	Quarterly
7	*bm_ia*	Industry-adjusted book to market	Asness, Porter & Stevens	2000, WP	CSMAR	Quarterly
8	*cash*	Cash holdings	Palazzo	2012, JFE	CSMAR	Quarterly
9	*cashdebt*	Cash flow to debt	Ou & Penman	1989, JAE	CSMAR	Quarterly
10	*cashspr*	Cash productivity	Chandrashekar & Rao	2009, WP	CSMAR	Quarterly
11	*cfp*	Cash flow to price ratio	Desai, Rajgopal & Venkatachalam	2004, TAR	CSMAR	Quarterly
12	*cfp_ia*	Industry-adjusted cash flow to price ratio	Asness, Porter & Stevens	2000, WP	CSMAR	Quarterly
13	*chato*	Change in asset turnover	Soliman	2008, TAR	CSMAR	Quarterly
14	*chatoia*	Industry-adjusted change in asset turnover	Soliman	2008, TAR	CSMAR	Quarterly
15	*chcsho*	Change in shares outstanding	Pontiff & Woodgate	2008, JF	CSMAR	Monthly
16	*chempia*	Industry-adjusted change in employees	Asness, Porter & Stevens	1994, WP	CSMAR	Annual
17	*chinv*	Change in inventory	Thomas & Zhang	2002, RAS	CSMAR	Quarterly
18	*chmom*	Change in 6-month momentum	Gettleman & Marks	2006, WP	WIND	Monthly
19	*chpm*	Change in profit margin	Soliman	2008, TAR	CSMAR	Quarterly
20	*chpmia*	Industry-adjusted change in profit margin	Soliman	2008, TAR	CSMAR	Quarterly
21	*chtx*	Change in tax expense	Thomas & Zhang	2011, JAR	CSMAR	Quarterly
22	*cinvest*	Corporate investment	Titman, Wei & Xie	2004, JFQA	CSMAR	Quarterly
23	*currat*	Current ratio	Ou & Penman	1989, JAE	CSMAR	Quarterly
24	*depr*	Depreciation / PP&E	Holthausen & Larcker	1992, JAE	CSMAR	Semi-annual
25	*divi*	Dividend initiation	Michaely, Thaler & Womack	1995, JF	CSMAR	Annual
26	*divo*	Dividend omission	Michaely, Thaler & Womack	1995, JF	CSMAR	Annual
27	*dolvol*	Yuan trading volume	Chordia, Subrahmanyam & Anshuman	2001, JFE	CSMAR	Monthly
28	*dy*	Dividend to price	Litzenberger & Ramaswamy	1982, JF	CSMAR	Annual
29	*ear*	Earnings announcement return	Kishore, Brandt, Santa-Clara & Venkatachalam	2008, WP	CSMAR	Quarterly
30	*egr*	Growth in common shareholder equity	Richardson, Sloan, Soliman & Tuna	2005, JAE	CSMAR	Quarterly
31	*gma*	Gross profitability	Novy-Marx	2013, JFE	CSMAR	Quarterly
32	*grCAPX*	Growth in capital expenditures	Anderson & Garcia-Feijoo	2006, JF	CSMAR	Semi-annual
33	*herf*	Industry sales concentration	Hou & Robinson	2006, JF	CSMAR	Quarterly
34	*hire*	Employee growth rate	Bazdresch, Belo & Lin	2014, JPE	CSMAR	Annual
35	*idiovol*	Idiosyncratic return volatility	Ali, Hwang & Trombley	2003, JFE	WIND	Monthly

Table C.1: Details on Stock Characteristics (Continued)

No.	Acronym	Stock Characteristics	Author(s)	Year, Journal	Data Source	Frequency
36	*ill*	Illiquidity	Amihud	2002, JFM	WIND	Monthly
37	*invest*	Capital expenditures and inventory	Chen & Zhang	2010, JF	CSMAR	Annual
38	*lev*	Leverage	Bhandari	1988, JF	CSMAR	Quarterly
39	*lgr*	Growth in long-term debt	Richardson, Sloan, Soliman & Tuna	2005, JAE	CSMAR	Quarterly
40	*maxret*	Maximum daily return	Bali, Cakici & Whitelaw	2011, JFE	WIND	Monthly
41	*mom12m*	12-month momentum	Jegadeesh	1990, JF	WIND	Monthly
42	*mom1m*	1-month momentum	Jegadeesh & Titman	1993, JF	WIND	Monthly
43	*mom6m*	6-month momentum	Jegadeesh & Titman	1993, JF	WIND	Monthly
44	*mom36m*	36-month momentum	Jegadeesh & Titman	1993, JF	WIND	Monthly
45	*ms*	Financial statement score	Mohanram	2005, RAS	CSMAR	Annual
46	*mve*	Size	Banz	1981, JFE	CSMAR	Monthly
47	*mve_ia*	Industry-adjusted size	Asness, Porter & Stevens	2000, WP	CSMAR	Monthly
48	*nincr*	Number of earnings increases	Barth, Elliott & Finn	1999, JAR	CSMAR	Quarterly
49	*operprof*	Operating profitability	Fama & French	2015, JFE	CSMAR	Quarterly
50	*orgcap*	Organizational capital	Eisfeldt & Papanikolaou	2013, JF	CSMAR	Quarterly
51	*pchcapx_ia*	% Industry adjusted % change in capital expenditures	Ou & Penman	1989, JAE	CSMAR	Annual
52	*pchcurrat*	% change in current ratio	Ou & Penman	1989, JAE	CSMAR	Quarterly
53	*pchdepr*	% change in depreciation	Holthausen & Larcker	1992, JAE	CSMAR	Semi-annual
54	*pchgm_pchsale*	% change in gross margin - % change in sales	Abarbanell & Bushee	1998, TAR	CSMAR	Quarterly
55	*pchquick*	% change in quick ratio	Ou & Penman	1989, JAE	CSMAR	Quarterly
56	*pchsale_pchinvt*	% change in sales - % change in inventory	Abarbanell & Bushee	1998, TAR	CSMAR	Quarterly
57	*pchsale_pchrect*	% change in sales - % change in A/R	Abarbanell & Bushee	1998, TAR	CSMAR	Quarterly
58	*pchsale_pchxsga*	% change in sales - % change in SG&A	Abarbanell & Bushee	1998, TAR	CSMAR	Quarterly
59	*pchsaleinv*	% change sales-to-inventory	Ou & Penman	1989, JAE	CSMAR	Quarterly
60	*pctacc*	Percent accruals	Hafzalla, Lundholm & Van Winkle	2011, TAR	CSMAR	Semi-annual
61	*pricedelay*	Price delay	Hou & Moskowitz	2005, RFS	WIND	Monthly
62	*ps*	Financial statements score	Piotroski	2000, JAR	CSMAR	Quarterly
63	*quick*	Quick ratio	Ou & Penman	1989, JAE	CSMAR	Quarterly
64	*rd*	R&D increase	Eberhart, Maxwell & Siddique	2004, JF	CSMAR	Quarterly
65	*rd_mve*	R&D to market capitalization	Guo, Lev & Shi	2006, JBFA	CSMAR	Quarterly
66	*rd_sale*	R&D to sales	Guo, Lev & Shi	2006, JBFA	CSMAR	Quarterly
67	*realestate*	Real estate holdings	Tuzel	2010, RFS	CSMAR	Quarterly
68	*volatility*	Return volatility	Ang, Hodrick, Xing & Zhang	2006, JF	WIND	Monthly
69	*roaq*	Return on assets	Balakrishnan, Bartov & Faurel	2010, JAE	CSMAR	Quarterly
70	*roavol*	Earnings volatility	Francis, LaFond, Olsson & Schipper	2004, TAR	CSMAR	Quarterly

