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

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 8 月 A 股市场的 74 个交易类股票异象性特征，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 2.975%(2.459%)的月度收益，月度波动率为 4.276%(5.973%)，年化夏普比率为 2.410(1.426)，经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 2.810(2.333)。

关键词：收益率预测；机器学习；金融科技；异象性因子
JEL 分类号: G10; G14　**文献标识码:** A

作者信息：吴辉航，清华大学五道口金融学院，电子邮箱：wuhh@pbcsf.tsinghua.edu.cn，联系电话：18512171805；魏行空，清华大学五道口金融学院，电子邮箱：weixk@pbcsf.tsinghua.edu.cn，联系电话：13552864664；张晓燕(通讯作者)，清华大学五道口金融学院，电子邮箱：zhangxiaoyan@pbcsf.tsinghua.edu.cn，联系电话：18510388372。
邮寄地址：北京市海淀区成府路 43 号，邮编 100083。

本文感谢国家自然科学基金重大项目(71790605)的资助。**张晓燕为本文通讯作者。**

机器学习视角下中国股票资产收益率可预测性研究

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 8 月 A 股市场的 74 个交易类股票异象性特征，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 2.975%(2.459%)的月度收益，月度波动率为 4.276%(5.973%)，年化夏普比率为 2.410(1.426)，经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 2.810(2.333)。

关键词：收益率预测；机器学习；金融科技；异象性因子
JEL 分类号: G10; G14　**文献标识码:** A

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[①]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)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 8 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 74 个交易类股票异象性特征[②]；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-0.178%，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，这意味着基于个股交易数据中蕴含的历史信息能够预测未来的个股资产收益率。其中最好的两层神经网络模型的样本外 R 方高达 0.633%；(2)中国股市中流动性的指标对未来收益率的预测效果最好，成交量的方差、换手率的方差、Amihud 流动性这三个流动性指标的重要性排名靠前。(3)**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 8 月期间，平均能获得 2.975%(2.459%)的月度收益，月度波动率为 4.276(5.973)，年化夏普比率为 2.41(1.426)，经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 2.81(2.33)。

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 74 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)》[③]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。

本文余下部分的安排如下。第二部分是数据说明，第三部分介绍本文使用的主要机器学习方法，第四部分是主要的实证结果，最后总结全文。

二、数据说明

（一） 数据来源

本文使用的股票收益率数据和股本数据均来源于 Wind 金融数据库。本文选取的数据 1997 年 1 月至 2019 年 8 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。

文本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。

（二） 变量构造

本文参考 Hou et al.(2019) 和 Qiao (2019) 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 74 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三大类：(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 22 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 15 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。本文仅基于股票交易数据构造了所有预测因子，并没有纳入财务信息，原因主要有两点：第一，本文仅探索过去的交易信息能否预测未来的股票收益率，因此没有纳入额外的财务信息。第二，由于中美会计准则差异的原因，对于很多基于财务指标构建的预测因子并无法完全复现。具体调整细节和指标构建说明见表 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(22 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
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(15 个)				
1	Srev	反转	参考 Liu et al. (2019)，过去 1 个月的累计收益率	1
2	Mom_6	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	Mom_6	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	Mom_12	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	Lrev	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	im11	11 月动量残差	过去 11 个月 FF3 动量残差	1
8	im6	6 月动量残差	过去 6 个月 FF3 动量残差	1
9	ra	季节性 1	参考 Heston and Sadka (2008)，过去 t-12 个月收益率	2
10	rn	季节性 2	参考 Heston and Sadka (2008)，过去 t-11 到 t-1 个月收益率的平均值	2
11	intra	日内收益率	参考 Lou et al.(2019)，过去一个月股票的日内累计收益率	1
12	night	隔夜收益率	参考 Lou et al.(2019)，过去一个月股票的隔夜累计收益率	1
13	52w	52 周最高值	52 周月度股价的最高值	1

（三） 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]；(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. 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。

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

其中：$c_{i,t}$代表标准化过后的交易异象性特征；$c_{i,t}^{r}$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。

（四） 变量描述性统计

1997 年 1 月到 2019 年 8 月收益率有效的样本数如图 1 所示，其中蓝色阴影面积内的为有效样本数量，总面积内的为全部样本数量。平均每个交易日有 4.93%的股票收益率缺失，75%分位数 7.39%，最大值 51.85%(由于 2015 年股灾期间千股停牌导致)。会有以下两种情况导致数据缺失：股票停牌和股票暂停上市(很少)。如果一只股票当月完全停牌将会被直接

删除。A 股上市公司数量从 1997 年的 500 家持续上升，截止至 2019 年 6 月 30 日，总计 3732 只股票发行，103 只股票退市，A 股上市公司目前市场上共有 3629 只股票。表 2 展示了 1997 年 1 月到 2019 年 8 月分年度 A 股日度收益率描述性统计。

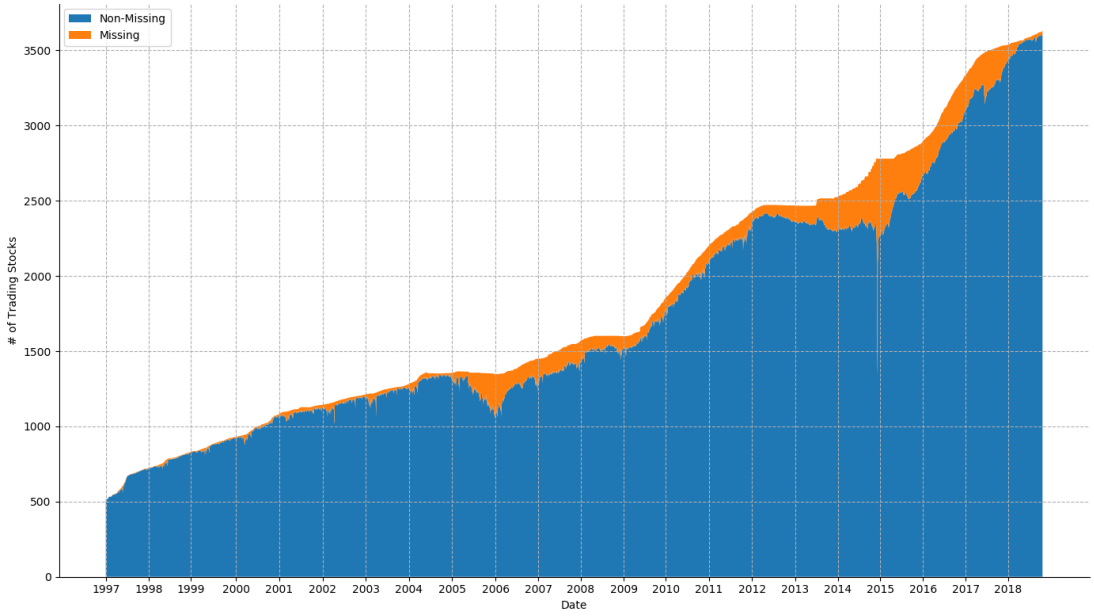


图 1：1997 年 1 月到 2019 年 8 月有效样本数

表 2：1997 年 1 月到 2019 年 8 月 A 股日度收益率描述性统计 (单位：%)

	min	median	max	mean	std	skewness	kurtosis	OBV
1997	-19.57	0.07	12.73	0.12	3.36	-0.04	1.22	153202
1998	-45.27	-0.09	12.54	0.05	2.66	0.34	2.44	188997
1999	-15.68	-0.09	10.62	0.1	2.89	0.44	1.97	206650
2000	-25.85	0.07	15.88	0.24	2.71	0.56	2.48	230255
2001	-31.82	-0.06	10.18	-0.1	2.15	0.13	5.02	260021
2002	-62.5	-0.12	38.13	-0.08	2.35	0.41	6.46	269816
2003	-12.36	-0.11	87.96	-0.06	2.01	0.56	15.83	290457
2004	-30.9	0	70.25	-0.07	2.47	0.2	4.22	312866
2005	-34.15	0	179	-0.04	2.81	1.64	88.5	314129
2006	-34.8	0.25	128.46	0.32	3.16	1.73	38.15	287804
2007	-29.93	0.63	1227.19	0.55	5.83	80.13	14466.28	323245
2008	-68.73	0	1544.12	-0.3	5.27	74.39	21027.52	360251
2009	-22.21	0.47	2068.42	0.43	5.46	180.48	60557.23	375209
2010	-12.41	0.16	374.41	0.07	3.02	10.13	1168.73	431218
2011	-10.08	0	548.81	-0.14	2.72	20.3	3586.52	511173
2012	-25.57	0	1010.79	0.03	2.98	77.71	24573.11	565520
2013	-14.84	0.11	511.23	0.12	2.88	14.83	2303.41	564166
2014	-10.47	0.16	148.97	0.19	2.69	0.94	27.61	569823
2015	-10.22	0.47	986.07	0.39	5.1	17.85	3211.81	569475

