TSINGHUA PBCSF 清华五道口 | IFR 金融科技研究院

研究报告

2020 年 第 12 期 （总第 65 期） 2020 年 5 月 10 日

机器学习视角下中国股票资产收益率可预测性研究

吴辉航 魏行空 张晓燕

（鑫苑房地产金融科技研究中心）

【摘要】股票收益率的可预测性一直以来都是金融学的核心研究问题之一，本文尝试引入机器学习的方法来探索收益率可预测问题在中国的答案。基于 1997 年 1 月到 2019 年 12 月 A 股市场的 108 个股票异象性特征，本文比较了传统计量经济学模型与最小偏二乘回归、主成分回归、弹性网络回归、随机森林、梯度提升树和神经网络模型 6 大主流机器学习算法在 A 股个股样本外可预测性问题上的表现。研究主要发现有三点：（1）历史交易数据信息对下个月个股股票收益率依然有预测效果，且机器学习算法的样本外预测效果优于传统计量经济学模型。(2)在中国 A 股市场上，流动性类特征变量的预测能力较强，而动量类特征较弱。（3）机器学习算法与资产定价研究结合有显著的经济意义，两层神经网络等权重（市值加权）多空策略资产组合的绩

效表现在所有模型中表现最好，在样本外测试期内平均能获得 3.03%（2.94%）的月度收益，月度波动率为 4.65%（6.88%），年化夏普比率为 2.26（1.48），经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 3.03（2.95）。

TSINGHUA PBCSF 清华五道口 | IFR 金融科技研究院

Research Report

2020-2 edition 65 May 10th 2020

Are Stock Returns Predictable in China? A Machine Learning Approach

Wu huihang, Xingkong Wei, Xiaoyan Zhang

XIN Real Estate Fintech Research Center

Abstract： The predictability of stock returns has always been one of the core research questions in finance. This paper attempts to introduce machine learning method to answer whether stock returns are predictable in China. With 108 trading characteristics data in China A share market from January 1997 to December 2019, this paper compares the out of sample predictability of the traditional econometric model with that of 6 major machine learning models, including partial least squares, principal component regression, elastic net regression, random forests, gradient boosted regression trees and neural networks. The main findings of this study are as follows: (1) historical trading data can predict individual stock returns in the next month, and the out-of-sample prediction of machine learning algorithm is better than that of traditional econometrics model; (2) in China A share market, liquidity characteristics have strong predictive power, while momentum characteristics are weak in out of sample prediction; (3) the combination of machine learning algorithm and asset pricing research can generate significant

economic value. During the out of sample test period, the performance of two-layer neural network equal-weight (market value weighted) long-short strategy is the best among all models, with the average monthly return of 3.03% (2.94%), the monthly volatility of 4.65% (6.88%), the annualized Sharpe ratio of 2.26 (1.48), and the significant monthly adjusted Alpha of 3.03 (2.95) in terms of FF5 factor. We present results that demonstrate that machine learning algorithm does indeed have clear merit over traditional techniques in China.

Keywords: Return Prediction, Out-of-sample forecasts, Machine Learning, Fintech

目录

机器学习视角下中国股票资产收益率可预测性研究

吴辉航　魏行空　张晓燕

（鑫苑房地产金融科技研究中心）

一、引言

股票收益率的可预测性一直以来都是金融学界研究的焦点。经典的有效市场理论认为股票市场不能被公开市场信息预测(Fama, 1970)，然而越来越多的研究表明，很多变量（例如：利率、通货膨胀、投资者情绪、方差风险溢价等）都能显著的预测未来的股票市场收益率(Bollerslev et al., 2014; Ang & Bekaert, 2007; Campbell & Thompson, 2008)。除了市场收益率，能够预测横截面个股的收益率预测的股票特征更是超过 400 个，被戏称为“因子动物园”(Cochrane, 2011; Harvey et al., 2016; Green et al., 2017)。在有了这么多因子后，个股收益率到底能在多大程度上被预测？到底哪些股票特征真正为样本外收益率预测提供了有效信息？这些预测结果能够用于股票资产配置并赚取超额收益吗？探索以上问题在中国资本市场的答案对于提升中国股票市场 54 万亿资金的有效配置至关重要。

研究中国股票样本外收益率可预测性的难点有以下三点。第一，影响股票收益率的因素非常多，且信噪比非常低，在这种面临高维稀

疏矩阵的情况下，传统计量经济模型会拟合过多的噪音，导致十分难以提取有效信息。第二，股票预测特征变量与股票收益率之间的函数关系并不确定(Campbell & Cochrane, 1999; He & Krishnamurthy, 2013)，如何捕捉预测变量与收益率之间的非线性结构是第二个难点。第三，中国股票市场从成立到现在只有短短的二十几年，股票市场制度依然处于不断完善的阶段，有着自身的特殊性。在中国股票市场，构造有预测能力的股票特征，并探索哪些个股特征包含的信息含量更高都是十分有挑战性的问题。

机器学习模型在降维、惩罚项和泛函数等技术上的突破在解决以上前两个问题上具有天然的优越性，最近很多论文探索了不同类型的机器学习算法在股票收益率预测的效果。第一类是金融学中较为常用的降维类模型，这类模型的优点是既能将高维度数据压缩成低维，同时还能保留较多的信息。例如：Rapach & Zhou (2018) 和 Maio & Philip (2015) 基于主成分分析的方法使用美国宏观变量来预测股票市场未来收益率；Kelly & Pruitt (2015) 基于最小偏二乘模型使用风格因子收益率资产组合来预测预测股票市场。第二类是带惩罚项的线性模型，其优点是通过加入惩罚项，降低噪音信息的因子荷载，从而提高预测效果。例如 Chinco et al. (2018) 基于套索回归（LASSO）分析了一分钟频率的个股收益率预测。第三类是非线性模型，这类模型的优点在于能够基于历史数据信息拟合预测变量与收益率之间的非线性结构。例如有学者基于随机森林、模糊神经网络和长短期记忆神经网络模型等人工智能算法后，检验了技术和宏观预测因