Table C.1: Details on Stock Characteristics (Continued)

No.	Acronym	Stock Characteristics	Author(s)	Year, Journal	Data Source	Frequency
71	*roeq*	Return on equity	Hou, Xue & Zhang	2015, RFS	CSMAR	Quarterly
72	*roic*	Return on invested capital	Brown & Rowe	2007, WP	CSMAR	Quarterly
73	*rsup*	Revenue surprise	Kama	2009, JBFA	CSMAR	Quarterly
74	*salecash*	Sales to cash	Ou & Penman	1989, JAE	CSMAR	Quarterly
75	*saleinv*	Sales to inventory	Ou & Penman	1989, JAE	CSMAR	Quarterly
76	*salerev*	Sales to receivables	Ou & Penman	1989, JAE	CSMAR	Quarterly
77	*sgr*	Sales growth	Lakonishok, Shleifer & Vishny	1994, JF	CSMAR	Quarterly
78	*sp*	Sales to price	Barbee, Mukherji, & Raines	1996, FAJ	CSMAR	Quarterly
79	*std_dolvol*	Volatility of liquidity (yuan trading volume)	Chordia, Subrahmanyam & Anshuman	2001, JFE	CSMAR	Monthly
80	*std_turn*	Volatility of liquidity (share turnover)	Chordia, Subrahmanyam, & Anshuman	2001, JFE	CSMAR	Monthly
81	*stdacc*	Accrual volatility	Bandyopadhyay, Huang & Wirjanto	2010, WP	CSMAR	Quarterly
82	*stdcf*	Cash flow volatility	Huang	2009, JEF	CSMAR	Quarterly
83	*tang*	Debt capacity/firm tangibility	Almeida & Campello	2007, RFS	CSMAR	Quarterly
84	*tb*	Tax income to book income	Lev & Nissim	2004, TAR	CSMAR	Quarterly
85	*turn*	Share turnover	Datar, Naik & Radcliffe	1998, JFM	CSMAR	Monthly
86	*zerotrade*	Zero trading days	Liu	2006, JFE	CSMAR	Monthly
87	*atr*	Abnormal Turnover Ratio	Pan, Tang & Xu	2015, RF	WIND, CSMAR	Monthly
88	*er_trend*	Trend factor	Liu, Zhou & Zhu	2020, WP	WIND, CSMAR	Monthly
89	*largestholderrate*	Largest shareholder ownership	Gul, Kim & Qiu	2010, JFE	CSMAR	Annual
90	*top10holderrate*	Top 10 shareholders ownership	Gul, Kim & Qiu	2010, JFE	CSMAR	Annual
91	*soe*	State owned enterprise indicator			CSMAR	Annual
92	*private*	Private enterprise indicator			CSMAR	Annual
93	*foreign*	Foreign enterprise indicator			CSMAR	Annual
94	*others*	Other enterprise indicator			CSMAR	Annual

C.2 Variable construction

We closely follow the definitions in Green et al. (2017) and the original papers to construct the stock-level characteristics.

(1) ***acc***: We follow the definition of accruals in Sloan (1996) to construct *acc*, i.e.,

$$acc = [(\Delta\text{CA} - \Delta\text{CASH}) - (\Delta\text{CL} - \Delta\text{STD} - \Delta\text{TP}) - \text{Dep}]/\text{Total Assets},$$

where Δ represents the difference between two consecutive periods, CA, CASH, CL, STD, TP, Dep denote current assets, cash/cash equivalents, current liabilities, debt included in current liabilities, income tax payable, depreciation and amortization expense, respectively. These data are acquired from CSMAR.

(2) ***absacc***: Absolute value of *acc*.

(3) ***agr***: Annual percent change in total assets. Data of total assets are acquired from CSMAR.

(4) ***beta***: We estimate stock-level beta using weekly returns and value weighted market returns for 3 years ending month $t-1$ with at least 52 weeks of returns. Stock returns are acquired from WIND database.

(5) ***betasq***: Stock-level market beta squared.

(6) ***bm***: Book-to-market ratio, which equals the book value of equity divided by market capitalization. Data are acquired from CSMAR.

(7) ***bm_ia***: This is the industry-adjusted book-to-market ratio introduced in Asness et al. (2000),

$$bm\_ia_{it} = bm_{it} - bm_{Iit},$$

where bm_{Iit} is the equally weighted average book-to-market ratio of firms in firm i's industry. As firms' industries are reported annually in CSMAR, we let firm i's current industry to be the one reported in the year prior to the current month.

(8) ***cash***: Cash and cash equivalents divided by average total assets. Related data are reported in quarterly reports and are acquired from CSMAR.

(9) ***cashdebt***: Earnings divided by total liabilities, which is defined similar to that in Ou and Penman (1989). Data are acquired from CSMAR.

(10) ***cashspr***: Cash productivity, which is defined as quarter-end market capitalization plus long-term debt minus total assets divided by cash and equivalents. Related data are contained in quarterly reports and acquired from CSMAR.

(11) ***cfp***: Operating cash flows divided by quarter-end market capitalization. Related data are contained in quarterly reports and acquired from CSMAR.

(12) ***cfp_ia***: This is the industry-adjusted operating cash flows. The way of adjustment is similar to that for ***bm_ia***. Data are acquired from CSMAR.

(13) ***chato***: Change in sales divided by average total assets. Quarterly data on sales and total assets are acquired from CSMAR.

(14) ***chato_ia***: Industry-adjusted change in sales divided by average total assets. Data are acquired from CSMAR.

(15) ***chcsho***: Monthly percent change in shares outstanding. Monthly data on shares outstanding are acquired from CSMAR.

(16) ***chempia***: Industry-adjusted change in the number of employees. Related data are available annually on CSMAR and the way of industry adjustment is similar as that for ***bm_ia***.

(17) ***chinv***: Change in inventory scaled by total assets. Data are available quarterly on CSMAR.

(18) ***chmom***: Cumulative returns from months $t-6$ to $t-1$ minus months $t-12$ to $t-7$. Stock returns are acquired from WIND database.

