Machine learning in the Chinese stock market

期刊：Journal of Financial Economics

汇报人：罗跃 李代猛

时间：2021-11-25

文章信息：于2021年4月投稿，2021年6月被接收，2021年8月正式刊出

引言

- **是否某些由投资者交易行为形成的新兴技术性指标比公司基础特征更具有定价能力（预测能力）？**

中国股市由所谓“散户”主导。根据2019年上海交易所的年报，散户数占所有账户的比例为99.8%。如此高的散户比例，导致整个市场的交易额会因为散户频繁的短期交易行为而增加。因此整个市场的高波动性有可能会造成股价在经济基础上的偏移。

- **是否上市国企的股价会因为政府调控因素而变得难以预测?**

中国股市的关键特征是集中调控、银行主导与单一关系驱动。比如上市国企的股价在低于基础价值时是被保护的。鉴于国企在中国资本市场中的重要性和唯一性，所以对于国企的研究需要更加细致和特殊的对待。

- **从无做空操作出发，使结果更易被中国股市投资者理解与应用。**

中国资本市场缺少做空机制，而金融领域经典的因子分析主要依赖于多空组合构造。

数据

数据来源

- **Wind数据库**
  沪深两市A股每日和每月股票回报
- **CSMAR**
  中国一年期政府债券的收益率、季度财务报表数据
- **国家统计局网站**
  宏观经济数据

特征

- **股票级别特征**
  94个
- **行业哑变量**
  80个
  《上市公司行业分类指南》
- **宏观经济变量**
  11个

样本

- **股票数量**
  2000年1月至2020年6 月期间交易的3900多只A股股票
- **训练集**
  2000 - 2008
- **验证集**
  2009 - 2011
- **测试集**
  2012 -2020

数据频率

- **每月更新**
  22个股票级别特征
- **季度更新**
  51个股票级别特征
- **半年更新**
  6个股票级别特征
- **年度更新**
  15个股票级别特征

研究方法

预期超额收益率和超额收益的关系式

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

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

模型输入向量

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

其中Zi,t是P维的预测因子向量，Ci,t是一个94维的股票特征向量，Xt是11维度的宏观特征向量，di,t是一个80维的行业哑变量向量，$\otimes$ 指两向量之间的克罗内克积。

g(·)的具体形式不做设定，寻找一个g(·)的具体形式，以达到最佳预测效果。
本文考虑g函数的2个简单线性模型和11个机器学习模型，分别是普通最小二乘、三因子（规模、账面市值比和动量）最小二乘、偏最小二乘、LASSO、Enet、GBRT、RF、VASA、NN（单层到五层）。

研究方法

- **PLS**

  利用转换矩阵将输入矩阵降维到线性空间；降噪，保留Y和X的相关性；

- **VASA**

  平均多个下采样模型预测结果；降维；偏差小；

- **Lasso**

  解决线性回归的过拟合；引入L1范数；某些系数会变成0；减少计算量；

- **Enet**

  正则化的线性模型；L1范数和L2范数；删除无效特征，稳定；擅长相关特征选择；

- **RF**

  集成学习；bagging；弱分类器：决策树；方差较小；

- **GBRT**

  集成学习；前向分步加法模型；弱分类器：回归树；偏差较小；

研究方法

$R^2_{oos,S}$ 用于衡量模型的拟合优度

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

通过比较R-square来判断模型的预测能力

Huber损失函数作为优化目标函数

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, & |y - f(x)| \le \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2, & |y - f(x)| > \delta \end{cases}$$