2016	-10.1	0.03	10.19	0.03	3.18	-0.12	2.47	641319
2017	-28.92	0	10.2	-0.02	2.47	0.4	4.58	742801
2018	-27.88	0	25.08	-0.14	2.86	0	2.79	816883
2019	-12.5	0.09	13.33	0.17	3.08	0.21	2.1	468933
all	-68.73	0	2068.42	0.07	3.44	59.38	25014.97	9722311

三、模型构建

（一） 基准设定

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。

（二） 机器学习算法

本课题将机器学习的方法引入资产定价领域的原因有以下三点：

第一，机器学习算法能够优化传统计量经济学中函数形式假定过强的问题。传统计量经济学方法假设因子与股票收益率之间是简单的线性关系，然而现实市场中有些因子和收益率之间的关系并不是严格单调的，可能存在非线性的关系。机器学习算法(尤其是神经网络模型)并不需要人为的假设因子与股票收益之间的具体函数形式，而是基于真实的历史数据去拟合两者的关系，这种非参数估计的方法能够更好的用于描述因子与股票收益率之间的关系，从而提高预测的准确性。

第二，机器学习算法能够优化当因子过多或因子之间相关系数过高导致估计系数方差过

大的问题。在传统计量经济学方法下，当因子数量过多(例如因子数接近或等于样本数)或因子之间相关系数过高时，会导致模型的自由度下降，估计系数的方差将会上升。当估计系数方差太大时，基于该估计系数进行的样本外预测结果的方差也会上升，导致预测的结果不佳。在预测问题中的主要目的并不是解释过去，而是在有了新的数据后如何更好的预测未来，因此模型样本外的预测能力才是重要的。为了解决这种样本内预测结果较好，样本预测结果较差的问题(这种模型的泛化能力较差在机器学习中称之为过度拟合问题)，机器学习算法引入了惩罚项机制。通过加入惩罚项来压缩变量维度或收缩估计系数方差，可以使得模型的样本外预测结果得到改善。这也是为何机器学习的方法在预测问题上强于传统计量经济学的方法的原因之一。

第三，机器学习算法能够优化当被解释变量信息含量较低导致估计系数偏差的问题。现实中影响股票收益率的因素往往十分复杂，不同类型的股票同一时期的影响因子不一样，同一类型的股票不同时期的影响因子也不一样，这就导致了股票收益率的影响因子很多，噪音更多。传统计量经济学的方法无法很好的区分那些因子是有效因子，哪些因子是噪音。这样会导致传统计量经济的方法获得的估计系数可能产生偏差。为了这种应对信噪比低的情况，机器学习模型引入特征选择机制，通过剔除噪音，保留更少的有效因子来提升模型的估计准确性。

以下本文将阐述不同机器学习算法在金融学应用中的核心的区别，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。

1. 最小二乘回归

传统的计量经济学预测方法基于最小二乘回归(Ordinary Least Square，OLS)模型，它假设被解释变量和解释变量之间是线性关系，并通过最小化误差的平方和寻找数据的最优回归系数。当满足若干条件的情况下，OLS 估计量的样本内估计是具有无偏性和一致性的。以股票市场收益率预测问题为例，机器学习就算法是指基于历史经验的股票特征和预期收益率数据，构建不同的算法模型拟合股票特征和预期收益率之间的对应关系，使得通过股票特征进行收益率预测与实际预期收益率之间的样本外误差平方和尽可能小。

假设(2)式中的 g*(.)能够用线性关系来表示：

$$\mathrm{g}\left(z_{i,t;}\theta\right) = z'_{i,t}\theta \qquad (3)$$

我们需要通过在样本中最小化 L2 损失函数 L(θ) 来获得参数θ的最优取值

$$\mathrm{L}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t+1} - \mathrm{g}\left(z_{i,t;}\theta\right))^2 \qquad (4)$$

此外，由于金融学的股票收益率和预测变量经常呈现出厚尾分布的特征，方程(4)由于引入了平方项来定义损失函数，会导致样本的异常值会大大影响 OLS 估计的稳健性。因此我们考虑引入 Huber 损失函数来替代传统的 L2 损失函数，Huber 损失函数的定义如方程(5)所示(Huber, 1992)：

$$\mathcal{L}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H\left(r_{i,t+1} - \mathrm{g}\left(z_{i,t;}\theta\right), \gamma\right) \qquad (5)$$

其中

$$H(\mathrm{x},\gamma) = \begin{cases} x^2, & if \ \ |\mathrm{x}| \le \gamma \\ 2\gamma|\mathrm{x}| - \ \gamma^2 \ , & if \ \ |\mathrm{x}| > \gamma \end{cases}$$

Huber 损失函数通过引入超参数γ来调节损失函数的拟合情况，当模型常规值时，模型依然使用 L2 损失函数；而当模型拟合异常值时，改为使用更加稳健的 L1 损失函数。通过这种方法来解决面临异常值过多导致 OLS 估计结果不稳定的问题。

2. 带惩罚项的线性模型

然而，当面临高维数据、解释变量之间存在相关性、样本之间并不是独立同分布等问题时，传统计量经济学在预测问题上面的表现就显得更为差强人意了。为了解决上述问题，各种改良 OLS 的机器学习算法产生了。带惩罚项的线性模型就通过加入惩罚项来降低变量过多导致方程(4)出现过拟合问题的第一类解决思路。方程(6)表示了带惩罚项的线性模型的损失函数，在原方程(5)中损失函数上额外添加的$\phi(\theta;\cdot)$为惩罚项，其中θ为需要调节的超参数。如方程(7)所示，不同的惩罚项设定代表了不同的机器学习算法，当惩罚项为 l1 范数时，该算法为套索回归(LASSO)；当惩罚项为 l2 范数时，该算法为岭回归(Ridge)；当惩罚项为 l1 和 l2 范数综合时，该算法为弹性网络模型(Enet)。弹性网络模型综合了 LASSO 和 Ridge 回归，因此本文以 Enet 模型作为带惩罚项线性模型的代表。

$$\mathcal{L}(\theta;\cdot) = \underbrace{\mathcal{L}(\theta)}_{\text{Loss Function}} + \underbrace{\phi(\theta;\cdot)}_{\text{Penalty}} \qquad (6)$$

$$\phi(\theta;\cdot) = \begin{cases} \frac{1}{2}\lambda\sum_{j=1}^{P}\theta_j^2, \ \text{Ridge;} \\ \lambda\sum_{j=1}^{P}|\theta_j|, \ \text{Lasso;} \\ \lambda(1-\rho)\sum_{j=1}^{P}\left|\theta_j\right| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2, \ \text{Elastic Net;} \end{cases} \qquad (7)$$

3. 主成分分析回归

主成分分析(PCA)是金融学中常用的降维方法，原理是设法将原来变量重新组合成一组新的相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的总和变量尽可能

多地反映原来变量的信息的统计方法。可以先基于PCA的方法对预测因子合集$z_{i,t}$进行降维，然后再构建OLS方法对方程4进行估计，这种方法能够通过减低数据维度来提高模型样本外估计的准确性。第j个PCA的权重由以下方程(8)给出：

$$\mathrm{w_j} = argmax_w \ Var(zw),\ s.t. w'w = 1, w'z'zw_l = 0, l = 1,2,\dots,j-1 \quad (8)$$

当K=1时，PCR回归会使用最能代表被解释变量的主成分信息来估计模型，而当PCA保留的K个权重等于原维度时，PCR回归会保留所有被解释变量的信息，此时模型的估计与OLS的估计结果一致。PCR模型保留成分的数量K的选择需要基于验证集来调整。

4. 最小偏二乘回归

PCA 的方法问题在于会保留方差最大的信息，而金融数据的信噪比太低，保留方差最大信息的同时也会同时保留很多噪音信息。第j个PLS的权重由以下方程(9)给出：

$$\mathrm{w_j} = argmax_w \ corr^2(Y, zw)Var(zw),\ s.t. w'w = 1, w'z'zw_l = 0, l = 1,2,\dots,j-1 \quad (9)$$