(19) ***chpm***: Change in income before extraordinary items scaled by sales. Related data are acquired from CSMAR.

(20) ***chpm_ia***: Industry-adjusted change in income before extraordinary items scaled by sales. Data are acquired from CSMAR.

(21) ***chtx***: Percent change in taxes from quarter $t-1$ to t. Data are acquired from CSMAR.

(22) ***cinvest***: Change over one quarter in fixed assets divided by sales - average of this variable for prior 3 quarters; if sales are zero, then scale by 0.01. Data are acquired from CSMAR.

(23) ***currat***: The ratio of current assets to current liabilities. Data are acquired from CSMAR.

(24) ***depr***: Depreciation divided by fixed assets. Data are acquired from CSMAR.

(25) ***divi***: A dummy variable that equals to 1 if company pays dividends this year but did not in prior year. Data are acquired from CSMAR.

(26) ***divo***: A dummy variable that equals to 1 if company does not pay dividends this year but did in prior year. Data are acquired from CSMAR.

(27) ***dolvol***: Natural logarithm of trading volume times price per share from month t-2. Data are acquired from CSMAR.

(28) ***dy***: Total dividends divided by market capitalization at year end. Data are acquired from CSMAR.

(29) ***ear***: Sum of daily returns in three days around earnings announcement. Data are acquired from CSMAR.

(30) ***egr***: Quarterly percent change in book value of equity. Data are acquired from CSMAR.

(31) ***gma***: Revenue minus cost of goods sold divided by lagged total assets. Quarterly data are acquired from CSMAR.

(32) ***grCAPX***: Percent change in capital expenditures from year $t-2$ to year t. Data are acquired from CSMAR.

(33) ***herf***: Sum of squared percent sales in industry for each company. Sales data and industry code are acquired from CSMAR.

(34) ***hire***: Percent change in number of employees. Related data are acquired from CSMAR.

(35) ***idiovol***: Standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end. Data are acquired from WIND database.

(36) ***ill***: Average of daily (absolute return/RMB volume) in month t. Daily data are acquired from WIND database.

(37) ***invest***: The sum of annual change in fixed assets and annual change in inventories divided by lagged total assets. Data are acquired from CSMAR.

(38) ***lev***: Total liabilities divided by quarter-end market capitalization. Quarterly data are acquired from CSMAR.

(39) ***lgr***: Quarterly percent change in total liabilities. Data are acquired from CSMAR.

(40) ***maxret***: Maximum daily return from returns during month $t-1$. Daily returns are acqired from WIND database.

(41) ***mom12m***: 11-month cumulative returns ending one month before month end. Stock returns are acqired from WIND database.

(42) ***mom1m***: 1-month cumulative return. Stock returns are acqired from WIND database.

(43) ***mom6m***: 5-month cumulative returns ending one month before month end. Stock returns are acqired from WIND database.

(44) ***mom36m***: Cumulative returns from months $t-36$ to $t-13$. Stock returns are acqired from WIND database.

(45) ***ms***: Sum of 8 indicator variables for fundamental performance following the corresponding definitions in Mohanram (2005). Data are acquired from CSMAR.

(46) ***mve***: Natural log of market capitalization at end of month $t-1$. Related data are acquired from CSMAR.

(47) ***mve_ia***: Industry adjusted natural log of market capitalization at end of month $t-1$. Related data are acquired from CSMAR.

(48) ***nincr***: Number of consecutive quarters (up to eight quarters) with an increase in earnings. Earnings data are acquired from CSMAR.

(49) ***operprof***: Quarterly operating profit divided by lagged common shareholders' equity. Related data are acquired from CSMAR.

(50) ***orgcap***: Capitalized management expenses. This characteristic uses expense data acquired from CSMAR and is constructed according to the definition in Eisfeldt and Papanikolaou (2013). Data are acquired from CSMAR.

(51) ***pchcapx_ia***: Industry adjusted percentage change in capital expenditure. Data are acquired from CSMAR.

(52) ***pchcurrat***: Percentage change in current ratio (current liabilities divided by current assets). Data are acquired from CSMAR.

(53) ***pchdepr***: Percentage change in depreciation. Data are acquired from CSMAR.

(54) ***pchgm_pchsale***: Percentage change in gross margin minus Percentage change in sales. Data are acquired from CSMAR.

(55) ***pchquick***: Percentage change in quick ratio. Data are acquired from CSMAR.

(56) ***pchsale_pchinvt***: Quarterly percentage change in sales minus quarterly percentage change in inventory. Data are acquired from CSMAR.

(57) ***pchsale_pchrect***: Quarterly percentage change in sales minus quarterly percentage change in receivables. Data are acquired from CSMAR.

(58) ***pchsale_pchxsga***: Quarterly percentage change in sales minus quarterly percentage change in management expenses. Data are acquired from CSMAR.

(59) ***pchsaleinv***: Quarterly percentage change in sales-to-inventory. Data are acquired from CSMAR.

(60) ***pctacc***: Same as *acc* except that the numerator is divided by the absolute value of net income; if net income = 0 then net income set to 0.01 for denominator. Data are acquired from CSMAR.

(61) ***pricedelay***: The proportion of variation in weekly returns for 36 months ending in month t explained by 4 lags of weekly market returns incremental to contemporaneous market return. Stock returns are acquired from WIND database.

(62) ***ps***: Sum of 9 indicator variables that are defined similarly as in Piotroski (2000). Related data are acquired from CSMAR.

(63) ***quick***: Quick ratio = (current assets - inventory) / current liabilities. Data are acquired from CSMAR.

(64) ***rd***: An indicator variable equal to 1 if R&D expense as a percentage of total assets has an increase greater than 5%. Data are acquired from CSMAR.

(65) ***rd_mve***: R&D expense divided by end-of-quarter market capitalization. Data are acquired from CSMAR.

(66) ***rd_sale***: R&D expense divided by quarterly sales. Data are acquired from CSMAR.

(67) ***realestate***: Investment real estates divided by fixed assets. Data are acquired from CSMAR.

(68) ***volatility***: Standard deviation of daily returns from month $t-1$. Stock returns are acquired from WIND.

(69) ***roaq***: Income before extraordinary items divided by one quarter lagged total assets. Related data are acquired from CSMAR.

(70) ***roavol***: Standard deviation of 16 quarters of income before extraordinary items divided by average total assets. Data are acquired from CSMAR.

(71) ***roeq***: Income before extraordinary items divided by lagged common shareholders' equity. Related data are acquired from CSMAR.

(72) ***roic***: Quarterly earnings before interest and taxes minus nonoperating income divided by non-cash enterprise value. Related data are acquired from CSMAR.

(73) ***rsup***: Sales from quarter t minus sales from quarter $t-1$ divided by quarter-end market capitalization. Related data are acquired from CSMAR.

(74) ***salecash***: Quarterly sales divided by cash and cash equivalents. Data are acquired from CSMAR.