既可导，又降低函数对异常值的敏感度

结合了MAE 和 MSE 的优点

研究方法

根据3个标准将样本分成6个子集，分别按月和按年进行预测

- 企业规模

  排序前70%

  排序后30%

- 股东平均市值

  排序前70%

  排序后30%

- 国企与非国企

  国企

  非国企

Table 1

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

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

实证分析——模型比较

- **在考虑所有公司时，最小二乘模型解释力达到0.81%，大于三因子最小二乘模型：**

  这表示传统的线性三因子模型（SMB/BM/Momentum）并不能囊括大部分的线性预测因子。

- **PLS、LASSO和Enet，解释力均上升到了1%以上：**

  这些模型能够从大量备选预测因子中提取最具有代表性的子集，也就更具有稳健性与泛化能力。

  结果表示，全体公司特征在进行月度收益预测时是冗余的，并没有必要考虑所有可能的公司特征。

- **树模型RF、GBRT和五个神经网络模型，解释力都上升到2%以上：**

  这表明机器学习算法在捕捉预测因子间复杂交互作用时的强大能力。

实证分析——月度预测

按月预测结果

- 所有模型对小盘股的预测能力都要优于大盘股。
- 大部分模型对股东平均市值较低的股票的预测能力都要优于股东平均市值较高的股票。
- 对非国企股票的预测能力要优于国企股票。

分析

- 市场结构以散户为主，散户偏好小盘股，短期看，市场存在散户投机行为。
- 中国股市普遍存在国有企业，而国有企业的透明度低于其他企业。

Table 1

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

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

实证分析——年度预测

按年预测结果

- 所有模型对大盘股的预测能力都要优于小盘股。
- 大部分模型对股东平均市值较低的股票的预测能力都要劣于股东平均市值较高的股票。
- 对非国企股票的预测能力要劣于国企股票。

分析

- 长期来看，国企受益于政府激励政策，大盘股具有某些优势。

Table 2

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

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

实证分析——中美市场对比

中美结果对比

- 无论是按月还是按年预测，对我国股市的预测结果远远好于美国股市。

Table 3

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

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

实证分析——变量重要性分析

指标的重要性度量：对于一个特定的模型，当在每个训练样本中将给定指标的所有值设置为零时，R^2的减少量。

- infl: Inflation
- ntis: Net Equity Expansion (净股本扩张:中国A股市场12个月净发行额的移动总和除以A股股票年末总市值的比率)
- bm: Book-to-Market Ratio(账面市值比率:中国A股市场上所有上市股票的账面价值与市值之比)
- ep: Earnings Price Ratio(市盈率:中国A股市场加权平均每股收益对数与加权平均股价对数之差)