由于加入了$corr^2(Y, zw)$项，PLS方法会在压缩信息的同时考虑压缩信息与被解释变量之间的相关性，以保证在压缩信息时更多的保留那些与被解释变量相关的信息(Kelly & Pruitt, 2015)。

5. 树回归模型

以上机器学习模型依然无法包含解释变量之间的交互效应，如果在上面模型中直接加入解释变量之间的交互项的话，随着解释变量数量上升时，整个模型估计的复杂度会呈现指数级别的上升，不利于模型估计。回归树(Regression Trees)模型目前是机器学习算法中比较常用的捕获变量交互效应的非线性模型。回归树实际上是将空间用超平面进行划分的一种方法，每次分割的时候，都将当前的空间一分为二， 这样使得每一个叶子节点都是在空间中的一个不相交的区域，在进行决策的时候，会根据输入样本每一维 feature 的值，一步一步往下，最后使得样本落入N个区域中的一个。单个回归树模型往往容易出现过拟合的问题，此时衍生出了两种方式来改善：集成学习(Ensemble)和提升树(Boosting)。

随机森林实际上是一种特殊的集成学习方法，它将决策树用作基础中的模型。首先，用bootstrap 方法生成 m 个训练集，然后，对于每个训练集，构造一颗决策树，在节点找特征进行分裂的时候，并不是对所有特征找到能使得指标(如信息增益)最大的，而是在特征中随机抽取一部分特征，在抽到的特征中间找到最优解，应用于节点，进行分裂。通过随机减少样本和特征的方法来避免过拟合的问题。

梯度提升树(Gradient Boosting Regression Tree，GBDT) 是一种迭代的决策树算法，由多棵决策树组成，综合所有树的预测作为最终预测。该算法的核心在于每棵树学习之前所有树

的残差；而为了消除残差，模型在残差减少的梯度(Gradient)方向上建立一个新的模型。

6. 神经网络模型

最后本课题将使用神经网络模型拟合 g*(.)，神经网络模型是目前最强大的非线性机器学习算法。本课题仅测试最简单的全连接前馈神经网络模型(feed-forward networks)，这种神经网络结构并不复杂，便于数据拟合和计算，但是也能拟合任何函数形态。文章总共涉及五个神经网络模型，每个神经网络模型的大体结构一致，仅仅只是隐藏层和神经元的数量不同，五个神经网络模型分别部署了 1 到 5 个隐藏层，其他的模型构建情况基本一致。具体神经网络模型的设置如下表 3 所示。

表 3：神经网络模型的具体结构

参数	NN1	NN2	NN3	NN4	NN5
隐藏层数	1	2	3	4	5
每层神经元数	32	32+16	32+16+8	32+16+8+4	32+16+8+4+2
连接情况	Fully Connected				
激活函数	ReLU				
优化算法	Adaptive moment estimation algorithm(Adam)				
停止条件	Terminate the optimization by validation sample loss				
正则项	L1 penalization of weight parameters				
批标准化	Add batch normalization for each hidden unit				
集合方式	Random initialize weight and construct predictions by averageing				

（三） 模型评价指标

本文参考 Gu et al. (2019a)采用以下方程(10)来评价模型的表现。

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3} (r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}} \quad (10)$$

其中：$\mathcal{T}_3$代表样本外测试集模型，$\hat{r}_{i,t+1}$代表模型的预测值，$r_{i,t+1}$代表真实值。需要强调的是这个计算公式的样本外 R 方与传统的 R 方不同，分母并没有做去均值处理(即没有减去收益率的均值)。原因在于：股票历史的均值这一数据包含很多噪音，使用该方法会引入额外的无效信息损失，如果将样本的历史均值作为个股的预测收益率的基准，其效果可能还不如直接使用 0 作为个股的预测收益率。(本文也采用了传统的 R 方来度量模型的评价指标作为稳健性检验，实证结果并不会发生变化)

（四） 模型变量重要性程度计算

在训练好的模型中，保持模型的参数不变，将某个预测因子的值全部替换为 0，然后模

型其他输入变量不变，记录去掉某个预测因子后模型预测 R 方的减小量。对所有变量执行以上操作，并依照 R 方减小量进行标准化处理，得到不同模型每个特征的重要性水平。

（五） 超参数校准

在机器学习算法中一般都会有超参数需要人为决定，它决定了每个模型的复杂性，并且是模型建造者防止过度拟合、提升模型样本外表现的首要手段。常见的超参数有随机森林中树的个数、深度，LASSO 算法中的惩罚参数等。具体而言本文所有模型需要调节的超参数和取值范围如表 4 所示。

表 4：所有模型需要调节的超参数和取值范围

OLS3+H	PLS	PCR	Enet+H	RF	GBRT+H	NN1
Huber loss 99.9%	# of Component 1 - 50	# of Component 1 - 50	Huber loss 99.9% ρ = 0.5 $\lambda \in (10^{-4}, 10^{-1})$	Depth= 1 – 6 # Trees= 300 # Features in each split 3-50	Huber loss 99.9% Depth= 1 – 6 # Trees= 300 Learning Rate 0.01 – 0.1	L1 penalty $10^{-5} - 10^{-3}$ Learning Rate 0.001 – 0.01 Batch Size=10000 Epochs=100 Patience=5 Ensemble=10

本文将所有数据按照时间将样本集划分为三个部分：训练集、验证集和测试集。首先在训练集中拟合数据，再用第二部分验证样本中通过计算目标函数判断误差对模型进行超参数调整。第三部分样本将作为测试集来评估所得模型的样本外预测准确性。具体而言，文章把 22 年(1997 到 2019)的样本拆分为 3 部分：前 6 年为训练集(1997 - 2003)、中 6 年为验证集(2004 - 2009)、后 10 年为样本外预测集(2010 - 2019)。此外，为了尽可能的接近真实样本外投资过程，保留数据集为时间序列的特征，数据集的训练和验证过程没有采用机器学习中使用的交叉验证(cross-validation)方法，而是使用了更为复杂的混合递归的方法，这种方法的具体做法如下：

第 1 次，在 1997 - 2003 年的样本中拟合，在 2004 - 2009 年的样本中根据损失函数确定超参数，在 2010 年的样本中预测，保留预测结果。

第 2 次，在 1997 - 2004 年的样本中拟合，在 2005 - 2010 年的样本中根据损失函数确定超参数，在 2011 年的样本中预测，保留预测结果。

...

第 10 次，在 1997 - 2012 年的样本中拟合，在 2013 - 2018 年的样本中根据损失函数确定超参数，在 2019 年的样本中预测，保留预测结果。

四、实证结果

（一） 个股的可预测性实证结果

表 5 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 OLSH 代表基于 OLS + Huber Loss 方程(5)使用所有变量进行拟合的结果，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[⑤]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 5：R 方度量下不同机器学习模型样本外预测准确度

	OLSH	OLS3H	PLS	PCR	Enet	RF	GBRT	NN1	NN2	NN3	NN4	NN5
ALL	-0.411	-0.178	0.194	0.152	0.234	0.083	0.338	0.361	0.633	0.283	0.215	0.045
Top 300	-0.626	-0.121	0.175	0.545	0.432	0.529	0.566	0.065	0.44	0.354	0.435	0.189
Bottom 300	-0.204	0.07	0.75	0.425	0.608	0.424	0.359	0.717	0.928	0.51	0.412	-0.22

注：样本外测试时间：2010 年到 2019 年 8 月

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。OLSH 模型的全样本 R 方仅为-0.411%，OLS3H 方法模型的样本外 R 方仅为-0.178%。这表明了：(1)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。(2)传统的 OLS 模型使用全部变量的预测结果还不如仅仅使用三个特征的预测结果。

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 0.194%、0.152%和 0.234%。 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 0.083%和 0.338%，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 0.361%、0.633%、0.283%、0.215%和 0.045%。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