(75) ***saleinv***: Quarterly sales divided by total inventory. Data are acquired from CSMAR.

(76) ***salerev***: Quarterly sales divided by accounts receivable. Data are acquired from CSMAR.

(77) ***sgr***: Quarterly percentage change in sales. Data are acquired from CSMAR.

(78) ***sp***: Quarterly sales divided by quarter-end market capitalization. Data are acquired from CSMAR.

(79) ***std_dolvol***: Monthly standard deviation of daily RMB trading volume. Data are acquired from CSMAR.

(80) ***std_turn***: Monthly standard deviation of daily share turnover. Data are acquired from CSMAR.

(81) ***stdacc***: Standard deviation of 16 quarters of accruals from month $t-16$ to $t-1$. Data are acquired from CSMAR.

(82) ***stdcf***: Standard deviation for 16 quarters of net cash flows divided by sales. Data are acquired from CSMAR.

(83) ***tang***: Cash holdings + 0.715 $\times$ receivables + 0.547$\times$ inventory + 0.535 $\times$ fixed assets/total assets. Data are acquired from CSMAR.

(84) ***tb***: Tax income, defined as current tax expense divided by enterprise income tax rate in China (25%), divided by total income. Data are acquired from CSMAR.

(85) ***turn***: Average monthly trading volume for month $t-3$ to $t-1$ scaled by number of shares outstanding in month t. Related data are acquired from CSMAR.

(86) ***zerotrade***: Turnover weighted number of zero trading days in month $t-1$. Related data are acquired from CSMAR.

(87) ***atr***: The abnormal turnover ratio (atr) is constructed following the definition in Pan et al. (2016). Specifically, for stock i in month t, we run the following regression using daily data from month $t-7$ to $t-1$,

$$\mathrm{DTR}_{i,t} = \beta_1 + \beta_2 \times \mathrm{DMRT}_t + \sum_{j=1}^{K} c_j \times \mathrm{Dummy\_Event}(j)_{i,t} + \epsilon_{i,t},$$

where $\mathrm{DTR}_{i,t}$ is stock i's daily turnover ratio, DMRT_t is the market turnover ratio, $\mathrm{Dummy\_Event}(j)$, $j = 1, ..., K$ is a sequence of event dummy variables. The aggregated $\hat{\epsilon}_{i,t}$ for the entire month t is defined to be the abnormal turnover ratio (atr). Data are acquired from both WIND database and CSMAR.

(88) ***er_trend***: This trend factor is constructed following the definition in Liu et al. (2020). For the sake of simplicity, we refer readers to the original paper for more details. Data are acquired from both WIND database and CSMAR.

(89) ***largestholderrate***: Percentage of common shares owned by the largest shareholder. Data are acquired from CSMAR.

(90) ***top10holderrate***: Percentage of common shares owned by top 10 shareholders. Data are acquired from CSMAR.

(91) ***soe***: A dummy variables that equals 1 if the firm is state-owned. Data are acquired from CSMAR.

(92) ***private***: A dummy variable that equals 1 if the firm is privately-owned. Data are acquired from CSMAR.

(93) ***foreign***: A dummy variable that equals 1 if the firm is controlled by foreign investors. Data are acquired from CSMAR.

(94) ***others***: A dummy variable that equals 1 if the firm is not state-owned, private, or foreign. Data are acquired from CSMAR.

C.3 Discussions on accounting treatment in China

In China, companies are required to follow the Chinese Accounting Standards (CAS), *a.k.a.*, the Chinese Generally Accepted Accounting Principles (China GAAP), when running their businesses. Specifically, the Chinese Accounting Standards mainly consist of two sets of accounting standards: (1) the Accounting Standards for Business Enterprises (ASBEs) for general companies; (2) the Accounting Standards for Small-sized Business Enterprises (ASSBEs). We focus on the ASBEs as they are most relevant to publicly traded companies.

The revised ASBEs were introduced by the Ministry of Finance of PRC in 2006, which consist of one Basic Standard, 38 Specific Standards and the related application guidance. Interestingly, the ASBEs are more than 90% the same as the International Financial Reporting Standards (IFRS), making most account titles readily comparable across these two systems.[58] In China, all publicly traded companies are required by the government to follow the ASBEs when filing their financial statements. It is worth pointing out that even though public companies in China used a set of more traditional accounting standards before 2006, most account titles needed for signal construction in this paper are still available, and more importantly, comparable to those under the current ASBEs.[59]

On the other hand, as the U.S. GAAP are also very similar to the IFRS, most account titles mentioned in Subsection C.2 can be linked to their counterparts under the ASBEs clearly. Hence, we can follow the definitions in original papers to construct these stock-level characteristics in most cases. In rare situations, account titles may not directly have their counterparts in the ASBEs, such as the SG&A expenses. When this happens, we conduct some simple calculations to get their equivalents under the ASBEs accordingly.

Even though the difference between the ASBEs and the U.S. GAAP looks inessential in our case, it is still helpful to get a more comprehensive picture of these two systems. There are several key differences between the ASBEs and the U.S. GAAP. Firstly, the Chinese Accounting Standards only allow the historical cost valuation method when valuating fixed assets, while the U.S. GAAP

[58] See Qu and Zhang (2010) for a comprehensive comparison between the ASBEs and the IFRS.

[59] The ASBEs was first established in 1992 when China was transforming towards a market-oriented economy. The initial standards absorbed many important contexts from the IFRS but still kept some conventional accounting practice, especially when dealing with debt. After China entered WTO, the Ministry of Finance of China issued the Accounting Regulations for Enterprises in 2000, which was already very close to the current ASBEs.

allows companies to choose between the historical cost valuation method and re-evaluating the assets. In this sense, the CAS are more conservative when dealing with fixed assets. Secondly, regrading inventory, the CAS bans the usage of 'Last in, First out' (LIFO) method, which is nevertheless allowed by the U.S. GAAP. Thirdly, the CAS requires fiscal year in accounts must begin on January 1st, while the U.S. GAAP let companies to determine the starting date of their fiscal years. Fourthly, companies submit their financial statements to government and file their tax returns on a monthly basis in China, while returns can be filed on a quarterly or bi-monthly basis under U.S. GAAP. Last but not the least, the CAS stipulate that expenses are classified according to function, whereas the U.S. GAAP generally classify expenses by nature. Overall, we find that these differences haven't changed the construction of stock-level characteristics much in our empirical study.

Finally, it is worth noting that all publicly traded companies in China are required to provide a set of complete financial statements in their half-year reports, including the balance sheet, the income statement, and the cash flow statement, whose formats are the exactly same as that for annual reports (Accounting Standards for Enterprises No.32). In addition, a large proportion of public companies in China also provide such information in their quarterly reports. In contrast, the U.S. GAAP only requires public companies to provide a condensed set of financial statements, in which many account titles are not reported. Although interim statements are not necessarily audited both in China and the U.S., the extra financial information may still help investors to learn about companies better.