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

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

实证分析——变量重要性分析

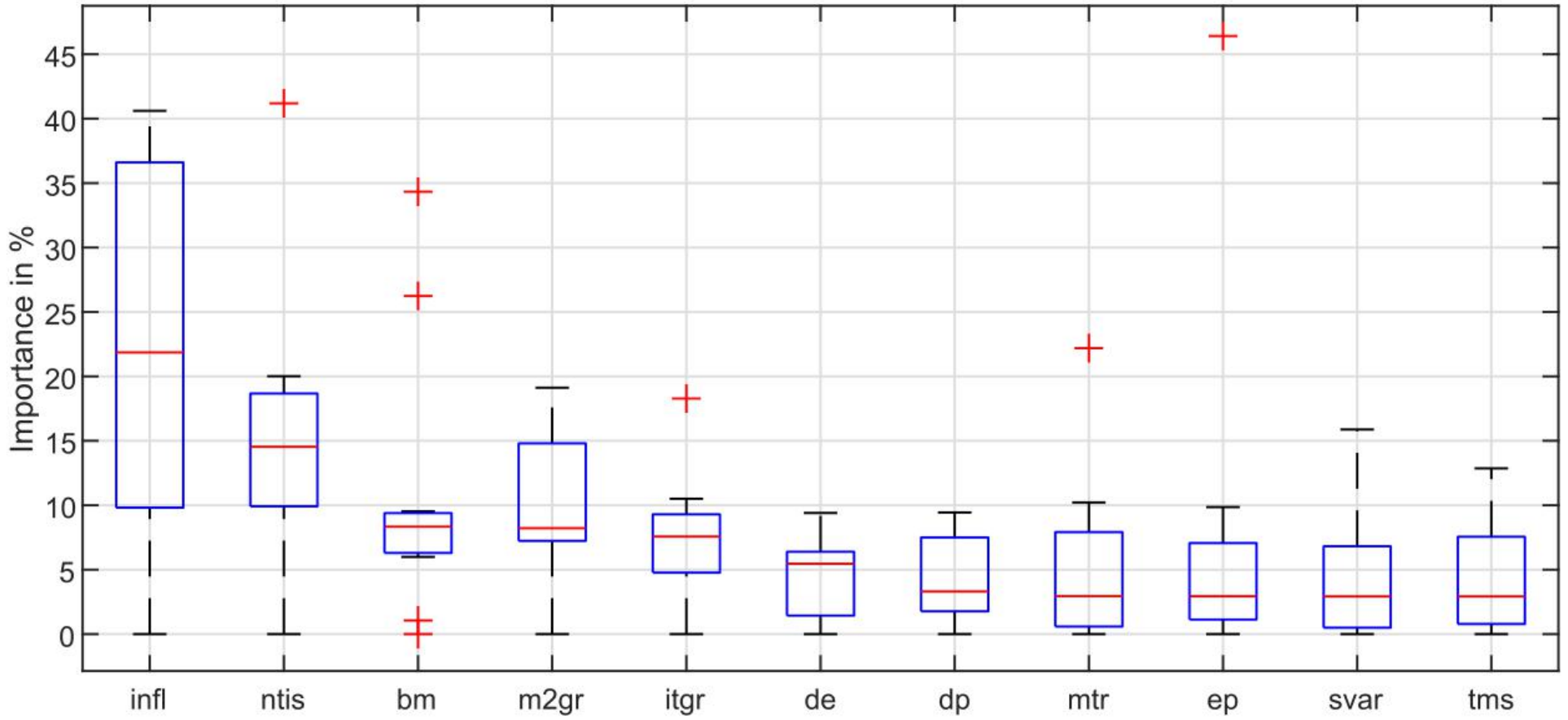
- 对于PLS而言，衡量发行活动水平的nti（净股本扩张）最为重要。中国过去长期采用基于审批的IPO制度，证监委经常在股市下跌时暂停或减少IPO数量，这使得nti在预测月度回报方面发挥重要作用是合理的。
- nti也是GBRT模型中最重要的宏观经济变量，也是神经网络模型的第二重要变量。
- 与其他基于回归的方法相比，树模型GBRT和RF中宏观经济变量的重要性分布相对更为均匀，表明这两种方法可以检测宏观经济变量与股票特征之间潜在的复杂非线性交互作用。