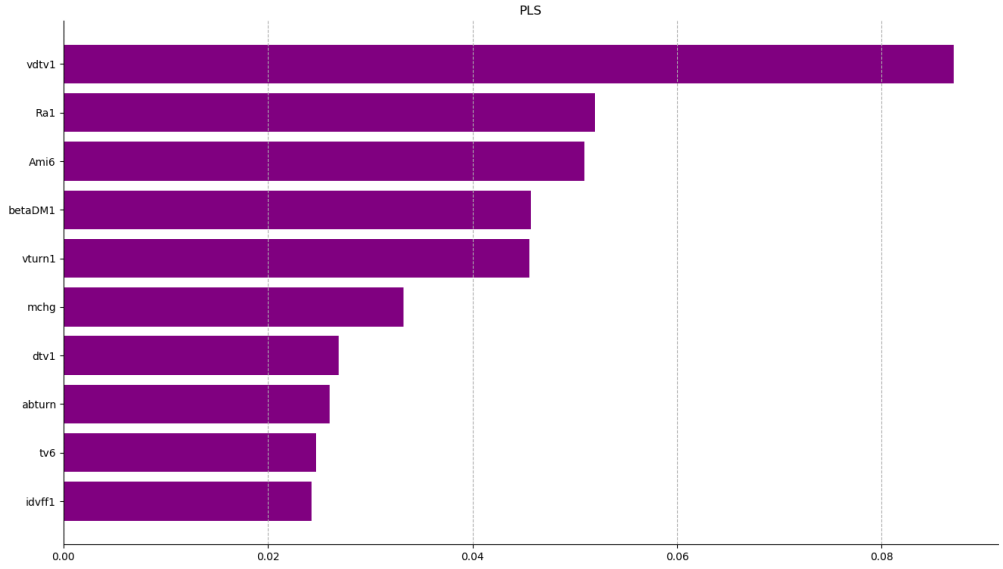
神经网络模型的结果最好为 0.663%，而 5 层神经网络模型的结果却 0.045%。Top (Bottom)300 是指最大(小)的 300 只股票预测结果，最好的模型为 GBRT(NN2)，样本外 R 方为 0.566%(0.928%)。本文的预测结果与美国文献类似，对比 Gu et al. (2019)基于美国机器学习的预测结果，其表现最好的随机森林的样本外 R 方为 0.33%。

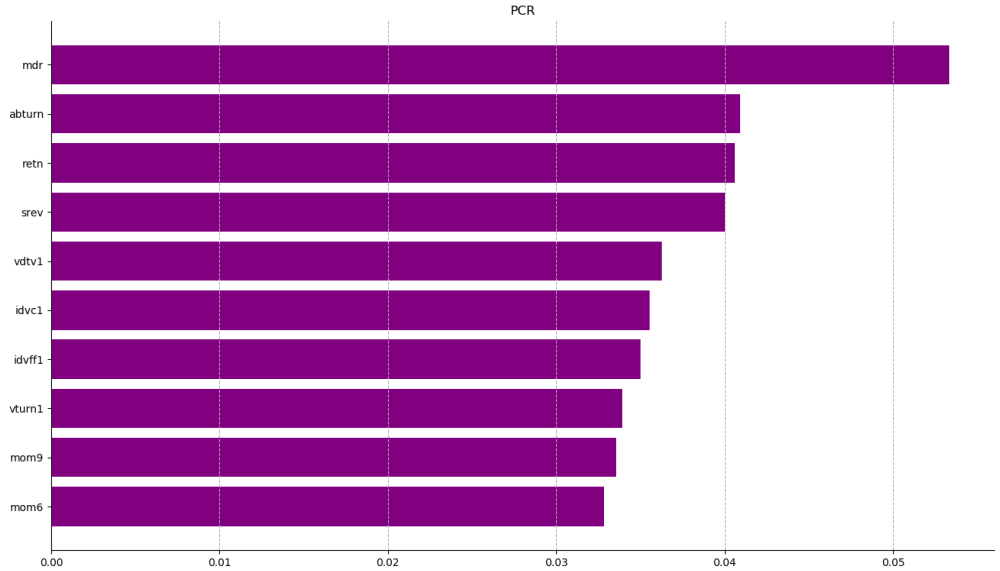
（二） 预测因子的重要性

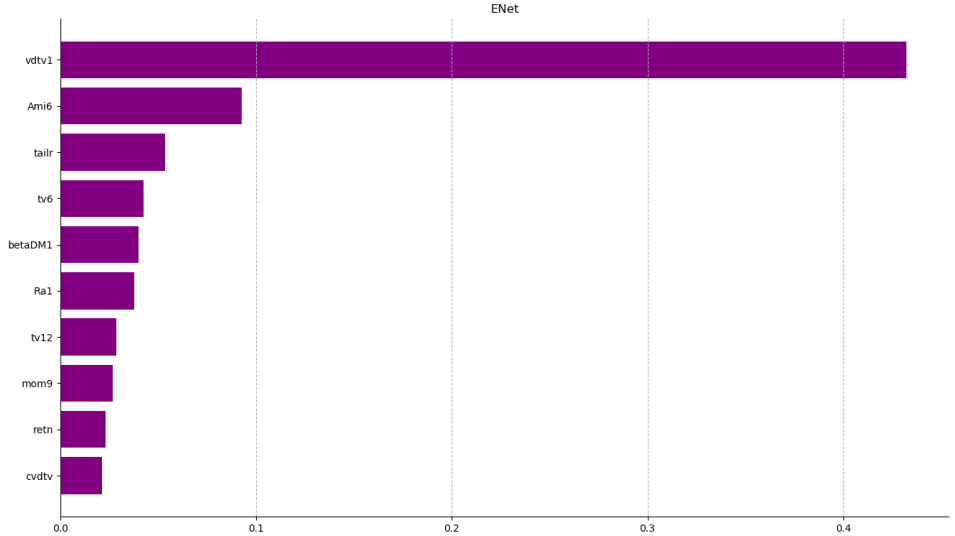
表 6、图 2 分别展示了不同机器学习模型算法不同股票预测特征的重要性排序。所有机器学习模型的不同变量的重要程度排序都十分类似，其中流动性的指标 vdtv1 成交量的方差[⑥]、vturn 换手率的方差变量、Ami6 等三个流动性指标的重要性排名靠前，平均重要性为 0.132、0.039、0.031。重要性程度前 20 的特征中，流动性特征有 5 个，动量特征有 5 个，波动率(风险)特征有 10 个。

表 6：不同变量的重要性程度(前 20 个因子)

	PLS	PCR	ENet	RF	GBRT	mean
vdtv1	0.087	0.036	0.432	0.034	0.069	0.132
vturn1	0.046	0.034	0.007	0.048	0.061	0.039
Ra1	0.052	0.000	0.038	0.027	0.071	0.038
Ami6	0.051	0.001	0.093	0.007	0.002	0.031
abturn	0.026	0.041	0.016	0.033	0.037	0.030
mom11	0.011	0.030	0.019	0.031	0.039	0.026
tailr	0.020	0.002	0.053	0.015	0.038	0.026
mom9	0.010	0.034	0.027	0.024	0.022	0.023
idvff1	0.024	0.035	0.006	0.024	0.026	0.023
turn1	0.017	0.022	0.001	0.039	0.036	0.023
tv6	0.025	0.005	0.042	0.019	0.020	0.022
mdr	0.017	0.053	0.000	0.019	0.012	0.020
betaDM1	0.046	0.003	0.040	0.005	0.006	0.020
retn	0.010	0.041	0.023	0.010	0.016	0.020
cvdtv	0.019	0.030	0.021	0.018	0.010	0.020
Ra25	0.001	0.007	0.002	0.029	0.055	0.019
lrev	0.001	0.006	0.010	0.030	0.044	0.018
idvc1	0.015	0.036	0.000	0.018	0.023	0.018
mchg	0.033	0.011	0.008	0.016	0.023	0.018


PLS
vdtv1
Ra1
Ami6
betaDM1
vturn1
mchg
dtv1
abturn
tv6
idvff1
0.00
0.02
0.04
0.06
0.08


PCR
mdr
abturn
retn
srev
vdtv1
idvc1
idvff1
vturn1
mom9
mom6
0.00
0.01
0.02
0.03
0.04
0.05