子在日度股票价格收益率预测的效果外 R 方(Fischer & Krauss, 2018; Sirignano et al., 2018; Bao et al., 2017; Butaru et al., 2016)。Gu et al. (2019a;2019b) 探索了神经网络模型、自编码机等深度学习模型在个股月度收益率的效果，获得非常好的样本外预测准确率。由于以上方面的优势，机器学习技术已经成为金融领域中的应用前沿之一，特别是在预测金融市场运动、处理文本信息、改进交易策略方面 (苏治等, 2017)。

中国股票市场依然处于不断发展和完善的阶段，不成熟的市场是不是更加容易被预测？很多国内学者也尝试结合机器学习技术解释中国股票市场的预期收益率预测问题。姜富伟等(2011)研究了中国市场投资组合和根据公司行业、规模、面值市值比和股权集中度等划分的各种成分投资组合的股票收益的可预测性；陈卫华和徐国祥(2018)发现深度学习预测沪深 300 指数的效果明显好于传统计量经济学模型；李斌等(2017, 2019)分别采用了支持向量机、神经网络、Adaboost 等机器学习算法，利用 19 项技术指标预测股价方向，发现基于机器学习算法预测所构建的投资组合也确实能取得更好的投资收益。现有文献并没有回答机器学习算法到底能在多大程度上预测中国股票横截面股票收益率，这个问题的探索有助于深入了解中国股票市场的运行特点。

本文尝试引入机器学习的方法来探索收益率可预测问题在中国个股资产收益率的答案。具体而言，本文首先基于 1997 年 1 月到 2019 年 12 月中国股市日度收益率交易数据，构造了文献中对股票横截面

收益率有预测能力的 108 个交易类股票异象性特征；其次，本文比较了传统计量经济学模型与最小偏二乘回归、主成分回归、弹性网络回归、随机森林、梯度提升树和神经网络模型 6 大主流机器学习算法在 A 股个股样本外可预测性问题上的表现；再次，本文详细分析了动量类、流动性和波动率三大不同类别股票异象性特征在中国股票横截面收益率预测的重要性排序；最后，本文根据股票预测收益率构建交易策略，探索机器学习算法的实际经济价值。

本文研究的主要发现有三点：（1）机器学习算法能够显著提升传统计量经济学模型的样本外预测结果。OLS 模型的样本外预测 R 方仅为-0.35%，而所有机器学习模型的样本外预测 R 方都为正，预测效果都在统计上显著的好于 OLS 模型，其中最好的两层神经网络模型的样本外 R 方高达 0.76%；（2）机器学习算法构建的交易策略能创造显著的经济意义。两层神经网络等权（市值）加权多空策略资产组合的绩效表现最好，在样本外测试时间 2010 年到 2019 年 12 月期间，平均能获得 3.03%（2.94%）的月度收益，月度波动率为 4.65%（6.88%），年化夏普比率为 2.26（1.48），经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 3.03（2.95）。（3）中国股市中流动性的指标对未来收益率的预测效果最好，其中成交量的方差（vdtv1）、换手率的方差变量（vturn）、去零交易日调整后换手率（LM1）等三个流动性指标的重要性排名靠前，平均重要性分别为 7.00%、3.79%、3.30%。

本文的创新点和贡献主要体现在以下三点：

第一，构建了与交易数据相关的 108 个中国股票异象性特征(因

子)。目前在学术期刊正式发表已表明对股票收益率有预测能力的特征数量多达 400+(Hou et al., 2019)，然而大多数发现的因子都是基于美国股票市场。中美两个股票市场在金融规律上有些很多共同的特点，但是中国股票市场由于特殊的制度环境、发展阶段也必然有其特殊性，因此需要进一步检验不同的因子在中国市场上表现情况是怎样的。Gu et al. (2019a) 研究发现美国个股预测能力最强的因子是动量类因子。本文基于中国日度收益率交易数据，重新构造了美国经典文献中的异象性特征。

第二，比较了不同机器学习算法在中国个股资产收益率预测中的效果。已经有研究表明在美国股票市场个股收益率预测问题上，机器学习算法能够显著的改进传统计量经济学的预测结果，获得更好的预测。那么到底在中国股票市场利用个股历史数据来预测个股收益率能做获得多高的准确率呢？机器学习算法又能否比传统计量经济学方法获得更好的样本外预测结果呢？不同的机器学习算法里面哪些算法表现会更好？是不是越复杂的模型预测效果越好呢？本文清晰揭示了机器学习算法能够提升传统计量经济学方法背后的经济原理。

第三，根据股票预测收益率构建交易策略，探索机器学习算法的实际经济价值。机器学习技术作为人工智能核心技术之一，历史性地站在了时代的风口，将对人类经济社会发展带来智能化浪潮的颠覆性猛烈冲击。全球各国都不遗余力的大力推动人工智能技术在各个行业中的应用，中国政府也高度重视。2019 年 8 月，中国人民银行印发《金融科技(FinTech)发展规划(2019-2021 年)》中明确指出金融科

技发展的重点任务之一是，合理运用金融科技手段丰富服务渠道、完善产品供给、降低服务成本、优化融资服务，提升金融服务质量与效率，使金融科技创新成果更好地惠及百姓民生。尽管如今金融科技正在如火如荼的发展，本文探索了人工智能技术如何在金融投资产业中落地。

二、数据说明

2.1 数据来源

本文使用的股票收益率、股本和财务报表数据均来源于 Wind 金融数据库。本文选取的数据 1997 年 1 月至 2019 年 12 月，虽然上海证券交易所在 1991 年就有交易记录了，但是 1996 年底上海证券交易所决定对证券交易方式进行了重大调整，其中包括设定 10%的涨跌停板沿用至今（Hu et al.，2019）。鉴于这个交易规则对股票收益率存在系统性影响，所以本文研究选取的数据区间为 1997 年 1 月开始。

本文以沪深两市上市并交易的 A 股为研究对象。A 股包括上海、深圳两市以人民币计价交易的所有股票，具体有上海主板股票（600 开头），深圳主板股票（000 开头）、深圳中小板股票（002 开头）、深圳创业板股票（300 开头）。为了保证数据库数据的准确性，我们还会结合国泰安数据的相同指标，对 Wind 数据库的数据完整性和准确性进行对比研究，尽量减少由于数据错误导致的模型构建失败问题。本文使用的股票收益率数据为考虑现金股利在投资的股票月度收益率。