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

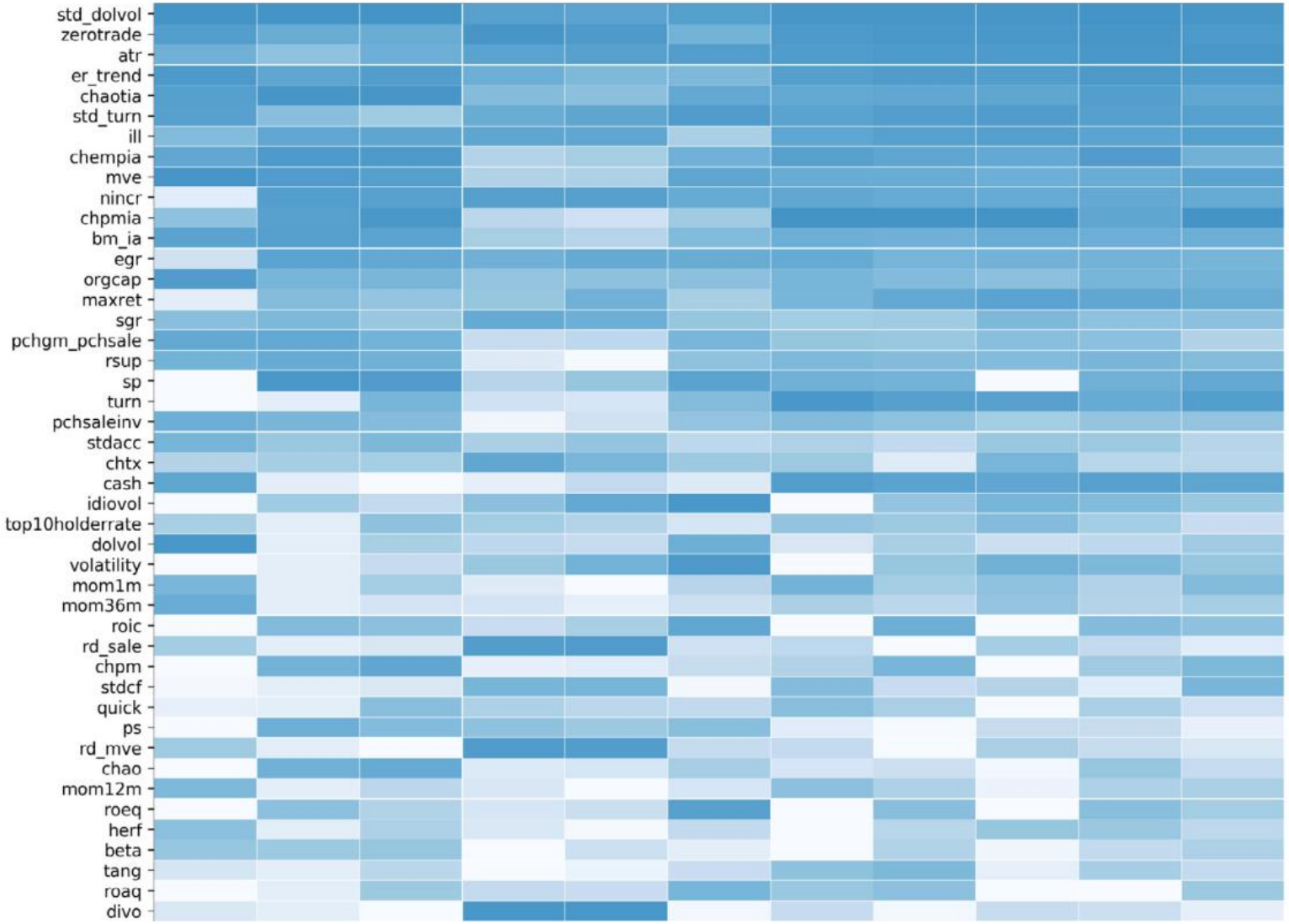
	PLS	LASSO +H	Enet +H	GBRT +H	RF	VASA	NN1	NN2	NN3	NN4	NN5
dp	0.00	8.65	4.07	9.11	9.44	1.34	2.17	2.96	3.31	4.01	1.63
de	0.00	1.06	1.78	9.40	8.59	1.32	5.46	5.86	5.28	6.57	5.78
bm	1.06	34.33	26.24	8.97	8.34	0.00	8.46	7.23	5.99	7.99	9.53
svar	0.00	0.00	0.13	7.76	8.86	15.88	2.12	2.93	3.23	3.97	1.59
ep	0.00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

实证分析——变量重要性分析

- infl和nti是预测中国股市收益率最具影响力的两个宏观经济变量，尤其是使用神经网络时。
- 股息价格比（dp）、市场波动率（svar）、每股总收益（ep）、期限利差（tms）和市场流动性（mtr）都被大多数模型所忽略，因而算不上重要变量。



实证分析——变量重要性分析



左方方热力图展示了公司层面特征在不同模型中的重要程度。

- 与市场流动性相关的股票特征最为重要，即流动性波动性（std_dolvol和std_turn）、零交易日（zerotrade）和不确定性度量（ill）。
- 第二具有影响力的变量集包括基本信号和估值比率，如行业调整后的资产周转率变化（chaotia）、行业调整后的员工变化（chempia）、总市值（mve）、最近收入增长数（nincr）、行业调整后的利润率变化（chpmia）和行业调整账面市值（bm_ia）。
- 对于美国资本市场的研究指出，传统的价格趋势指标是最具影响力的预测指标，而对中国股市来说，除了最高每日回报（maxret）之外，其他指标的重要性都较低。
- 基本面因素的突出作用同样是这一部分的显著结论，Gu et al.（2020）表明这些因素对美国市场的重要性很小，但在预测中国市场时就很重要。

投资组合分析——投资策略

多空投资组合：在每个月末，生成提前一个月的样本外股票回报。然后，根据预测的回报率将股票分成十分位，并使用价值权重每月重新组合投资组合。因此，通过买入预期回报最高的股票（十分之一）并卖出预期回报最低的股票（十分之一）来构建多空投资组合。