C.4 Classification of stock characteristics

Table C.4. Classification of stock characteristics

Category	Variable
C_size	mve, mve_ia, herf, chinv, chcsho
C_beta	beta, betasq
C_mom	mom1m, mom6m, mom12m, mom36m, chmom, er_trend, maxret
C_liq	std_dolvol, zerotrade, atr, chaotia, std_turn, ill, turn, dolvol, pricedelay
C_vol	idiovol, ear, volatility, roavol
C_own	top10holderrate, LargestHolderRate
C_bpr	bm, bm_ia, cfp, cfp_ia, sp, cashspr, invest, realestate, depr
C_ey	roeq, roaq, divo, absacc, divi, salerev, chempia, nincr, chpmia, stdacc, chtx, cash, roic, chpm, stdcf, chao, dy, acc, pctacc, saleinv, operprof, pchsale_pchrect, salecash, tb, gma, pchdepr
C_growth	egr, orgcap, sgr, pchgm_pchsale, rsup, pchsaleinv, rd_sale, rd_mve, rd, cinvest, pchsale_pchxsga, pchsale_pchinvt, agr, grCAPX, hire
C_lever	lev, pchquick, pchcapx_ia, lgr, quick, ps, tang, currat, ms, pchcurrat, cashdebt

Table C.5. Details on Macroeconomic State Variables

No.	Acronym	Variable	Definition	Frequency	Reference
1	*dp*	Dividend Price Ratio	Dividends are 12-month moving sums of dividends paid in the A-share market. The dividend price ratio is the difference between the log of dividends and the log of weighted average stock price in China's A-share market.	Monthly	Welch (2008), Gu et al. (2020b)
2	*de*	Dividend Payout Ratio	The dividend payout ratio is the difference between the log of dividends and the log of earnings of all stocks listed in China's A-share market.	Annual	Welch (2008), Gu et al. (2020b)
3	*bm*	Book-to-Market Ratio	The book-to-market ratio is the ratio of book value to market value for all stocks listed in China's A-share market.	Monthly	Welch (2008), Gu et al. (2020b)
4	*svar*	Stock Variance	The stock variance is computed as sum of squared daily returns on the SSE Composite Index.	Monthly	Welch (2008), Gu et al. (2020b)
5	*ep*	Earnings Price Ratio	Earnings are 12-month moving sums of earnings of all stocks in the China's A-share market. The earnings price ratio is the difference between the log of weighted average earnings per share and the log of weighted average stock price in China's A-share market.	Monthly	Welch (2008), Gu et al. (2020b)

Table C.6. Details on macroeconomic state variables

No.	Acronym	Variable	Definition	Frequency	Reference
6	*ntis*	Net Equity Expansion	The net equity expansion is the ratio of 12-month moving sums of net issues in China's A-share market divided by the total end-of-year market capitalization of A-share stocks.	Monthly	Welch (2008), Gu et al. (2020b)
7	*tms*	Term Spread	The term spread is the difference between the yield on 10 year government bond and the 1 year government bond.	Monthly	Welch (2008), Gu et al. (2020b)
8	*infl*	Inflation	The inflation is the monthly *Consumer Price Index* from 2000-2020 from the National Bureau of Statistics of China.	Monthly	Welch (2008), Gu et al. (2020b)
9	*mtr*	Monthly Turnover	The monthly turnover is the ratio of monthly trading volume measured in Chinese yuan to the average daily market value of all stocks in China's A-share market.	Monthly	Baker and Stein (2004)
10	*m2gr*	M2 Growth Rate	The monthly M2 growth rate (YoY) from 2000 to 2020 is from the National Bureau of Statistics of China.	Monthly	Chen (2009)
11	*itgr*	International Trade Volume Growth Rate	The international trade volume growth rate (YoY) from 2000 to 2020 is from the National Bureau of Statistics of China.	Monthly	Rapach et al. (2013)

D Time variations of R^2_{oos} and variable importance

To obtain a better understanding of model predictability, we also explore the time variations in the out-of-sample R^2_{oos} of our models. In Figure D.1, we plot the yearly out-of-sample R^2_{oos}, which we define as

$$R^2_{\text{oos},S}(t) = 1 - \frac{\sum_{i \in \mathcal{I}_t} (r_{i,t} - \hat{r}^{(S)}_{i,t})^2}{\sum_{i \in \mathcal{I}_t} r^2_{i,t}}, \tag{21}$$

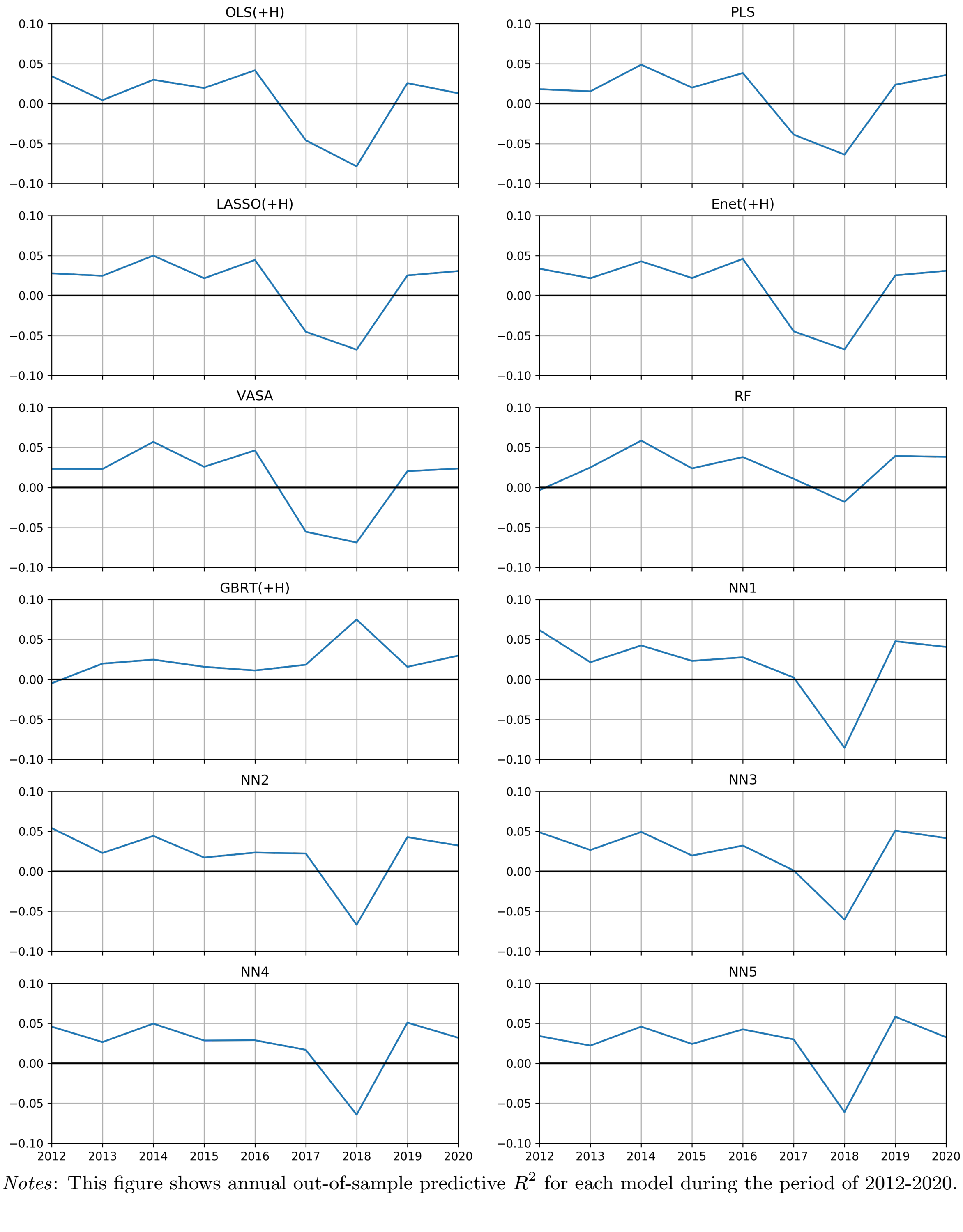
where $\mathcal{I}_t$ is the set of tradable stocks in year t in the testing sample.