本文使用的 FF3 和 FF5 因子来源于国泰安数据库，无风险收益率数据为一年定期存款利率的月度收益率，数据来源为国泰安数据库。

2.2 变量构造

本文参考 Hou et al.(2019) 和 Hou et al. (2020) 文章对股票异象性特征的构造方法，还原了美国股票市场至今在文献中发现全部的 108 个量价相关的异象性特征。本文所有的与交易数据相关的异象性特征可以分为五大类：1）波动率（风险）类，例如 beta、波动率、异质性波动率等，共计 37 个；2）流动性类，例如规模、换手率、Amihud 等，共计 23 个；3）动量类，例如 11 个月动量、6 个月动量、动量的变化、动量的残差等，共计 9 个；4）财务类，例如净资产收益率、毛利率、资产增长率等，共计 31 个；5）价值类，例如市净率、红利与股价比等，共计 8 个。

沿用美国股票异象性因子的原因在于，中国股票市场建立了完整的交易制度，因此部分美国股票市场的经济规律在中国也许也是成立的，例如规模和价值因子的规律在中国依然成立(Liu et al., 2019)，巴菲特价值投资策略在中国股票市场依然适用(胡熠、顾明, 2018)。然而，中国作为发展中国家，其股票市场机制依然处于不断完善的阶段，自然会与发达国家成熟的股票市场不同。此外中国股票市场还有着很多特殊的规章制度，例如 IPO、涨跌停板、T+1 等等，这些特殊的规章制度也会对中国股票预期收益率产生影响。这也导致很多在美国文献中非常显著的预测因子，例如：动量因子(Asness et al.,

2013)、投资因子(Li & Zhang, 2010)在中国股票市场可能不显著。

具体指标构建说明见表 1：异象性因子构造说明。

表 1：异象性因子构造说明

No.	Name	因子名称	构建说明	数量
Panel A. 流动性类因子（23 个）				
1	Size	企业市值	参考 Liu et al. (2019)，月末收盘价乘以总股本（流通 A 股）	2
2	Turn	换手率	参考 Liu et al. (2019)，基于过去 1、6、12 个月的日度换手率的平均值，其中日度换手率等于交易量除以总股本	3
3	vturn	换手率的方差	参考 Chordia et al. (2001)，基于过去 1、6、12 个月的日度换手率计算换手率的方差	3
4	dtv	成交量	过去 1、6、12 个月的交易量	3
5	vdtv	成交量的方差	过去 1、6、12 个月的交易量的方差	3
6	Ami	Ami 流动性	过去 1、6、12 个日度收益的绝对值除以交易量来度量流动性	3
7	Lm	去零交易日调整后换手率	参考 Liu (2006) ，基于过去 1、6、12 个月的去零交易日调整后换手率	3
8	mdr	最大日度回报	平均 5 日最高回报	1
9	Pr	股价	参考 Miller and Scholes (1982)，月底股票价格	1
10	abturn	异常换手率	参考 Liu et al. (2019)，过去一个月平均换手率与过去一年换手率之差	1
Panel B. 波动率（风险）类因子（37 个）				
1	idvc	异质性波动率-CAPM	参考 Ang et al. (2006)，基于过去 1、6、12 个月的日度收益率计算 CAPM 模型下个股的异质性波动率	3
2	idvcff	异质性波动率-FF3	参考 Ang et al. (2006)，基于过去 1、6、12 个月的日度收益率计算 FF3 模型下个股的异质性波动率	3
3	tv	总波动率	参考 Ang et al. (2006)，基于过去 1、6、12 个月的日度收益率计算总波动率	3
4	idsc	异质性波动率偏度-CAPM	参考 Boyer et al. (2009)，基于过去 1、6、12 个月的日度收益率计算 CAPM 模型下个股的异质性波动率偏度	3
5	idff	异质性波动率偏度-FF3	参考 Boyer et al. (2009)，基于过去 1、6、12 个月的日度收益率计算 FF3 模型下个股的异质性波动率偏度	3
6	Ts	总偏度	参考 Amaya et al. (2015)，基于过去 1、6、12 个月的日度收益率计算总偏度	3
7	cs	协偏度	参考 Harvey and Siddique (2000)，基于过去 1、6、12 个月的日度收益率计算协偏度	3
8	betam1	月度贝塔	参考 Fama and MacBeth (1973)，基于过去 1、6、12 个月的月度收益率计算市场贝塔	3

9	beta	日度贝塔	参考 Fama and MacBeth (1973)，基于过去 1、6、12 个月的日度收益率计算市场贝塔	3
10	dbeta	下行贝塔	参考 Ang et al. (2006b)，基于过去 1、6、12 个月的日度收益率计算熊市时期的下行贝塔	3
11	betaFP	FP 贝塔	参考 Frazzini and Pedersen (2013)，基于过去 1、6、12 个月的日度收益率计算贝塔	3
12	tailr	尾部风险	参考 Kelly and Jiang(2014)，计算股票尾部风险	1
13	betaDM	Dimsion	参考 Dimson (1979)，基于过去 1、6、12 个月的日度收益率计算贝塔	3
Panel C. 动量类因子（9 个）				
1	Mom1	反转	参考 Liu et al. (2019)，过去 1 个月的累计收益率	1
2	Mom6	6 个月动量	参考 Jegadeesh and Titman (1993)，过去 6 个月的累计收益率，并剔除最近的 1 个月	1
3	Mom9	9 个月动量	参考 Jegadeesh and Titman (1993)，过去 9 个月的累计收益率，并剔除最近的 1 个月	1
4	Mom11	12 个月动量	参考 Jegadeesh and Titman (1993)，过去 12 个月的累计收益率，并剔除最近的 1 个月	1
5	Mom24	长期反转	参考 Jegadeesh and Titman (1993)，过去 24 个月的累计收益率，并剔除最近的 1 个月	1
6	Mchg	动量变化	参考 Gettleman and Marks (2006)，过去 1 到 6 个月的累计收益率减去过去 7 到 12 个月的累计收益率	1
7	imom11	11 月动量残差	过去 11 个月 FF3 动量残差	1
8	imom6	6 月动量残差	过去 6 个月 FF3 动量残差	1
9	52w	52 周最高值	52 周月度股价的最高值	1
Panel D. 财务类因子（31 个）				
1	rdmq	研发支出占比 1	季度研发支出除以市值	1
2	rdsq	研发支出占比 2	季度研发支出除以营业收入	1
3	age	企业年龄	企业上市时间	1
4	cta	现金占比	企业现金及其等价物除以总资产	1
5	olq	运营杠杆	季度运营支出除以总资产	1
6	vcf	资金波动率	季度现金流的波动率	1
7	tan	无形资产率	季度无形资产占总资产比	1
8	cagq	流动资产增长率 1	季度流动资产增长率	1
9	ncagq	非流动资产增长率	季度非流动资产增长率	1
10	cashgq	现金流增长率	季度现金流增长率	1
11	fagq	固定资产增长率	季度固定资产增长率	1
12	agq	总资产增长率	季度总资产增长率	1
13	nccagq	流动资产增长率 2	流动资产（不含现金）增长率	1
14	oagq	其它资产增长率	季度其它资产增长率	1

