机器学习驱动的基本面量化投资研究

李 斌， 邵新月， 李玥阳

[摘要] 基本面量化投资是近年来金融科技和量化投资研究的新热点。作为人工智能的代表性技术，机器学习能够大幅度提高经济学和管理学中预测类研究的效果。本文系统性地运用机器学习，来提升基本面量化投资中的股票收益预测模块。基于 1997 年 1 月至 2018 年 10 月 A 股市场的 96 项异象因子，本文采用预测组合算法、Lasso 回归、岭回归、弹性网络回归、偏最小二乘回归、支持向量机、梯度提升树、极端梯度提升树、集成神经网络、深度前馈网络、循环神经网络和长短期记忆网络等 12 种机器学习算法，构建股票收益预测模型及投资组合。实证结果显示，机器学习算法能够有效地识别异象因子—超额收益间的复杂模式，其投资策略能够获得比传统线性算法和所有单因子更好的投资绩效，基于深度前馈网络预测的多空组合最高能够获得 2.78%的月度收益。本文进一步检验了因子在预测模型中的重要性，发现交易摩擦因子在 A 股市场具有较强的预测能力，深度前馈网络在筛选因子数据上的多空组合月度收益达到了 3.41%。本文尝试将机器学习引入基本面量化投资领域，有助于促进人工智能、机器学习与经济学和管理学的交叉融合研究，为推进国家人工智能战略的有效实施提供参考。

[关键词] 基本面量化投资； 市场异象因子； 机器学习； 深度学习

[中图分类号]F424 [文献标识码]A [文章编号]1006-480X(2019)08-0061-19

DOI:10.19581/j.cnki.ciejournal.2019.08.004

一、引言

随着居民财富和可投资资产的快速增长，如何让资产保值增值，是每个家庭和个人都会面临的问题。更重要的是，资产管理也是金融服务实体经济的重要手段之一，积极为实体经济提供金融“血液”和资本力量。作为金融科技创新形式之一的智能量化投资，通过人工智能技术开展资产管理业务，能够大幅度提升资产管理的效率，正在成为中国金融业高质量发展的重要组成部分。例如，平安资产管理公司认为智能量化投资将是未来的主流投资方式，于 2018 年进行量化转型和科技赋能，

[收稿日期] 2019-04-26

[基金项目] 教育部人文社会科学研究青年项目“机器学习与技术分析融合视角下资产收益预测与投资组合策略研究”(批准号 18YJCZH072)；武汉大学人文社会科学青年学者学术团队建设计划项目“大数据驱动的投资管理研究团队”(批准号 WHU2016012)；国家自然科学基金重大研究计划“大数据驱动的管理与决策研究”重点支持项目“基于知识关联的金融大数据价值分析、发现及协同创造机制”(批准号 91646206)。

[作者简介] 李斌，武汉大学经济与管理学院教授，博士生导师，计算机博士；邵新月，武汉大学经济与管理学院硕士研究生；李玥阳，武汉大学经济与管理学院硕士研究生。通讯作者：李斌，电子邮箱：binli.whu@whu.edu.cn。

感谢岳阳和陈梦玄在收集数据中提供的帮助，感谢张学勇、唐礼智、赵沛霖、李正洋、谭明奎等学者的中肯建议，感谢匿名评审专家和编辑部的宝贵意见，当然文责自负。

致力于打造科技型资产管理公司。与此同时，人工智能的迅速发展正在深刻地改变着人类社会。2017年中国发布《新一代人工智能发展规划》,将发展人工智能上升为国家战略。在此背景下,如何发挥以机器学习为代表的人工智能技术优势,推动资产管理水平提升,逐渐引起中国金融业界和学术界的兴趣和广泛关注。

基本面量化投资(Quantamental Investing)融合了量化投资(计算机驱动)与价值投资(人为驱动),是近年来备受关注的一种智能量化投资方式(Lee and So,2015;张然和汪荣飞,2017)。其核心是分析股票的基本面因素和风险溢价(或超额收益)之间的关系,或股票收益的准确预测。当前学术研究中的基本面因素通常来源于市场异象的研究(Lee and So,2015),即能够提供超额收益的公司特征。尽管现有研究提出了数以百计的被认为能够提供超额收益的市场异象因子,但后续的样本外检验发现大部分因子难以持续地提供超额收益(Green et al.,2017;Hou et al.,2019)。异象因子的大量涌现也对传统的资产定价方法提出了技术挑战:①资产风险溢价的候选因子多达数百个,且很多因子极为相近(Gu et al.,2018),而传统的组合排序(Portfolio Sorts)和Fama-MacBeth回归并未综合考虑各因子,也未考虑因子间的交互作用(Bali et al.,2016);②当因子维度变大时,线性和非线性的考虑使得预测函数形式的搜索复杂度急剧增加,几乎无法通过人工去指定,但现有研究方法并未提供高维因子与预测函数形式选择的建议。以上两个技术难题呼唤着新研究工具的介入,前美国金融学会会长Cochrane认为,在处理如此众多的因子时,必须使用"不同的研究工具"(Cochrane,2011)。