ENet
vdtv1
Ami6
tailr
tv6
betaDM1
Ra1
tv12
mom9
retn
cvdtv
0.0
0.1
0.2
0.3
0.4

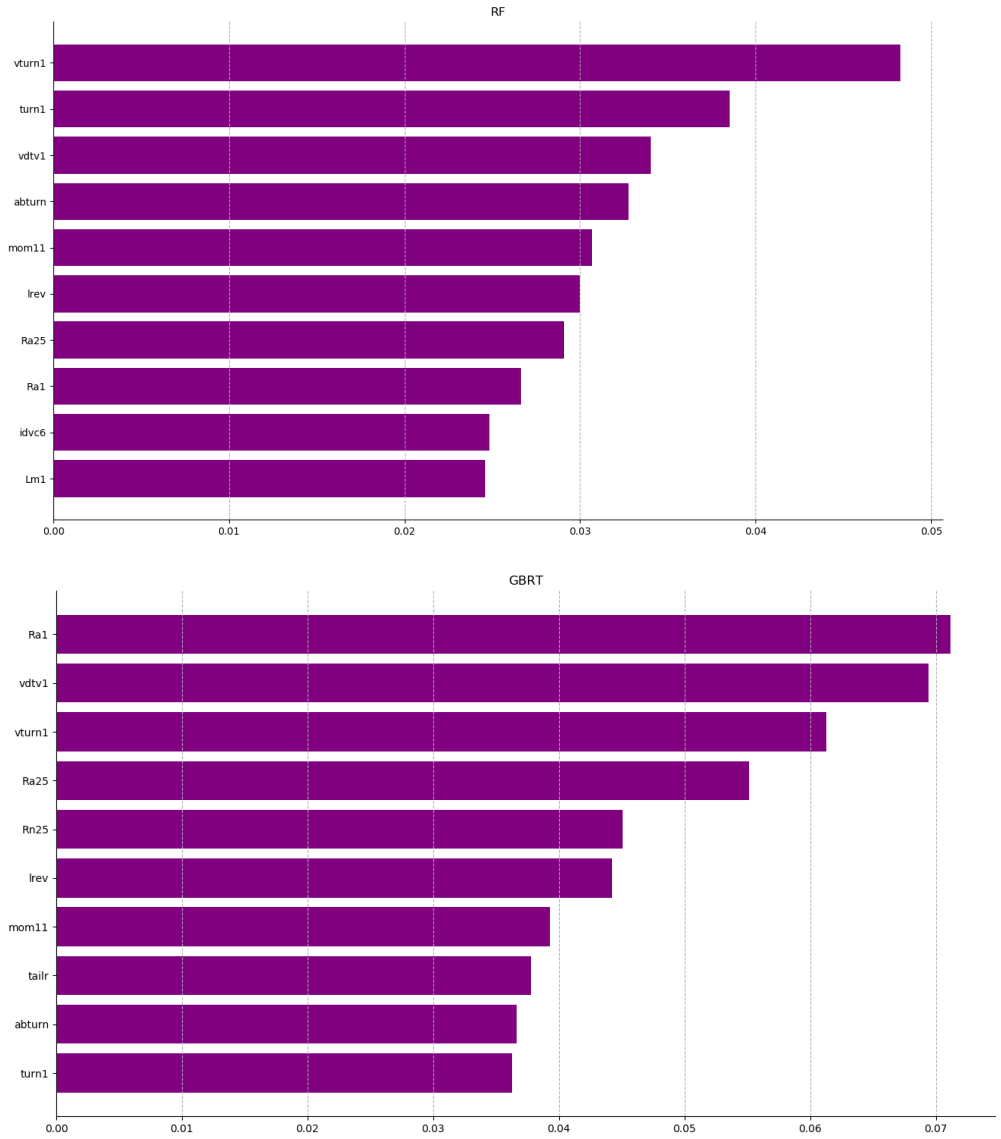


图 2：不同机器学习模型算法不同股票预测特征的重要性排序

尽管本文已经剔除了中国市值最小的 30%股票，但是在中国市场上，流动性指标还是表现除了最强的预测能力，可能的原因有两点：首先，流动性因子的溢价来源于流动性低的股票会比流动性高的股票更难达成交易，当面临股市衰退时，持有流动性低的股票资产面临无法处置资产的损失会远远大于持有流动性高的股票。非流动性溢价就是为了弥补这种低流动资产的风险从而带来的风险溢价。其次，由于中国股票市场制度依然处于不断完善的阶段，涨跌停板和公司随意停牌制度进一步加大了中国流动性低股票的溢价程度。正是由于这些市场特征，导致的中国非流动股票特征对于下一期的股票收益率会产生更大的预测能力。动量类指标的预测能力最弱，这也与中国 A 股市场动量异象性因子不显著的研究发现十分类似(鲁臻和邹恒甫, 2007)

（三） 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 8 月。表 7 显示了不同机器学习算法市值加权构建资产组合策略的结果。其中 OLS3 是指基于 OLS + Huber Loss 预测方法得到的资产组合，Hi_10 是指纯多头资产组合策略平均能获得 0.369%的月度收益，月度标准差为 7.151%，年化夏普比率为 0.179。Lo_10 是指纯空头资产组合策略平均能获得 0.644%的月度收益，月度标准差为 9.504%，年化夏普比率为 0.235。H-L 是指多空资产组合策略平均能获得 1.013%的月度收益，月度标准差为 6.871%，年化夏普比率为 0.511。

表 7：不同机器学习模型等权加权构建资产组合的绩效表现

Panel A. 市值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	-0.644	0.141	0.238	0.173	0.353	0.064	0.521	0.698	0.392	0.369	1.013
PCR	-0.591	-0.139	0.087	0.114	0.470	0.303	0.350	0.445	0.433	0.680	1.271
PLS	-0.737	-0.098	0.239	0.628	0.413	0.247	0.556	0.591	0.886	0.734	1.471
ENet	-0.878	-0.074	0.186	-0.153	0.289	0.341	0.383	0.687	0.918	0.969	1.847
RF	-1.263	-0.167	0.070	0.442	0.356	0.270	0.184	0.688	0.746	0.366	1.629
GBRT	-1.576	-0.135	0.213	0.298	0.086	0.151	0.521	0.612	0.866	1.019	2.595
NN1	-1.403	-0.298	-0.152	0.524	0.434	0.755	0.815	0.917	1.153	0.834	2.237
NN2	-1.611	-0.372	0.194	0.476	0.711	0.268	0.519	0.326	0.866	0.849	2.459
NN3	-1.121	-0.481	0.205	0.136	0.447	0.186	0.527	0.744	0.644	0.718	1.839
NN4	-1.397	-0.383	-0.383	0.258	0.882	0.402	0.951	0.532	0.650	0.854	2.251
NN5	-0.712	-0.167	0.000	0.133	0.198	0.249	0.430	0.543	0.764	0.816	1.528
Panel B. 市值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	9.504	8.859	8.169	7.582	7.129	6.811	6.652	6.956	6.900	7.151	6.871
PCR	9.141	8.723	7.719	7.577	6.922	6.745	6.789	6.510	6.286	6.665	6.021
PLS	8.622	7.741	7.671	7.774	7.742	7.658	7.526	7.617	7.300	6.905	6.995
ENet	8.956	8.152	7.321	6.989	6.886	6.950	6.605	6.842	7.204	7.107	5.922
RF	9.780	8.560	7.233	7.135	7.146	6.676	7.073	7.286	7.251	7.383	5.740
GBRT	9.583	8.936	8.019	7.395	6.809	6.596	6.929	6.902	7.262	7.506	5.246
NN1	8.732	7.793	7.417	7.221	7.197	7.740	8.017	7.839	8.527	7.859	6.124
NN2	9.256	8.006	7.868	7.161	6.804	6.827	6.688	7.365	8.130	8.309	5.973
NN3	9.579	8.622	8.068	6.680	7.170	6.637	6.891	7.777	7.928	8.010	6.650
NN4	8.378	7.809	7.312	7.290	7.942	7.221	7.707	7.301	7.662	8.193	5.144

NN5	8.797	8.041	8.283	8.008	7.795	7.517	7.505	7.483	7.643	7.466	6.699
Panel C. 市值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	-0.235	0.055	0.101	0.079	0.171	0.033	0.271	0.348	0.197	0.179	0.511
PCR	-0.224	-0.055	0.039	0.052	0.235	0.156	0.179	0.237	0.239	0.353	0.731
PLS	-0.296	-0.044	0.108	0.280	0.185	0.112	0.256	0.269	0.421	0.368	0.729
ENet	-0.340	-0.031	0.088	-0.076	0.145	0.170	0.201	0.348	0.442	0.472	1.081
RF	-0.448	-0.068	0.034	0.215	0.173	0.140	0.090	0.327	0.357	0.172	0.983
GBRT	-0.570	-0.052	0.092	0.139	0.044	0.079	0.261	0.307	0.413	0.470	1.714
NN1	-0.556	-0.133	-0.071	0.251	0.209	0.338	0.352	0.405	0.468	0.368	1.265
NN2	-0.603	-0.161	0.085	0.230	0.362	0.136	0.269	0.153	0.369	0.354	1.426
NN3	-0.405	-0.193	0.088	0.071	0.216	0.097	0.265	0.332	0.281	0.311	0.958
NN4	-0.578	-0.170	-0.182	0.122	0.385	0.193	0.427	0.253	0.294	0.361	1.516
NN5	-0.281	-0.072	0.000	0.058	0.088	0.115	0.198	0.252	0.346	0.378	0.790