15	roe	净资产收益率	季度公司税后利润除以净资产	1
16	droe	净资产收益率变化	前后两期净资产收益率差值	1
17	roa	总资产收益率	季度公司税后利润除以总资产	1
18	droa	总资产收益率变化	前后两期总资产收益率差值	1
19	rnaq	运营资产收益率	季度公司收入除以运营资产	1
20	pmaq	净利润率	季度公司收入减成本除以收入	1
21	atoq	资产周转率	季度公司收入除以总资产	1
22	ctq	运营资产周转率	季度公司收入除以运营资产	1
23	gplaq	毛利率	季度公司毛利润除以上一期期末总资产	1
24	opleq	主营业务利润率 1	季度公司主营业务利润除以上一期期末所有者权益	1
25	oplaq	主营业务利润率 2	季度公司主营业务利润除以上一期期末总资产	1
26	tbiq	会税差异	季度应纳税收入除以账面收入	1
27	blq	杠杆率	季度负债除以总资产	1
28	sgq	销售增长率	季度主营业务收入增长率	1
29	fscoreq	F 值	参考 Piotroski (2000)，基本面综合评分 F	1
30	oscoreq	0 值	参考 Ohlson (1980)，基本面综合评分 0	1
31	zscoreq	Z 值	参考 Dichev (1998)，基本面综合评分 Z	1
Panel E. 价值性类因子（8 个）				
1	am	账面总资产与市值比	季度总资产除以企业市值	1
2	dm	账面总负债与市值比	季度账面总负债与市值比	1
3	bm	账面总权益与市值比	季度账面总权益与市值比	1
4	ep	盈利与股价比	季度企业每股盈余除以股价	1
5	ocfp	现金与股价比 1	季度企业经营现金流除以股价	1
6	cfp	现金与股价比 2	季度企业现金流除以股价	1
7	sp	销售与股价比 2	季度企业销售收入除以股价	1
8	dp	红利与股价比	季度企业分红除以股价	1

2.3 特别处理

1. 删除特别样本

中国现代股票市场从 1990 年上海、证券交易所成立至今共计 30 年。这 30 年时间中国的股票市场制度从无到与国际接轨，几乎走完了西方发达国家股票市场 200 多年的发展历程，经历了多变的制度变迁。很多重大的股票市场制度可能会导致微观金融市场结构的变迁。例如：中国股票发行是审核制，由于证监会对股票 IPO 发行定价审核有着明确的规定，不可以超过 23 倍的发行市盈率，这就导致了中国股票市场存在 IPO 抑价问题(Lee et al.，2019)。这些由于外生政策扭曲的非市场定价行为，会导致股票收益率价格的异常，需要在数据清洗的步骤剔除。除此之外，还有壳资源、ST 制度、股权分置改革和暂停上市等特殊的制度规定也会导致股票收益率不符合正常的市场定价规律，导致股票收益率产生异常，都需要细致清洗。

为了解决以上问题，本文参考 Liu et al. (2019)处理方式在原始样本中剔除了以下五种特殊的股票：（1）被特别处理的股票（ST、ST*、PT）；（2）过去 12 个月交易日小于 120 天；（3）过去一个月小于当月总交易天数 75%的股票 ；（4）30%市值最小的股票（市值用收盘价乘以总股本计算）；（5）最后一个交易日换仓时停牌或一字涨停等无法交易的股票。

2. 财务因子构建

本文的财务因子主要来自企业的三张调整前财务报告主表，我们利用 wind 数据库中提供的财务报表公布日期作为索引，与收益率数

据进行合并。由于财务报告为季度频率，我们通过向下填充的方式变频到月度，这样我们能在获取每个月企业最新的财务信息的同事又避免用到未来信息。此外，由于中美会计准则差异的原因，我们对一些财务指标构建的预测因子进行了调整。具体指标构建说明见表 1：异象性因子构造说明。

3. 标准化处理

本文经过以上特别样本删除后，如果收益率依然存在异常值，我们不再进行调整。对于构建好的交易异象性特征，本文采取下面横截面排序标准化算法进行处理。

$$c_{i,t} = \frac{2}{N+1} CSrank\left(\mathrm{c}_{i,t}^{r}\right) - 1$$

其中：$c_{i,t}$代表标准化过后的交易异象性特征；$\mathrm{c}_{i,t}^{r}$代表标准化前交易异象性特征；$CSrank$代表每个月横截面排序函数；N 代表本月上市公司数。通过使用该横截面排序算法可以将所有指标值缩放到[-1,1]的值内，使用该标准化方法有以下三点好处：1）移除不同财务指标或公司特征的量纲差异，使得不同财务指标横向可比；2）移除财务指标或公司特征数据异常值给模型带来的影响；3）移除量纲的差异能大大加快一些机器学习算法的收敛速度。如果某观测值某月收益率缺失（比如整月停牌），我们将删除该观测值，如果交易异象性特征值存在缺失，本文采用每个月在横截面生成该变量的中位数进行替换操作。