We make the following two observations. First, all models, except GBRT, have experienced a significant drop in R^2_{oos} in 2018. In that year, all models excluding GBRT produce negative R^2_{oos}, indicating that naive predictions of zero returns would have beaten them. However, the dysfunctionality of machine learning models in 2018 is likely due to the Chinese stock market's persistent fall caused by the severe trade conflicts between China and the United States. This finding points 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 case, the political risk related to a trade war between the U.S. and China. From this perspective, it is even more surprising that GBRT still achieves a positive out-of-sample R^2_{oos} in 2018, which is even larger than those in other periods. We conjecture that its generic model properties cause GBRT's unusual performance. For example, if the model heavily relies on predictors relating to price trends, it may still generate a good performance in 2018. Still, compared to other machine learning methods such as neural networks, GBRT generally produces a lower out-of-sample R^2 in periods other than 2018.

Second, regularized linear models and VASA based on linear submodels not only produce negative R^2_{oos}'s in 2018 but also in 2017. In that year, the Chinese stock market went through a persistent boom for large stocks as the CSI 300 Index[60] has increased by almost 30% within the year, and most large stocks have positive returns in almost every month. As noted previously, monthly returns of large stocks in the Chinese market are more challenging to predict, especially for linear models, which is likely why PLS, LASSO, Enet, and VASA attain negative R^2_{oos}'s in 2017. On the hand, neural networks, especially NN2, NN4, and NN5, produce positive R^2_{oos}'s in 2017, indicating their prediction

[60] The CSI 300 is a capitalization-weighted stock market index designed to replicate the performance of the top 300 stocks traded on the Shanghai Stock Exchange and the Shenzhen Stock Exchange.

Figure D.1. Annual out-of-sample predictive R^2

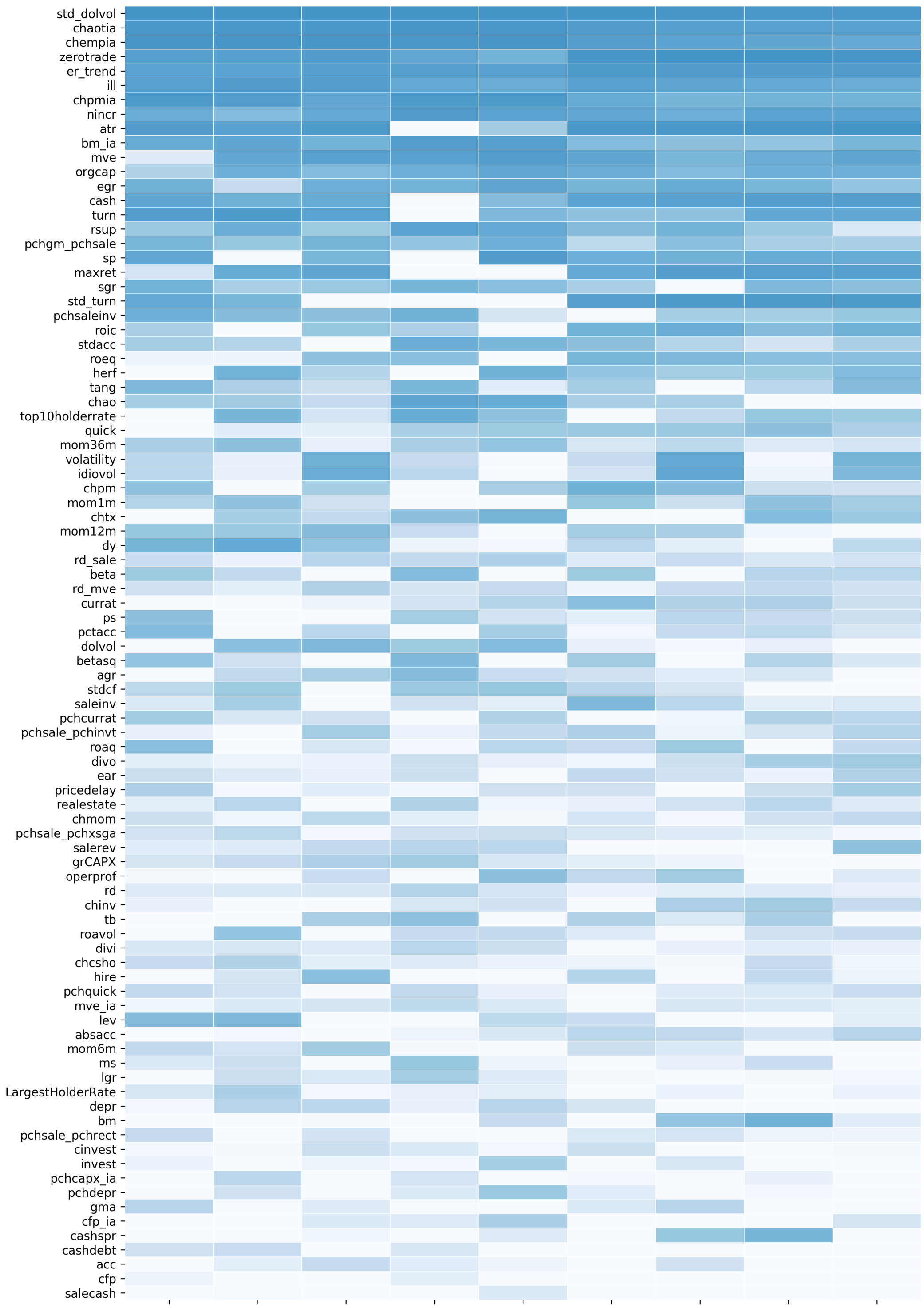


Notes: This figure shows annual out-of-sample predictive R^2 for each model during the period of 2012-2020.

performance is quite robust to specific market conditions associated with certain subgroups of stocks.