表 7 中可以发现以下规律：(1)比较不同模型的纯多策略的绩效表现的话，2 层神经网络选股策略的结果最好，平均能获得 0.849%的月度收益，月度标准差为 8.309%，年化夏普比率为 0.354；(2)比较不同模型的多空策略的绩效表现的话，两层神经网络选股策略的结果最好，平均能获得 2.459%的月度收益，月度标准差为 5.973%，年化夏普比率为 1.426；(3)神经网络模型优于树类模型，树类模型优于线性机器学习模型，机器学习模型都比 OLS 结果要好；(4)神经网络模型并不是层数越多越好，意味着模型并不是越复杂越好。

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，表 7 所有的规律在表 8 依然存在，稍有不同点在于同一模型下等权重加权的资产组合结果会优于市值加权构建资产组合，这也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，纯多头资产组合策略平均能获得 1.204%的月度收益，月度标准差为 8.748%，年化夏普比率为 0.477。多空资产组合策略平均能获得 2.975%的月度收益，月度标准差为 4.276%，年化夏普比率为 2.410。

表 8：不同机器学习模型市值等权构建资产组合的绩效表现

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	-0.753	0.008	0.178	0.427	0.455	0.501	0.631	0.755	0.723	0.548	1.301
PCR	-0.959	-0.432	-0.053	0.240	0.524	0.558	0.765	0.824	0.915	1.091	2.051
PLS	-1.330	-0.396	-0.048	0.393	0.418	0.526	0.757	0.891	1.062	1.200	2.530
ENet	-1.260	-0.360	-0.107	0.216	0.354	0.678	0.673	1.003	1.113	1.162	2.422
RF	-1.280	-0.271	0.206	0.491	0.467	0.584	0.639	0.797	0.964	0.875	2.156
GBRT	-1.467	-0.143	0.402	0.228	0.387	0.570	0.784	0.857	0.898	0.961	2.428

NN1	-1.525	-0.396	-0.009	0.400	0.476	0.720	0.792	0.850	1.115	1.057	2.582
NN2	-1.771	-0.395	0.144	0.402	0.676	0.652	0.776	0.725	1.034	1.204	2.975
NN3	-1.302	-0.394	0.143	0.186	0.411	0.536	0.777	0.951	1.049	1.115	2.417
NN4	-1.654	-0.413	-0.196	0.467	0.567	0.660	0.845	0.916	1.006	1.275	2.928
NN5	-1.003	-0.196	0.003	0.301	0.485	0.593	0.609	0.597	0.954	1.128	2.130
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	9.331	8.906	8.685	8.535	8.367	8.256	8.133	8.240	8.294	8.431	4.705
PCR	9.360	9.124	8.835	8.711	8.373	8.328	8.202	8.059	7.933	7.929	4.452
PLS	9.245	8.752	8.689	8.676	8.419	8.362	8.310	8.269	8.074	7.970	4.279
ENet	9.408	8.950	8.661	8.522	8.390	8.322	8.209	8.271	8.299	7.878	4.527
RF	9.856	9.021	8.508	8.395	8.116	8.084	8.165	8.091	8.069	8.449	3.866
GBRT	9.872	9.116	8.740	8.280	8.075	8.039	7.912	8.043	8.254	8.536	3.755
NN1	9.021	8.713	8.514	8.343	8.019	8.281	8.489	8.395	8.702	8.678	4.712
NN2	9.375	8.919	8.658	8.374	8.200	8.187	8.113	8.109	8.344	8.748	4.276
NN3	9.337	9.095	8.751	8.254	8.111	8.149	8.226	8.221	8.500	8.655	4.913
NN4	8.968	8.833	8.541	8.516	8.501	8.261	8.152	8.019	8.521	8.730	3.879
NN5	9.219	9.050	8.934	8.776	8.546	8.215	7.918	8.035	8.131	8.318	5.060
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	-0.280	0.003	0.071	0.173	0.189	0.210	0.269	0.317	0.302	0.225	0.958
PCR	-0.355	-0.164	-0.021	0.095	0.217	0.232	0.323	0.354	0.399	0.477	1.595
PLS	-0.499	-0.157	-0.019	0.157	0.172	0.218	0.316	0.373	0.455	0.521	2.048
ENet	-0.464	-0.139	-0.043	0.088	0.146	0.282	0.284	0.420	0.465	0.511	1.853
RF	-0.450	-0.104	0.084	0.203	0.199	0.250	0.271	0.341	0.414	0.359	1.932
GBRT	-0.515	-0.054	0.159	0.095	0.166	0.245	0.343	0.369	0.377	0.390	2.240
NN1	-0.586	-0.157	-0.004	0.166	0.206	0.301	0.323	0.351	0.444	0.422	1.898
NN2	-0.654	-0.153	0.058	0.166	0.286	0.276	0.332	0.310	0.429	0.477	2.410
NN3	-0.483	-0.150	0.056	0.078	0.176	0.228	0.327	0.401	0.427	0.446	1.704
NN4	-0.639	-0.162	-0.080	0.190	0.231	0.277	0.359	0.396	0.409	0.506	2.615
NN5	-0.377	-0.075	0.001	0.119	0.197	0.250	0.267	0.257	0.406	0.470	1.458

表 9 展示了不同机器学习模型构建资产组合的风险调整后绩效表现(样本外测试时间：2010 年到 2019 年 8 月)，等权(市值)加权的机器学习资产组合 FF3 因子调整后的月度 Alpha 为 3.602(3.349)，模型 R 方仅有 16.607%(16.158%)，说明 FF3 因子对于本文机器学习选股的资产组合解释力较低，FF5 因子调整后的月度 Alpha 为 3.602(3.349)，模型 R 方仅有 26.676%(29.322%)，说明 FF5 因子对于本文机器学习选股的资产组合解释力较低，所有资产组合 Alpha 的 T 值都超过了 5，说明均在统计上显著。

表 9：不同机器学习模型构建资产组合的风险调整后绩效表现

	OLS3	PLS	PCR	Enet	RF	GBRT	NN1	**NN2**	NN3	NN4	NN5
Panel A1.	市值加权机器学习资产组合 FF3 因子模型调整后收益										
Mean_Ret	1.01	1.47	1.27	1.85	1.63	2.60	2.24	2.46	1.84	2.25	1.53
Alpha	0.78	1.33	1.14	1.59	1.71	2.93	2.30	2.57	1.72	2.22	1.62
Alpha_t	(1.401)	(2.491)	(2.536)	(4.141)	(3.472)	(6.472)	(4.126)	(4.814)	(3.481)	(5.128)	(2.771)
R_2	29.08	16.60	33.30	41.04	15.09	16.01	1.52	0.72	16.51	0.86	1.02
Panel A2.	市值加权机器学习资产组合 FF5 因子模型调整后收益										
Alpha	0.73	1.35	1.01	1.42	1.59	2.81	2.05	2.33	1.60	2.07	1.38
Alpha_t	(1.386)	(2.683)	(2.389)	(3.74)	(3.339)	(5.991)	(4.084)	(4.941)	(3.196)	(5.085)	(2.442)
R_2	70.92	31.78	32.30	33.52	45.93	36.73	2.97	5.27	16.51	5.97	4.05
Panel B1.	等权加权机器学习资产组合 FF3 因子模型调整后收益										
Mean_Ret	1.30	2.53	2.05	2.42	2.16	2.43	2.58	2.98	2.42	2.93	2.13
Alpha	1.17	2.49	1.98	2.30	2.15	2.48	2.64	2.97	2.31	2.94	2.33
Alpha_t	(3.004)	(7.597)	(5.766)	(6.442)	(6.975)	(7.588)	(6.624)	(8.261)	(5.621)	(8.615)	(5.497)
R_2	17.84	17.68	25.09	27.35	19.67	17.21	0.35	0.72	8.67	0.36	0.80
Panel B2.	等权加权机器学习资产组合 FF5 因子模型调整后收益										
Alpha	1.06	2.42	1.84	2.13	2.04	2.37	2.44	2.81	2.17	2.77	2.08
Alpha_t	(2.714)	(7.413)	(5.492)	(6.014)	(6.592)	(7.065)	(6.727)	(8.552)	(5.379)	(8.688)	(5.356)
R_2	18.30	17.49	27.92	31.23	21.34	19.27	5.22	3.97	9.98	4.02	7.83

注：Alpha_t 为经过 White (1980)异方差调整过后的 T

图 4-7 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(对数)，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 8 年累计收益率(对数)约为 1.6(1.45)，而同期沪深 300 收益率仅为-0.05。