三、模型构建

本文的基准的实证模型从最一般的函数形式出发，资产的超额收益可以由以下模型刻画：

$$r_{i,t+1} = E_t(r_{i,t+1}) + \varepsilon_{i,t+1} \qquad (1)$$

其中：

$$E_t(r_{i,t+1}) = g^*(z_{i,t}) \qquad (2)$$

$r_{i,t+1}$代表第 i 只股票(i = 1, ..., N)第 t+1 个月(t = 1, ..., T)的真实超额回报率；$E_t(r_{i,t+1})$ 代表在根据 t 时期的信息合集，在第 t 期对 t+1 期股票超额收益率的期望收益；$z_{i,t}$代表第 i 只股票 t 时期的预测变量（公司特征）合集，是一个 P 维向量。

$g^*()$是一个灵活的函数形式，用来建立$z_{i,t}$与$E_t(r_{i,t+1})$之间的映射关系。

当$g^*(.)$为线性函数形式时，该模型即为最基本的 OLS 回归，该结果将作为基准模型提供比较的参考基准，此外我们还将考虑 6 种不同的机器学习算法：最小偏二乘回归（以下缩写 PLS）、主成分回归（以下缩写 PCR）、弹性网络（以下缩写 Enet）、随机森林（以下缩写 RF）、梯度提升树（以下缩写 GBRT）、神经网络模型（以下缩写 NN），对比不同机器学习模型的预测效果。

具体的机器学习算法实现的伪代码和统计理论上的特性请参考 Gu et al. (2019a)的附录 B。

四、实证结果

4.1 个股的可预测性实证结果

表 2 展示了 R 方度量下不同机器学习模型样本外预测准确度。其中 OLS3 代表基于 OLS + Huber Loss 方程（5）且仅使用企业市值、总波动率、反转三个特征进行拟合的结果。PLS、PCR、ENet、RF、GBRT 分别代表使用最小偏二乘回归、主成分回归、弹性网络、随机森林和梯度提升树模型使用所有变量拟合的结果。NN1 到 NN5 分别代表使用 1 到 5 层神经网络模型使用所有变量拟合的结果。

表 2：R 方度量下不同机器学习模型样本外预测准确度（样本外测试时间：2010 年到 2019 年 12 月）

	OLS3	PLS	PCR	Enet	RF	GBRT	NN1	NN2	NN3	NN4	NN5
ALL	-0.35	0.43	0.17	0.31	0.35	0.31	0.27	0.76	0.21	0.67	0.17
Top 300	0.08	0.19	0.43	0.54	0.43	0.57	0.02	0.15	0.04	0.05	0.08
Bottom 300	-0.53	0.67	0.28	0.02	0.22	0.16	0.36	0.98	0.31	0.91	0.15

其中 All 是指全部样本的样本外 R 方，Top （Bottom）300 是指最大(小)的 300 只股票预测结果。OLS 模型的全样本 R 方仅为-0.35%，这说明基于传统的 OLS 模型，中国 A 股个股的收益率的预测十分困难，OLS 模型的预测结果在统计上还不如直接用 0 作为预测结果跟接近真实值。这也说明了中国 A 股个股收益率难被以预测。

反观其他机器学习算法所有的模型的样本外 R 方都为正，其中

PLS、PCR 和 Enet 三类线性模型的样本外 R 方分别为 0.43%、0.17%和 0.31%。这说明变量信息压缩和添加惩罚项两种机器学习方法都能显著改善传统 OLS 模型估计不稳定的问题，从而提升模型的样本外预测结果。随机森林和提升树算法的样本外 R 方分别为 0.35%和 0.31%，这说明基于树类机器学习算法的非线性特征也能提升 OLS 模型的样本外预测结果。

NN1 到 NN5 五类模型的样本外 R 方分别为 0.27%、0.76%、0.21%、0.67%和 0.17%。这说明：1）基于神经网络类机器学习算法的非线性特征也能提升 OLS 模型的样本外预测结果；2）神经网络模型算法的样本外 R 方并没有展现出越复杂的模型越好的特征，其中两层神经网络模型的结果最好为 0.76%，而 5 层神经网络模型的结果却 0.17%。Top （Bottom）300 是指最大（小）的 300 只股票预测结果，最好的模型为 GBRT（NN2），样本外 R 方为 0.57%（0.98%）。本文的预测结果与美国文献类似，对比 Gu et al. (2019)基于美国机器学习的预测结果，其表现最好的随机森林的样本外 R 方为 0.33%。

4.2 机器学习选股策略绩效表现

本文的机器学习选股策略是在每个月的最后一个交易日根据所有模型预测的下一期股票收益率预测结果进行排序，根据排序的结果来构建不同的资产组合。样本外的测试时间为 2010 年 1 月到 2019 年 12 月。表 8 为不同机器学习模型等权加权构建资产组合的绩效表现，说明在中国规模因子依然是有效的。例如最好的 2 层神经网络模型，

多空资产组合策略平均能获得 3.03%的月度收益，月度标准差为 4.65%，年化夏普比率为 2.26。图 1 和图 2 展示了不同机器学习模型构建资产组合的累计收益率曲线。可以看到等权（市值）加权的机器学习资产组合的纯多头策略 10 年累计收益率(对数)约为 1.35(1.12)，而同期沪深 300 收益率收益率仅为 0.05。