In Figure D.2, we take a closer look at the time variability of the variable importance for NN4.

Figure D.2. Characteristic importance for NN4



Notes: This figure shows the ordering of all stock-level characteristics ranked by NN4 across the different years. The vertical axes give the orderings of the NN4-specific R^2-based variable importance, which is defined similarly as in Figure 2. The horizontal axes indicate the periods in the testing sample, and the color gradient in each column reflects the explanatory power of predictors in a given evaluation period.

E Equal-weighted portfolio analysis

Table E.7. Performance of machine learning portfolios (equal-weighted)

	"1/N" Portfolio	*Machine Learning Portfolios* OLS-3 +H	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
Long-Short													
Avg	–	1.50	4.06	4.80	4.88	4.06	3.27	5.14	5.46	5.45	5.80	5.95	5.93
Std	–	4.49	4.69	5.17	5.22	5.02	4.12	5.29	4.72	4.51	5.01	5.02	4.95
S.R.	–	1.15	3.00	3.21	3.24	2.80	2.75	3.36	4.01	4.18	4.00	4.10	4.15
Skew	–	0.16	−0.44	1.29	0.76	1.11	−0.20	1.01	2.14	1.08	1.95	2.66	1.88
Kurt	–	0.37	1.61	6.80	6.27	0.93	0.65	6.22	9.65	3.29	8.81	13.87	7.93
Max DD	–	34.04	14.12	9.07	16.58	13.67	9.86	10.60	4.12	3.86	4.64	4.51	4.49
Max 1M Loss	–	11.40	14.12	9.07	15.26	8.83	9.69	10.60	3.25	3.86	4.64	3.45	4.49
Long-Only													
Avg	1.56	2.24	3.67	4.05	4.20	3.83	3.48	4.38	4.50	4.45	4.74	4.91	4.85
Std	8.44	9.16	7.75	9.08	9.22	8.40	8.50	9.14	9.55	8.73	9.29	9.57	9.71
S.R.	0.64	0.85	1.64	1.54	1.58	1.58	1.42	1.66	1.63	1.77	1.79	1.78	1.73
Skew	0.26	0.57	0.26	1.09	1.05	0.56	0.54	0.99	1.20	0.89	1.20	1.36	1.31
Kurt	1.26	1.38	1.50	4.86	4.50	3.09	1.05	4.38	5.24	3.97	5.41	5.49	5.43
Max DD	54.20	45.35	23.55	23.45	23.03	27.84	27.11	23.12	22.62	21.79	23.23	20.80	21.73
Max 1M Loss	25.56	22.89	20.82	21.85	22.42	23.39	20.25	21.99	22.62	21.26	22.33	19.45	20.11

Notes: This table reports the out-of-sample performance measures for all machine learning models of the equally-weighted long-short and long-only 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.": Sharpe ratio. "Skew": skewness. "Kurt": kurtosis. "Max DD": the portfolio maximum drawdowns (%). "Max 1M Loss": the most extreme negative monthly return (%).

E.1 Ex-ante selection of machine learning methods

A practical problem of real-world investment is to select the best prediction method *ex-ante*, which is especially relevant when the number of candidate models is large. We consider two model selection procedures and report their performance on the testing sample. The first one is via simple model averaging. We first predict monthly returns for all stocks using 11 machine learning methods (PLS-NN5) and then construct the long-short and long-only portfolios based on the average predicted stock returns. We report the out-of-sample performance for the model-averaging portfolios in Table E.8, which shows that this procedure works well in the Chinese market. For example, the model-averaging long-only portfolio achieves a Sharpe ratio of 1.76, which is in line with the highest Sharp ratio for a single method shown in Table 6.

The second procedure utilizes the CSPA test to select the method that performs the best con-

Table E.8. Performance of the model-averaging portfolio (value-weighted)

Portfolio	Avg	Std	S.R.	Skew	Kurt	Max DD	Max 1M
Long-short	5.10	5.76	3.06	1.02	4.24	8.51	8.51
Long-only	4.24	8.36	1.76	1.31	6.08	20.63	19.81

Notes: This table reports the out-of-sample performance measures for the model-averaging portfolio 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 (%).

ditional on the current macroeconomic condition. Given a pre-specified conditioning variable and the benchmark model, we estimate the conditional expected loss differential functions and the corresponding confidence bounds using the out-of-sample predictions in the testing sample on an annual basis. These plots then form a decision rule for selecting the forecast method month by month. More precisely, given the current macroeconomic condition X_t, the benchmark model is selected if the estimated loss differential to all other models is positive. If a loss differential is significantly negative for a set of alternative models, we chose the model with the most negative loss differential.

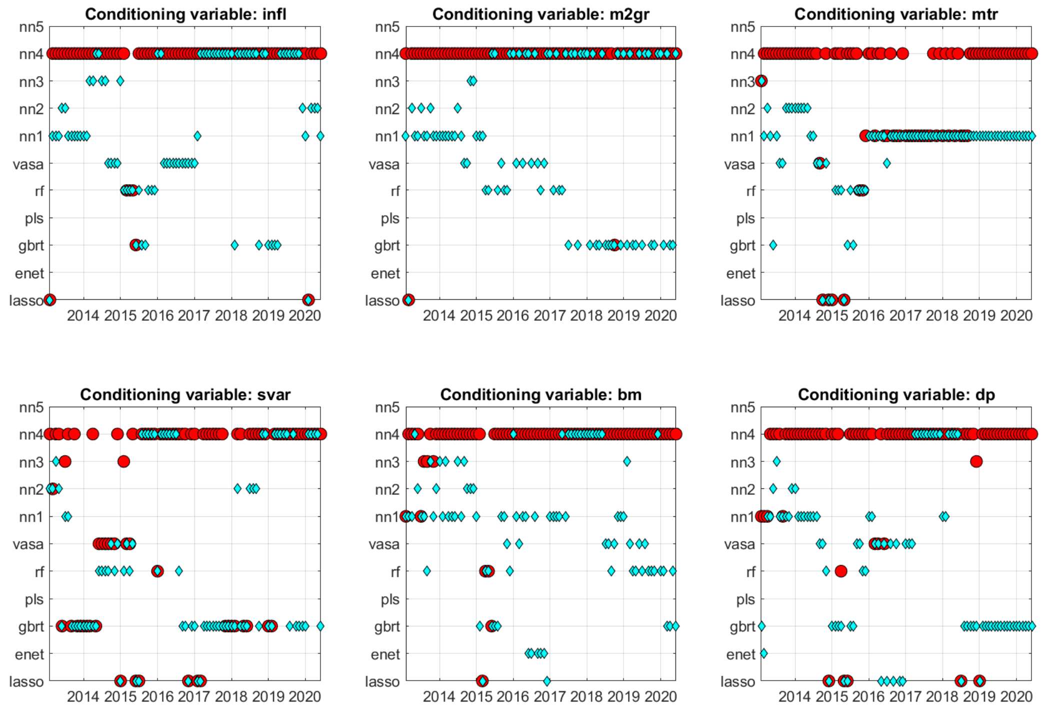
In Figure E.3, we plot the selected models for different conditioning variables across time when we choose NN4 as our benchmark model for the full sample. We either use directly the change in the loss function or its 90% confidence bound as a decision rule for model selection. At the end of each month, we check whether the expected loss differential or its 90% confidence bound is below zero at the conditioning variable's actual value. If this is the case, we switch to the model with the largest negative loss differential and use that model for investing in the following month. The resulting model choices are plotted in Figure E.3 as blue diamonds (when using the expected loss differential) and red circles (when using the 90% confidence bound). Clearly, using the confidence bound for the decision rule provides some additional robustness. Under this rule, there are only a few deviations from the benchmark model. Interestingly, these deviations happen for most conditioning variables during 2015, when the Chinese market in a crisis. During that time, the CSPA-based model selection particularly favors Lasso and, to some extent, GBRT. When using the expected loss differential, the model selection varies much more, but with some preferred candidates like NN1, GBRT, and VASA.