只做多投资组合：尽管长短组合是评估机器学习方法在投资组合层面绩效的有用工具，但由于严格的卖空限制，很难在中国股票市场上实现。因此，本文还包括只做多的投资组合，该投资组合只持有前十分位的股票。

注：中国证监会（CSRC）于2010年3月推出了融资融券交易。最初只有90只股票可供卖空，但截至2020年7月，已增至800只。然而，这一数字相对于中国市场的股票总数（超过4000只）仍然很小。

投资组合分析——结果分析

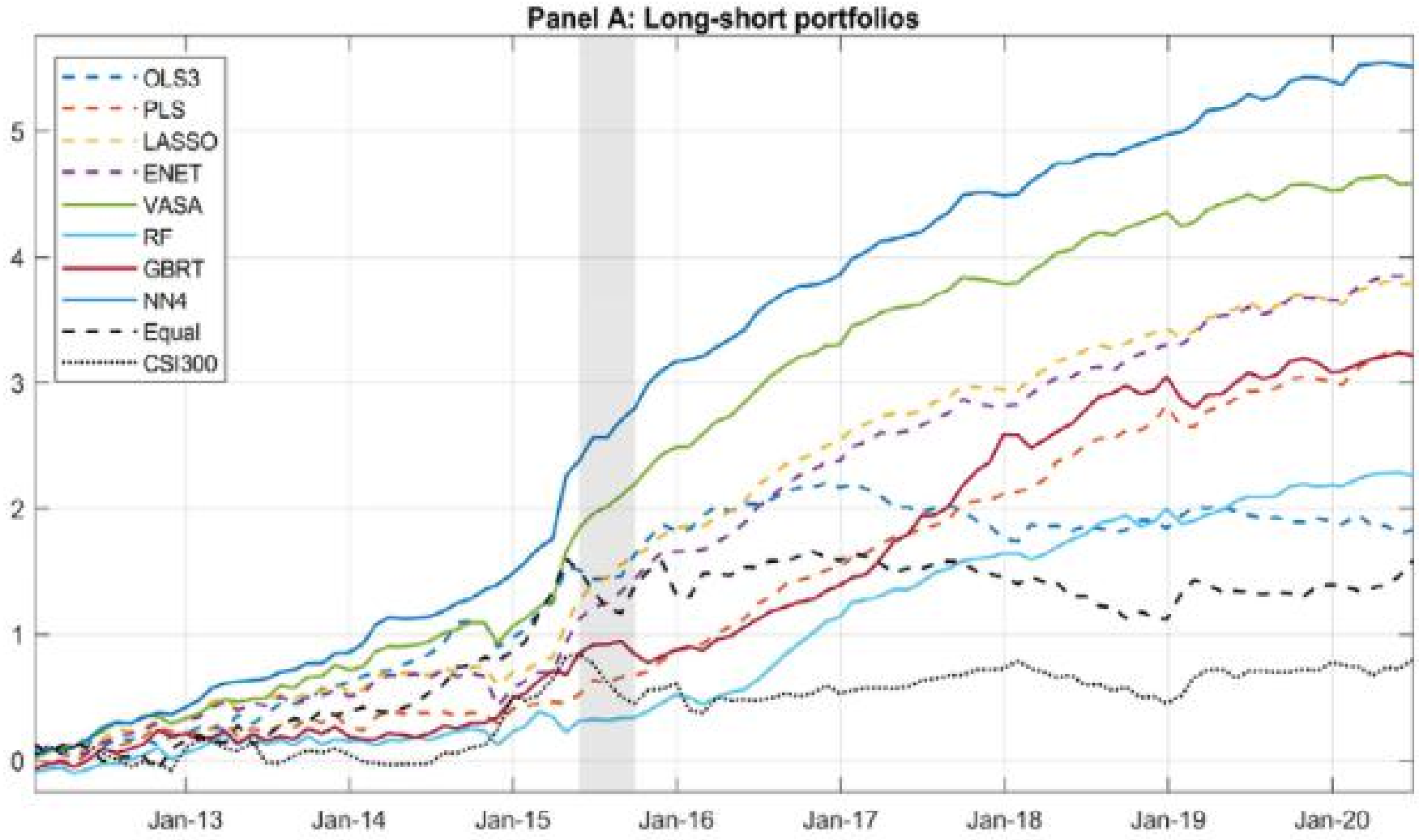
出于比较目的，本文还报告了所有股票权重相等的1/N投资组合的绩效。就平均预期月回报率、夏普比率和其他衡量指标而言，所有机器学习投资组合在OLS-3投资组合和1/N投资组合中占据主导地位。总体而言，机器学习技术，特别是神经网络模型，有利于投资组合水平的预测。