表 8：不同机器学习模型等权加权构建资产组合的绩效表现（样本外测试时间：2010 年到 2019 年 12 月）

Panel A. 等权加权机器学习资产组合分组收益率											
Ret	Lo_10	2_Dec	3_Dec	4_Dec	5_Dec	6_Dec	7_Dec	8_Dec	9_Dec	Hi_10	H_L
OLS3	-0.88	0.29	0.49	0.50	0.41	0.31	0.56	0.66	0.64	0.91	1.79
PLS	-1.14	-0.47	0.07	0.18	0.45	0.65	0.71	0.93	1.15	1.36	2.50
PCR	-1.08	-0.21	0.07	0.24	0.43	0.66	0.77	0.82	1.02	1.18	2.26
ENet	-1.22	-0.44	0.02	0.23	0.58	0.61	0.76	0.88	1.06	1.42	2.64
RF	-1.29	-0.38	0.03	0.35	0.62	0.59	0.89	0.94	0.97	1.18	2.48
GBRT	-1.19	0.08	0.29	0.47	0.65	0.51	0.55	0.84	0.67	1.01	2.20
NN1	-1.46	-0.37	-0.09	0.34	0.52	0.57	0.71	0.90	1.14	1.63	3.09
NN2	-1.47	-0.49	0.09	0.28	0.41	0.61	0.82	0.96	1.12	1.56	3.03
NN3	-0.93	-0.25	0.02	0.10	0.38	0.46	0.75	0.91	0.97	1.48	2.42
NN4	-1.59	-0.54	0.07	0.29	0.55	0.71	0.86	0.90	1.15	1.45	3.04
NN5	-0.99	-0.11	0.19	0.37	0.57	0.73	0.62	0.75	0.81	0.95	1.94
Panel B. 等权加权机器学习资产组合分组标准差											
STD	Lo_10	2_Dec	3_Dec	4_Dec	5_Dec	6_Dec	7_Dec	8_Dec	9_Dec	Hi_10	H_L
OLS3	9.56	8.56	8.31	8.43	8.43	8.21	8.12	8.04	8.04	7.88	4.11
PLS	9.56	9.12	8.93	8.57	8.35	8.32	7.99	7.84	7.62	7.25	4.50
PCR	9.71	9.22	8.96	8.67	8.41	8.17	7.86	7.72	7.57	7.27	4.37
ENet	9.39	8.93	8.79	8.52	8.42	8.23	7.90	7.98	7.77	7.64	4.20
RF	9.76	9.32	8.97	8.75	8.60	8.32	7.99	7.73	7.43	6.87	4.89
GBRT	9.55	8.72	8.56	8.48	8.21	8.10	8.18	8.17	7.79	7.76	4.14
NN1	9.74	9.30	8.64	8.10	8.18	8.02	7.98	8.04	7.85	8.06	4.88
NN2	9.32	8.92	8.67	8.37	7.96	7.74	7.93	8.11	8.34	8.69	4.65
NN3	9.70	9.33	8.93	8.50	8.44	8.32	7.97	7.85	7.31	7.32	4.48
NN4	9.21	8.37	8.34	8.27	8.13	8.32	8.40	8.31	8.40	8.36	4.71
NN5	10.03	9.16	8.71	8.51	8.38	8.00	7.84	7.79	7.68	7.80	5.00
Panel C. 等权加权机器学习资产组合分组夏普比率											
SR	Lo_10	2_Dec	3_Dec	4_Dec	5_Dec	6_Dec	7_Dec	8_Dec	9_Dec	Hi_10	H_L
OLS3	-0.32	0.12	0.20	0.21	0.17	0.13	0.24	0.28	0.27	0.40	1.51
PLS	-0.41	-0.18	0.03	0.07	0.19	0.27	0.31	0.41	0.52	0.65	1.93
PCR	-0.39	-0.08	0.03	0.10	0.18	0.28	0.34	0.37	0.47	0.56	1.79

ENet	-0.45	-0.17	0.01	0.09	0.24	0.26	0.33	0.38	0.47	0.64	2.18
RF	-0.46	-0.14	0.01	0.14	0.25	0.25	0.38	0.42	0.45	0.60	1.75
GBRT	-0.43	0.03	0.12	0.19	0.28	0.22	0.23	0.36	0.30	0.45	1.84
NN1	-0.52	-0.14	-0.04	0.15	0.22	0.25	0.31	0.39	0.51	0.70	2.19
NN2	-0.55	-0.19	0.04	0.12	0.18	0.27	0.36	0.41	0.47	0.62	2.26
NN3	-0.33	-0.09	0.01	0.04	0.15	0.19	0.33	0.40	0.46	0.70	1.87
NN4	-0.60	-0.22	0.03	0.12	0.24	0.29	0.35	0.38	0.48	0.60	2.24
NN5	-0.34	-0.04	0.08	0.15	0.23	0.31	0.27	0.33	0.37	0.42	1.34