Table E.9. Performance of CSPA-based model selection portfolios (value-weighted)

	infl	m2gr	mtr	svar	bm	dp	infl	m2gr	mtr	svar	bm	dp
	Long-short						Long-short, 10%					
Avg	4.54	4.88	4.80	5.17	5.08	4.98	5.67	6.10	5.96	5.39	5.64	5.74
Std	4.57	4.34	4.11	5.06	3.83	4.43	3.83	5.36	5.21	5.56	3.81	4.38
S.R.	3.44	3.89	4.05	3.54	4.59	3.89	5.13	3.94	3.96	3.35	5.12	4.53
Skew	-0.20	0.42	0.11	0.02	0.56	-0.05	0.49	2.36	2.66	0.88	0.53	0.12
Kurt	3.57	3.19	4.48	3.29	3.08	3.93	3.37	14.07	16.00	7.86	3.45	4.20
Max DD	18.62	4.46	8.62	7.94	3.79	8.67	4.43	4.43	4.43	8.62	4.43	8.67
Max 1M	9.01	3.47	9.01	7.98	3.86	9.07	3.47	3.47	3.47	9.01	3.47	9.07
	Long-only						Long-only, 10%					
Avg	4.13	4.26	4.28	4.46	4.27	4.40	4.67	5.07	4.87	4.64	4.63	4.74
Std	8.59	8.62	8.67	8.46	8.99	8.94	8.98	9.81	9.77	9.30	8.94	8.73
S.R.	1.67	1.71	1.71	1.82	1.65	1.70	1.80	1.79	1.73	1.73	1.79	1.88
Skew	0.35	0.61	0.41	0.52	0.42	0.68	0.43	1.40	1.36	1.18	0.41	0.44
Kurt	4.19	4.25	4.30	4.39	3.97	5.11	4.00	8.29	8.72	7.27	3.96	4.12
Max DD	25.49	19.55	24.05	20.05	25.74	20.30	23.91	18.81	20.92	20.05	23.91	18.73
Max 1M	20.25	19.46	20.25	19.46	20.25	20.25	19.46	19.46	22.61	19.46	19.46	19.46

Notes: This table reports the out-of-sample performance measures for the CSPA-based portfolio based on the full sample. All measures are based on 91 monthly out-of-sample returns from January 2013 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 (%).

Figure E.3. Dynamic model selection based on CSPA test



Notes: This figure shows the monthly model selection based on the CSPA test and the prevailing value of the conditioning variable. The diamond markers represent the model when using directly the change in the loss function, and the circles represent the models when requiring a 10% confidence level. Portfolio formation starts in January 2013 and ends in June 2020.

To explore whether the above ex-ante model selection rule also provides superior performance, we analyze the long-short and the long-only strategy for the full sample. Table E.9 reports the results. For the long-short strategy, we find that using the CSPA-based selection criterion is highly beneficial compared to the model-averaging approach in Table E.8. The portfolio performs best when we take the confidence bound as a decision criterion across the six conditioning variables examined; the average Sharpe ratio is 4.34, compared to a Sharpe ratio of 3.06 in Table X. Even if we take the expected loss differential, we still get an average Sharpe ratio of 3.9. The portfolios perform better than average when we condition on $infl$, bm, and dp.

A similar conclusion holds when we investigate the performance of the long-only strategies. However, it turns out that the superiority of the CSPA-based selection versus the model-averaging approach vanishes. When we select models based on the 90% confidence bound, the average Sharpe ratio is 1.79, only slightly above the Sharpe ratio of 1.76 for the model-averaging approach. Using the expected loss differential even leads to a slightly underperforming average Sharpe ratio of 1.71. Hence, for the long-only portfolio, the two ex-ante selection strategies lead to a similar performance. Nevertheless, the long-only portfolios, $infl$, bm, and dp perform well if we use the confidence bound as the decision criterion. Therefore, it would be interesting to explore the role of the conditioning variable further. We leave this issue as a potential avenue for future research.

Reference

Asness, C. S., Porter, R. B. and Stevens, R. L.: 2000, Predicting stock returns using industry-relative firm characteristics, *Available at SSRN 213872* .

Eisfeldt, A. L. and Papanikolaou, D.: 2013, Organization capital and the cross-section of expected returns, *The Journal of Finance* **68**(4), 1365–1406.

Green, J., Hand, J. and Zhang, F.: 2017, The characteristics that provide independent information about average US monthly stock returns, *The Review of Financial Studies* **30**, 4389–4436.

Liu, Y., Zhou, G. and Zhu, Y.: 2020, Trend factor in China, *Available at SSRN 3402038* .

Mohanram, P. S.: 2005, Separating winners from losers among lowbook-to-market stocks using financial statement analysis, *Review of accounting studies* **10**(2-3), 133–170.

Ou, J. A. and Penman, S. H.: 1989, Financial statement analysis and the prediction of stock returns, *Journal of accounting and economics* **11**(4), 295–329.

Pan, L., Tang, Y. and Xu, J.: 2016, Speculative trading and stock returns, *Review of Finance* **20**(5), 1835–1865.

Piotroski, J. D.: 2000, Value investing: The use of historical financial statement information to separate winners from losers, *Journal of Accounting Research* pp. 1–41.

Qu, X. and Zhang, G.: 2010, Measuring the convergence of national accounting standards with international financial reporting standards: The application of fuzzy clustering analysis, *The International Journal of Accounting* **45**(3), 334–355.

Sloan, R. G.: 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings?, *Accounting review* pp. 289–315.