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

Avg: 平均预测月回报率
Std: 预测月收益率的标准差
S.R.: 年化夏普比率
Skew: 偏度系数
Kurt: 峰度系数
Max DD: 投资组合最大亏损（%）
Max 1M Loss: 最极端的负月回报率（%）

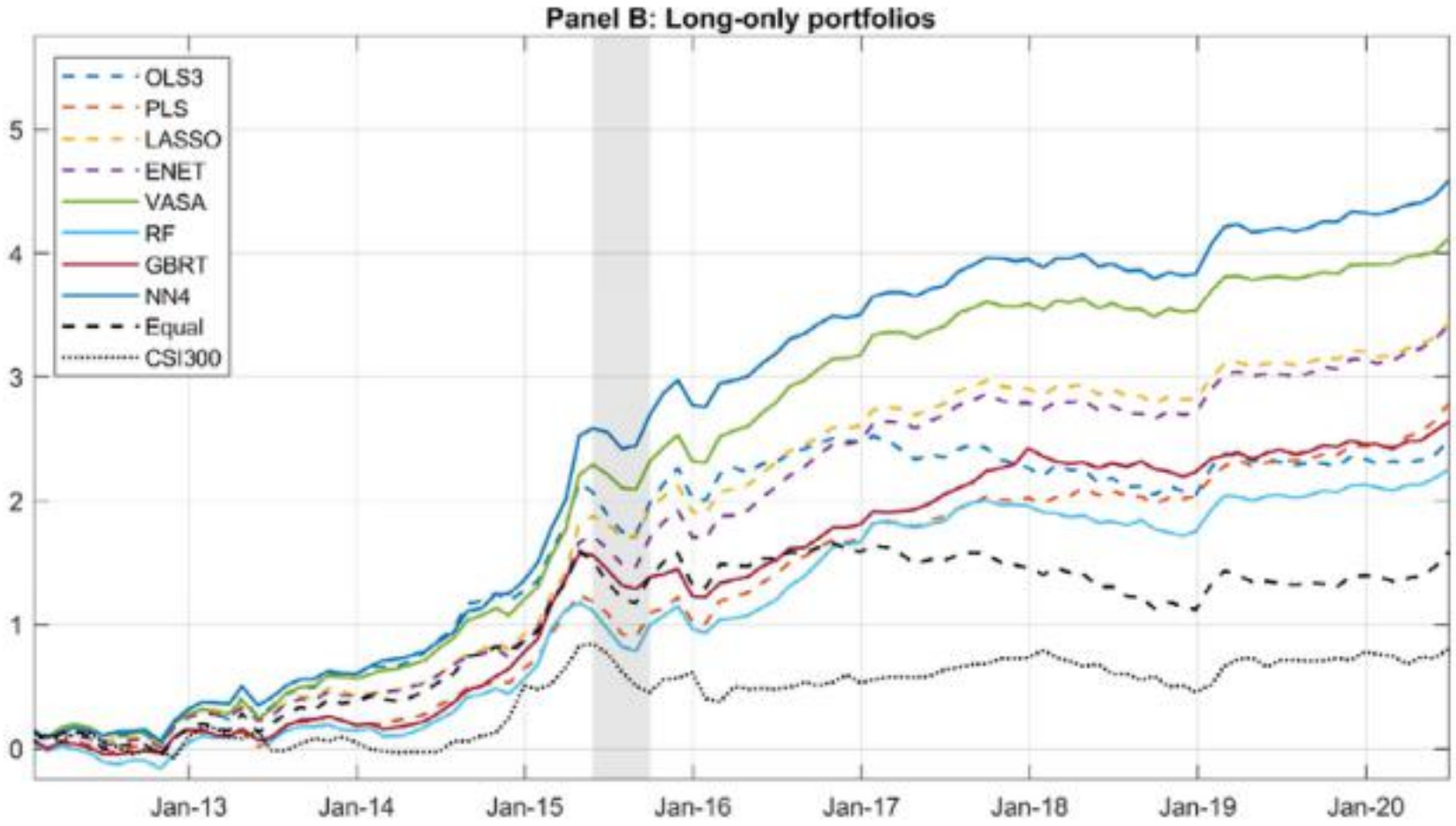
投资组合分析——结果分析

- 随着市场指数CSI300作为基准。在所有三种投资组合类型中，神经网络模型在其竞争对手中占据主导地位。
- VASA尽管简单，但被证明是仅次于NN4的第二好方法。请注意，正如阴影区域所示，这两种方法的长短组合在2015年股市崩盘期间表现非常好。此外，此外，由于2020 2019冠状病毒疾病爆发的全球冲击，并没有导致投资组合水平显著下降。
- 神经网络和VASA之后是惩罚线性模型，包括LASSO和Enet，它们的性能非常相似，因为这两种方法有许多共同点，而树模型的性能落后。



投资组合分析——结果分析

- 所有机器学习投资组合的表现都优于1/N投资组合和沪深300指数。
- 对于长短组合，本文在中国股市获得的夏普比率远远高于Gu等人（2020年）在美国股市中发现的夏普比率。例如，NN3在中国市场给出的最高夏普比率（SR=3.45）是NN4给出的最佳夏普比率（SR=1.35）的两倍多。如上所述，由于交易限制，多空策略几乎不可行，因此本文在解释这些结果时持谨慎态度。同时，仅长期投资组合的最高夏普比率为1.76，仍高于美国市场的多空策略。



投资组合分析——结果分析

基于前70%大型股的机器学习投资组合的表现在质量上与完整样本相似。然而，因为小型股被排除在外，所有投资组合的平均月回报率、夏普比率、标准差和极端负的月回报率都较低。然而，机器学习方法仍然在简单的OLS-3模型和1/N组合中占主导地位，神经网络表现最好，其次是正则化线性模型和树模型。因此，这些结果证实了机器学习方法在中国股票市场中也具有出色的投资组合水平预测能力。