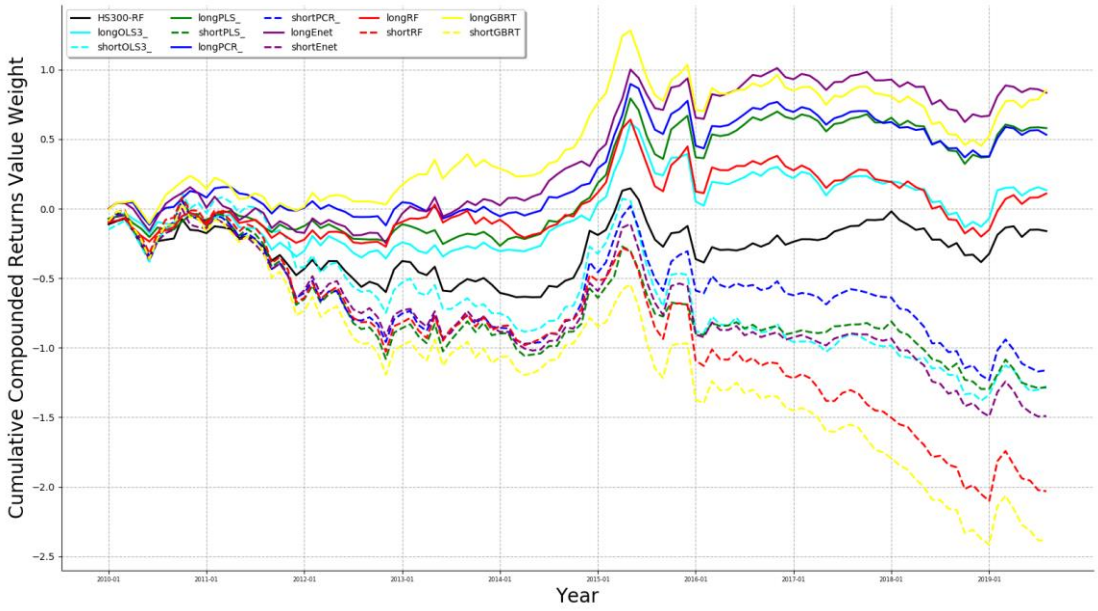


图 4：1997 年 1 月到 2017 年 12 月样本外机器学习资产组合策略累计收益率(市值加权)

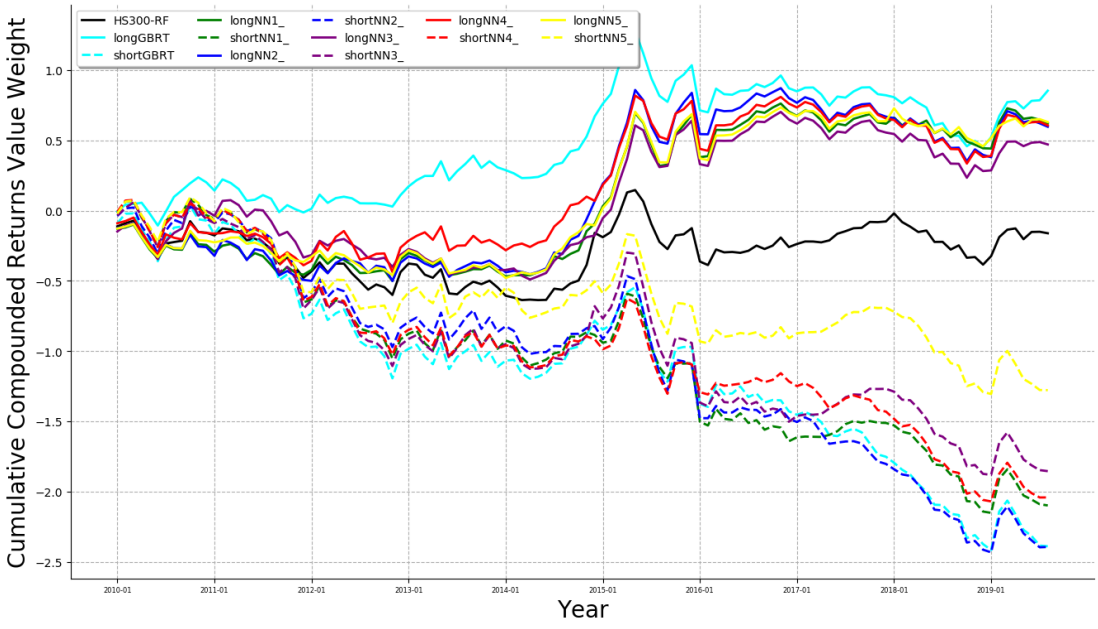


图 5：1997 年 1 月到 2017 年 12 月样本外机器学习资产组合策略累计收益率(市值加权)

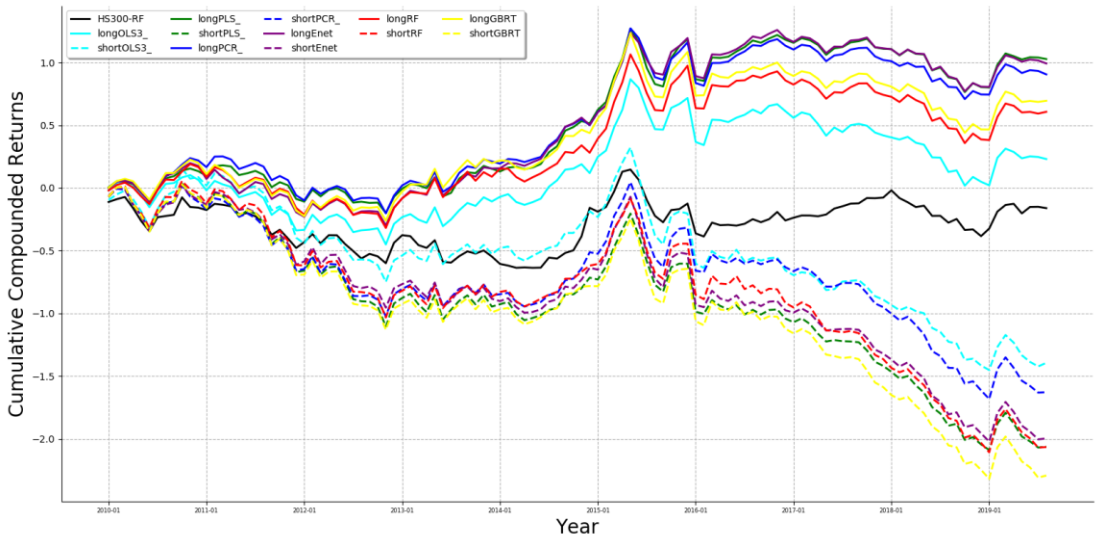


图 6：1997 年 1 月到 2017 年 12 月样本外机器学习资产组合策略累计收益率(等权加权)

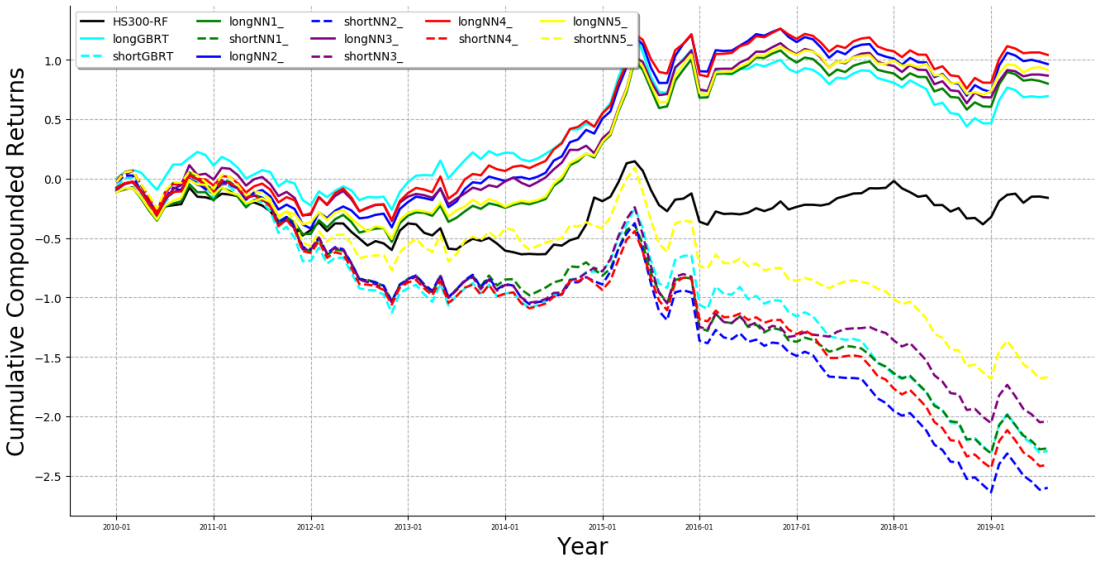


图 7：1997 年 1 月到 2017 年 12 月样本外机器学习资产组合策略累计收益率(等权加权)