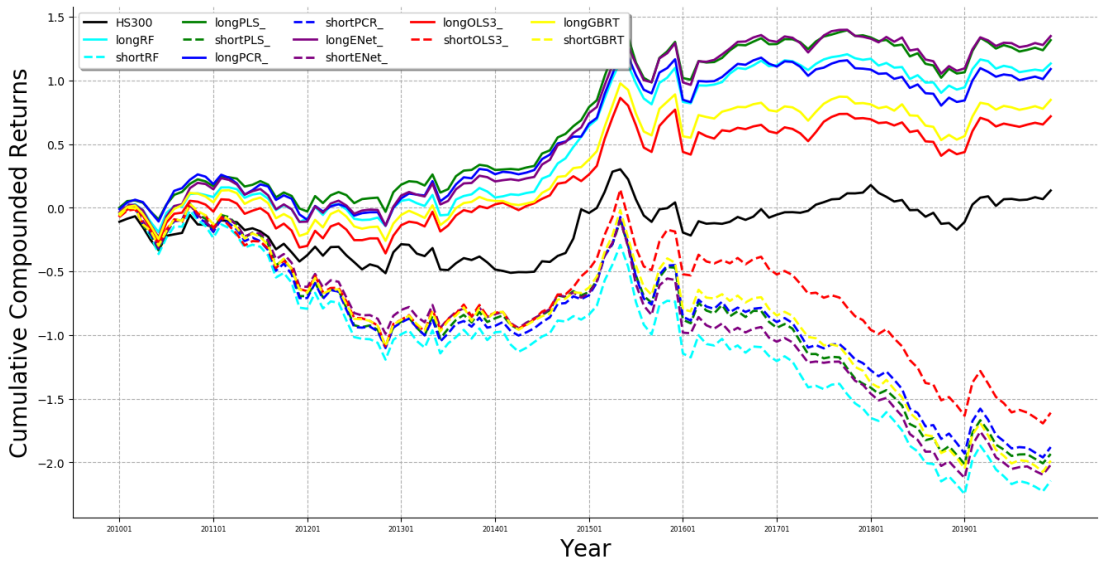

图 4：2010 年 1 月到 2019 年 12 月样本外机器学习资产组合策略累计收益率（等权加权）

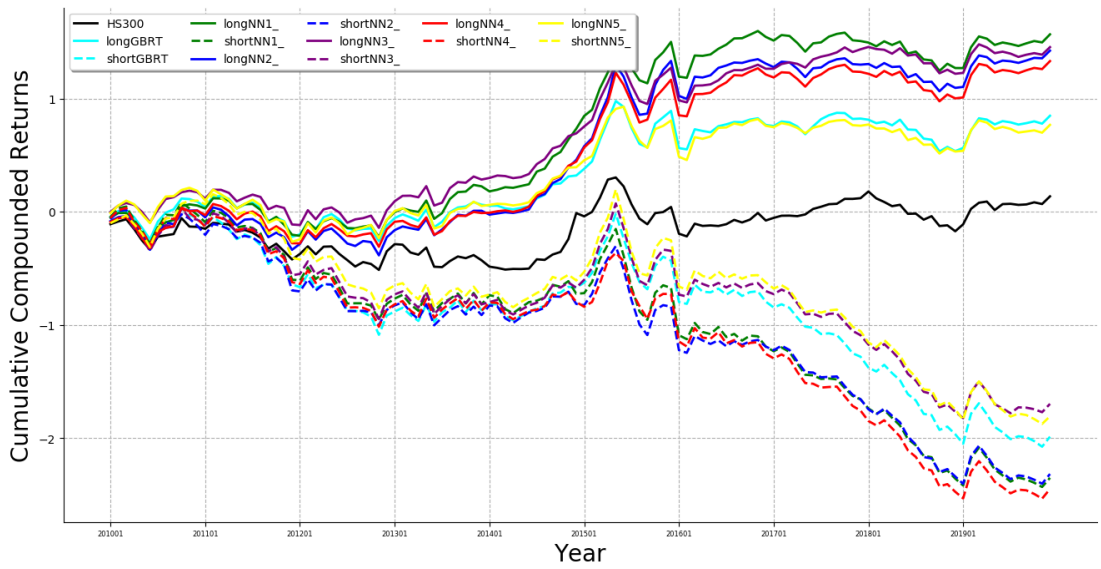

图 5：2010 年 1 月到 2019 年 12 月样本外机器学习资产组合策略累计收益率（等权加权）

五、研究结论

本文基于机器学习的视角检验了文献中提出的众多预测变量对中国个股收益率的预测能力。我们发现：（1）机器学习算法能够显著提升传统计量经济学模型的样本外预测结果。OLS 模型的样本外预测 R 方仅为-0.35%，而所有机器学习模型的样本外预测 R 方都为正，预测效果都在统计上显著的好于 OLS 模型，其中最好的两层神经网络模型的样本外 R 方高达 0.76%；（2）机器学习算法构建的交易策略能创造显著的经济意义。两层神经网络等权（市值）加权多空策略资产组合的绩效表现最好，在样本外测试时间 2010 年到 2019 年 12 月期间，平均能获得 3.03%（2.94%）的月度收益，月度波动率为 4.65%（6.88%），年化夏普比率为 2.26（1.48），经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 3.03（2.95）。（3）中国股市中流动性的指标对未来收益率的预测效果最好，其中成交量的方差（vdtv1）、换手率的方差变量（vturn）、去零交易日调整后换手率（LM1）等三个流动性指标的重要性排名靠前，平均重要性分别为 7.00%、3.79%、3.30%。这与中国 A 股市场停盘、T+1 等交易摩擦制度造成的非流动性资产溢价有关。厘清哪些股票特征能够有效的预测中国个股资产收益率有助于理解中国股票市场不同的交易性异象特征中的预测信息含量，有助于更加深入了解中国股票市场的运行特点。

参考文献

陈卫华、徐国祥,2018：《基于深度学习和股票论坛数据的股市波动率预测精度研究》,《管理世

界》，第 01 期。

胡熠、顾明, 2018：《巴菲特的阿尔法:来自中国股票市场的实证研究》,《管理世界》，第 08 期。

姜富伟、涂俊、Rapach David E.、Strauss Jack K.、周国富, 2011：《中国股票市场可预测性的实证研究》,《金融研究》，第 09 期。

苏治、卢曼、李德轩, 2017：《深度学习的金融实证应用:动态、贡献与展望》,《金融研究》，第 05 期。

李斌、林彦、唐闻轩, 2017：《MI-Tea:一套基于机器学习和技术分析的量化投资算法》,《系统工程理论与实践》，第 05 期。

李斌、邵新月、李玥阳，2019：《机器学习驱动的基本面量化投资研究》,《中国工业经济》，第 08 期。

Amaya, D., Christoffersen, P., Jacobs, K. and Vasquez, A., 2015, "Does Realized Skewness Predict the Cross-Section of Equity Returns?", *Journal of Financial Economics*, 118(1): 135-167.

Ang, A. and Bekaert, G., 2007, "Stock Return Predictability: Is It there?", *The Review of Financial Studies*, 20(3): 651-707.

Ang, A., Chen, J. and Xing, Y., 2006, "Downside Risk", *The Review of Financial Studies*, 19(4): 1191-1239.

Ang, A., Hodrick, R. J., Xing, Y. and Zhang, X., 2006, "The Cross-Section of Volatility and Expected Returns", *The Journal of Finance*, 61(1): 259-299.

Asness, C. S., Moskowitz, T. J. and Pedersen, L. H., 2013, "Value and Momentum Everywhere", *The Journal of Finance*, 68(3): 929-985.

Bao, W., Yue, J. and Rao, Y., 2017, "A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory", *Plos One*, 12(7): 18-24.

Bollerslev, T., Marrone, J., Xu, L. and Zhou, H., 2014, "Stock Return Predictability and Variance Risk Premia: Statistical Inference and International Evidence", *Journal of Financial and Quantitative Analysis*, 49(3): 633-661.

Boyer, B., Mitton, T. and Vorkink, K., 2010, "Expected Idiosyncratic Skewness", *The Review of Financial Studies*, 23(1): 169-202.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W. and Siddique, A., 2016, "Risk and Risk Management in the Credit Card Industry", *Journal of Banking & Finance*, 72(C): 218-239.

Campbell, J. Y. and Thompson, S. B., 2008, "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?", *The Review of Financial Studies*, 21(4): 1509-1531.

Chinco, A., Clark Joseph, A. D. and Ye, M., 2018, "Sparse Signals in the Cross-Section of Returns", *The Journal of Finance*.

Chordia, T., Subrahmanyam, A. and Anshuman, V. R., 2001, "Trading Activity and Expected Stock Returns", *Journal of Financial Economics*, 59(1): 3-32.

Cochrane, J. H., 2011, "Presidential Address: Discount Rates", *Journal of Finance*, 66(4): 1047-1108.