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

投资组合分析——结果分析

- 就夏普比率而言，国有企业的长短期策略表现远远高于前70%的股票，尤其是神经网络。
- 对于只做多的投资组合，1/N投资组合确实表明国有企业投资组合最大亏损大于前70%的股票（也包括国有企业）。然而，利用国有企业收益的可预测性，我们可以将长期策略的投资组合最大亏损降低到大大低于最大70%股票的水平。与此同时，长期国有企业投资组合的夏普比率也较高。

		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

投资组合分析——其他因素

本文还考虑了交易成本和每日限价（中国市场对主板和二板上市的普通股实行每日10%的限价（自2020年8月起，二板上市的股票限价为20%），对特殊处理（ST）股票限价为5%，对科技创新板上市的股票限价为20%。）情境下的投资组合性能评估，总的来说，考虑交易成本和引入现价规则之后不同策略的绩效仍然具有经济意义。

	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

结论

文章研究了几种机器学习方法在中国股市中的预测能力。
结论1：最关键的预测因素是基于流动性的交易信号。此外，中国股市正朝着允许和鼓励基本面投资这个方向发展。
结论2：**散户投资者的短期主义在短期投资期限内产生了实质性的可预测性，特别是对于小盘股。**同时，由于政府信号在中国市场中扮演着重要角色，研究也观察到国有企业的长期可预测性显著提高。
结论3：对于投资组合的分析表明，**短期内的高可预测性可以转化为多空投资组合的高夏普比率。**然而，因为中国市场的做空可操作性较低，作者还分析了仅做多的投资组合，发现其表现仍然具有经济意义。
总体而言，研究表明**机器学习方法可以成功应用于中国股票市场，尽管中国股票市场与美国股票市场具有完全不同的特征。**

感谢您的倾听！

Q&A