五、研究结论

本文将机器技术引入中国股票市场的个股收益率可预测问题，并在探索不同类别的交易性异象特征对于中国股票市场资产价格可预测性的贡献程度。本文发现：(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-0.178%，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 0.633%；(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 8 月期间，平均能获得 2.975%(2.459%)的月度收益，月度波动率为 4.276%(5.973%)，年化夏普比率为 2.41(1.426)，经过 FF5 因子调整后的依然能获得显著的月度 Alpha 值为 2.81(2.33)。(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**，vdtv1 成交量的方差、vturn 换手率的方差变量、Ami6 三个流动性指标的重要性排名靠前。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。

注释

①截止至 2019 年 9 月底，中国沪深两市股票市场总市值已达 54 万亿(数据来源 Wind)，位居全球股票市场第二，仅次于美国。

②股票异象性(Anomaly)是指按某种股票特征(例如：规模、价值、盈利性等)分组进行排序，构建出股票特征值最高与最低组的多空资产组合，获得的资产组合收益率不能被基准因子模型(CAPM 或 FF5)所解释，能产生有显著的 alpha 的现象。这些股票异象性特征也被认为是对未来一期股票收益率具有预测能力的变量，基于异象性特征排序构建的资产组合也被称为异象性因子(Novy-Marx & Velikov, 2015)。

③具体见：http://www.gov.cn/xinwen/2019-08/23/content_5423691.htm。

④这个条件与原文稍有不同，原文为剔除当月交易天数少于 15 天的样本，但是中国由于过年假期较长，如果根据原文的做法，会导致 4 个月的样本完全被删除。

⑤本文在流动性、波动率和动量类异象性特征中分别选了一个最显著的因子来代表该类异象性因子。

⑥本文对所有股票异象特征因子进行了单因子排序分析(single portfolio analysis)，单因子分组中 T 值最大的因子就是 vdtv1 成交量的方差，最高组减最低组的对冲资产组合平均月度收益率为-1.52%，t 值为-3.12。由于本文的研究重点并不是做中国异象性因子分析，所有这部分结果并没有展示，欢迎读者来信索取。

参考文献

(1) 陈卫华、徐国祥:《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》，2018年第01期。

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

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

(4) 李斌、林彦、唐闻轩:《ML-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》，2017年第05期。

(5) 鲁臻、邹恒甫:《中国股市的惯性与反转效应研究》，《经济研究》，2007年第09期。

(6) 马晓君、沙靖岚、牛雪琪:《基于LightGBM算法的P2P项目信用评级模型的设计及应用》，《数量经济技术经济研究》，2018年第05期。

(7) 苏治、卢曼、李德轩:《深度学习的金融实证应用:动态、贡献与展望》，《金融研究》，

2017年第05期。

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

(9) Amihud Y., Hameed A., Kang W., Zhang H., 2015, "The Illiquidity Premium: International Evidence", ***Journal of Financial Economics***, Vol. 117, pp.350~368.

(10) Ang A. & Bekaert G., 2007, "Stock Return Predictability: Is It there?", ***The Review of Financial Studies***, Vol. 20, pp.651~707.

(11) Ang A., Chen J., Xing Y., 2006, "Downside Risk", ***The Review of Financial Studies***, Vol. 19, pp.1191~1239.

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

(13) Asness C. S., Moskowitz T. J., Pedersen L. H., 2013, "Value and Momentum Everywhere", ***The Journal of Finance***, Vol. 68, pp.929~985.

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

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

(16) Boyer B., Mitton T., Vorkink K., 2010, "Expected Idiosyncratic Skewness", ***The Review of Financial Studies***, Vol. 23, pp.169~202.

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

(18) Campbell J. Y. & Cochrane J. H., 1999, "By Force of Habit: A Consumption Based Explanation of Aggregate Stock Market Behavior", Journal of Political Economy, Vol. 107, pp.205~251.

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

(20) Chinco A., Clark Joseph A. D., Ye M., 2018, "Sparse Signals in the Cross‐Section of Returns", ***The Journal of Finance***.

(21) Chordia T., Subrahmanyam A., Anshuman V. R., 2001, "Trading Activity and Expected Stock Returns", ***Journal of Financial Economics***, Vol. 59, pp.3~32.

(22) Cochrane J. H., 2011, "Presidential Address: Discount Rates", ***Journal of Finance***, Vol. 66, pp.1047~1108.

(23) Corazza M., Durbán M., Granéa., Perna C., Sibillo M., 2018, *Mathematical and Statistical Methods for Actuarial Sciences and Finance: Maf 2018*. Springer, Cham

(24) Dimson E., 1979, "Risk Measurement When Shares are Subject to Infrequent Trading", ***Journal of Financial Economics***, Vol. 7, pp.197~226.

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

(26) Fama E. F. & French K. R., 1992, "The Cross‐Section of Expected Stock Returns", ***The Journal of Finance***, Vol. 47, pp.427~465.

(27) Fama E. F. & French K. R., 2008, "Dissecting Anomalies", ***The Journal of Finance***, Vol. 63, pp.1653~1678.

(28) Fama E. F. & MacBeth J. D., 1973, "Risk, Return, and Equilibrium: Empirical Tests", ***Journal of Political Economy***, Vol. 81, pp.607~636.

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

(30) Frazzini A. & Pedersen L. H., 2014, "Betting Against Beta", ***Journal of Financial Economics***, Vol. 111, pp.1~25.

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

(32) Gu S., Kelly B., Xiu D., 2019, "Autoencoder Asset Pricing Models", ***Working Paper***.

(33) Gu S., Kelly B., Xiu D., 2019, "Empirical Asset Pricing Via Machine Learning", ***Working Paper***.

(34) Harvey C. R. & Siddique A., 2000, "Conditional Skewness in Asset Pricing Tests", ***The Journal of Finance***, Vol. 55, pp.1263~1295.

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

(36) He Z. & Krishnamurthy A., 2013, "Intermediary Asset Pricing", American Economic Review, Vol. 103, pp.732~770.

(37) Horvitz E. & Mulligan D., 2015, "Machine Learning: Trends, Perspectives, and Prospects", ***Science***, Vol. 349, pp.253~255.

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

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

(40) Huber P. J., 1992, *Robust Estimation of a Location Parameter*.

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

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

(43) Kelly B. & Jiang H., 2014, "Tail Risk and Asset Prices", ***The Review of Financial Studies***, Vol. 27, pp.2841~2871.

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

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

(46) Lee C. M. C., Qu Y., Shen T., 2019, "Going Public in China: Reverse Mergers Versus Ipos", ***Journal of Corporate Finance***, Vol. 58, pp.92~111.

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

(48) Li F., Zhang H., Zheng D., 2018, "Seasonality in the Cross Section of Stock Returns: Advanced Markets Versus Emerging Markets", ***Journal of Empirical Finance***, Vol. 49, pp.263~281.

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

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

(51) Lou D., 2014, "Attracting Investor Attention through Advertising", ***The Review of Financial Studies***, Vol. 27, pp.1797~1829.

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

(53) Maio P. & Philip D., 2015, "Macro Variables and the Components of Stock Returns", ***Journal of Empirical Finance***, Vol. 33, pp.287~308.

(54) Markowitz H., 1952, "Portfolio Selection", ***The Journal of Finance***, Vol. 7, pp.77~91.

(55) Merton R. C., 1973, "An Intertemporal Capital Asset Pricing Model", ***Econometrica***, Vol. 41, pp.867~887.

(56) Qiao F., 2019, "Replicating Anomalies in China", ***Working Paper***.

(57) Rapach D. & Zhou G., 2018, "Sparse Macro Factors", ***Working Paper***.

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

(59) Sirignano J., Sadhwani A., Giesecke K., 2018, "Deep Learning for Mortgage Risk"arXiv.org,Vol.

(60) Welch I. & Goyal A., 2008, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction", ***The Review of Financial Studies***, Vol. 21, pp.1455~1508.