Corazza, M., Durbán, M., Grané, A., Perna, C. and Sibillo, M., 2018, Mathematical and Statistical Methods for Actuarial Sciences and Finance: Maf 2018, Springer.

Dichev, I. D., 1998, "Is the Risk of Bankruptcy a Systematic Risk?", *The Journal of Finance*, 53(3): 1131-1147.

Dimson, E., 1979, "Risk Measurement When Shares are Subject to Infrequent Trading", *Journal of Financial Economics*, 7(2): 197-226.

Fama, E. F., 1970, "Efficient Capital Markets: A Review of Theory and Empirical Work *", *Journal of Finance*, 25(2): 383-417.

Fama, E. F. and French, K. R., 1992, "The Cross-Section of Expected Stock Returns", *The Journal of Finance*, 47(2): 427-465.

Fama, E. F. and French, K. R., 2008, "Dissecting Anomalies", *The Journal of Finance*, 63(4): 1653-1678.

Fama, E. F. and Macbeth, J. D., 1973, "Risk, Return, and Equilibrium: Empirical Tests", *Journal of Political Economy*, 81(3): 607-636.

Fischer, T. and Krauss, C., 2018, "Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions", *European Journal of Operational Research*, 270(2): 654-669.

Frazzini, A. and Pedersen, L. H., 2014, "Betting Against Beta", *Journal of Financial Economics*, 111(1): 1-25.

Green, J., Hand, J. R. M. and Zhang, X. F., 2017, "The Characteristics that Provide Independent Information About Average U.S. Monthly Stock Returns", *The Review of Financial Studies*, 30(12): 4389-4436.

Gu, S., Kelly, B. and Xiu, D., 2019a, "Autoencoder Asset Pricing Models", *Working Paper*.

Gu, S., Kelly, B. and Xiu, D., 2019b, "Empirical Asset Pricing Via Machine Learning", *Working Paper*.

Harvey, C. R. and Siddique, A., 2000, "Conditional Skewness in Asset Pricing Tests", *The Journal of Finance*, 55(3): 1263-1295.

Harvey, C. R., Liu, Y. and Zhu, H., 2016, "… and the Cross-Section of Expected Returns", *The Review of Financial Studies*, 29(1): 5-68.

Horvitz, E. and Mulligan, D., 2015, "Machine Learning: Trends, Perspectives, and Prospects", *Science*, 349(6245): 253-255.

Hou, K., Xue, C. and Zhang, L., 2019, "Replicating Anomalies", *The Review of Financial Studies*, (forthcoming).

Kewei Hou, Fang Qiao, and Xiaoyan Zhang(2019), "Finding Anomalies in China", *Working Paper*.

Hsu, J., Viswanathan, V., Wang, M. and Wool, P., 2018, "Anomalies in Chinese a-Shares", *The Journal of Portfolio Management*, 44(7): 108.

Huber, P. J., 1992, Robust Estimation of a Location Parameter.

Jegadeesh, N. and Titman, S., 1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency", *The Journal of Finance*, 48(1): 65-91.

Kelly, B. T., Pruitt, S. and Su, Y., 2019, "Characteristics are Covariances: A Unified Model of Risk and Return", *Journal of Financial Economics*.

Kelly, B. and Jiang, H., 2014, "Tail Risk and Asset Prices", *The Review of Financial Studies*, 27(10): 2841-2871.

Kelly, B. and Pruitt, S., 2013, "Market Expectations in the Cross-Section of Present Values", *The Journal of Finance*, 68(5): 1721-1756.

Kelly, B. and Pruitt, S., 2015, "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors", *Journal of Econometrics*, 186(2): 294-316.

Khandani, A. E., Kim, A. J. and Lo, A., 2010, "Consumer Credit-Risk Models Via Machine-Learning Algorithms", *Journal of Banking & Finance*, 34(11): 2767-2787.

Lee, C. M. C., Qu, Y. and Shen, T., 2019, "Going Public in China: Reverse Mergers Versus Ipos", *Journal of Corporate Finance*, 58: 92-111.

Li, D. and Zhang, L., 2010, "Does Q-Theory with Investment Frictions Explain Anomalies in the Cross Section of Returns?", *Journal of Financial Economics*, 98(2): 297-314.

Li, F., Zhang, H. and Zheng, D., 2018, "Seasonality in the Cross Section of Stock Returns: Advanced

Markets Versus Emerging Markets", *Journal of Empirical Finance*, 49: 263-281.

Liu, J., Stambaugh, R. F. and Yuan, Y., 2019, "Size and Value in China", *Journal of Financial Economics*, (forthcoming).

Liu, W., 2006, "A Liquidity-Augmented Capital Asset Pricing Model", *Journal of Financial Economics*, 82(3): 631-671.

Lou, D., 2014, "Attracting Investor Attention through Advertising", *The Review of Financial Studies*, 27(6): 1797-1829.

Lou, D., Polk, C. and Skouras, S., 2019, "A Tug of War: Overnight Versus Intraday Expected Returns", *Journal of Financial Economics*, 134(1): 192-213.

Maio, P. and Philip, D., 2015, "Macro Variables and the Components of Stock Returns", *Journal of Empirical Finance*, 33: 287-308.

Markowitz, H., 1952, "Portfolio Selection", *The Journal of Finance*, 7(1): 77-91.

Merton, R. C., 1973, "An Intertemporal Capital Asset Pricing Model", *Econometrica*, 41(5): 867-887.

Ohlson, J. A., 1980, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*: 109-131.

Piotroski, J. D., 2000, "Value Investing: The Use of Historical Financial Statement Information to Separate Winners From Losers", *Journal of Accounting Research*: 1-41.

Rapach, D. and Zhou, G., 2018, "Sparse Macro Factors", *Working Paper*, (October 1).

Rapach, D., Strauss, J. and Zhou, G., 2013, "International Stock Return Predictability: What is the Role of the United States?", *The Journal of Finance*, 68(4): 1633-1662.

Sirignano, J., Sadhwani, A. and Giesecke, K., 2018, "Deep Learning for Mortgage Risk", arXiv.org.

Welch, I. and Goyal, A., 2008, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction", *The Review of Financial Studies*, 21(4): 1455-1508.