作为人工智能的代表性技术,机器学习和深度学习是其中强有力的备选工具(Mullainathan and Spiess,2017;苏治等,2017;黄乃静和于明哲,2018;Athey and Imbens,2019)。机器学习和深度学习包含众多类型的研究方法,如监督学习(Supervised Learning)、无监督学习(Unsupervised Learning)、半监督学习(Semi-Supervised Learning)等。这三种研究方式的主要区别在于对数据样本标签的要求。监督学习需要样本的标签(比如股票收益),无监督学习无需标签,半监督学习则需要部分标签。本文选择(监督)机器学习来分析异象因子与超额收益之间关系,原因有三:①资产收益预测本质上是一个预测问题，而机器学习和深度学习旨在自动地寻找数据中的复杂结构和模式来辅助预测(Hastie et al.,2009;周志华,2016)。在资产收益预测中,收益数据的存在从本质上决定了该问题是一个监督学习的任务。因此,监督学习中的回归方法天然地适用于资产收益预测研究。②针对前述两个技术难题,机器学习的三个特性使其适用于该预测问题。通过众多备选的预测函数形式,无论线性模型还是非线性模型,机器学习提供了一系列丰富的方法来实现更加准确的预测;很多机器学习方法(如深度神经网络等)专门被设计用于逼近复杂的非线性关系(Goodfellow et al.,2016);参数正则化(Regularization)和模型选择(Model Selection)等技术使得在选择预测函数时不易过拟合(Overfitting)而导致虚假发现。③现有研究中丰富的异象因子为机器学习提供了有理论基础的输入变量,使得本文的研究区别于很多纯粹的金融数据挖掘。本文认为,在股票收益预测上运用机器学习方法,将有助于从全新的视角理解股票截面收益的决定因素,并能更好地预测股票收益和助力基本面量化投资。

尽管机器学习和深度学习研究方法天然地适用于解决股票收益预测面临的挑战，但根据机器学习理论中的"没有免费午餐定理"(No Free Lunch Theorem)(Wolpert,1996),本文并不能预知哪个算法会取得最好的预测效果①。因此,机器学习算法在中国股票收益预测问题上的表现也就成了

① 周志华(2016)也给出了该定理的一种推导。通俗地讲,"没有免费午餐定理"说明算法A在特定数据集上的表现优于另一种算法B的同时,一定伴随着算法A在另一个特定的数据集上的表现不如算法B。

一个实证问题,需要系统性检验。但是现有机器学习方法预测股票收益多从方法论的创新出发(李斌等,2017),仍缺乏系统性的研究来检视其作用和效果。就笔者所知,本文系首次系统性地运用机器学习方法检验中国市场的股票收益预测及基本面量化投资。

上述分析表明,本文研究的关键在于建立预测框架从而有效地度量异象因子与股票超额收益间的线性与非线性关系,并进行系统性检验。由此,本文提出两个研究问题:①机器学习算法能否有效地识别出异象因子和超额收益间的线性和非线性关系,从而依据预测构建的投资组合能够获得更好的绩效?②若机器学习算法的运用能够提升投资的绩效,哪些因子是真正重要的?基于机器学习算法筛选出的重要因子与传统单因子分析中显著的因子存在哪些差异?本文将针对这两个问题展开具体研究。本文收集了中国 A 股市场的 96 个异象因子,再基于 12 种代表性的机器学习算法分别构建了异象因子—超额收益预测模型,最后根据模型预测构建投资组合。这 12 种机器学习方法包括预测组合算法、Lasso 回归、岭回归、弹性网络回归、偏最小二乘回归、支持向量机、梯度提升树、极端梯度提升树、集成神经网络、深度前馈网络、循环神经网络和长短期记忆网络。

系统性的对比和分析发现:①机器学习算法能够自动识别异象因子间的线性和非线性模式并获得更好的投资绩效,其中非线性机器学习算法的绩效提升尤为突出。基于深度前馈网络预测的多空组合能够获得 2.78%的月度收益,而基于线性回归预测构建的多空组合月度收益仅为 2.01%。即便考虑到中国股票市场的做空限制和交易成本等因素,深度前馈网络模型构建的做多组合平均月度收益也达到 2.86%,而同期上证综合指数的月度收益仅为 0.61%。同时,所有机器学习模型的多空组合收益率均在 1%水平上显著。②异象因子在预测模型中重要性的分析结果显示,交易摩擦因子在 A 股市场具有较强的预测能力,而采用线性方法和非线性方法所得的重要因子区别于单因子检验所得的重要因子。进一步研究发现,机器学习驱动的基本面量化投资模型在重要因子集合上能够获得更好的绩效,多空组合月度收益率最高达 3.41%。

本文的研究有一定的现实意义与理论贡献:①丰富了经济学和管理学研究的工具箱。近年来,经济学和管理学界开始探讨机器学习对现有研究范式的价值和意义[①],也开始逐步地将其应用于经济学和管理学研究(Mullainathan and Spiess,2017;Kleinberg et al.,2018;伊志宏等,2019),但总体而言,国内外经济学和管理学的研究人员对该工具的使用仍在探索中,相关研究相对缺乏(苏治等,2017;黄乃静和于明哲,2018)。本文提供了一个典型的机器学习应用于金融问题的研究案例,为机器学习在经济学和管理学中的深入研究提供了参考。②丰富了量化投资的理论和实践研究。资产管理行业的科技化与智能化是未来的基本趋势,本文的研究有助于从新的机器学习视角厘清基本面量化投资的资产定价机理,为行业的高质量发展提供理论支撑。同时,本文的基本面量化投资模型可以无缝地运用于资产管理公司,具有潜在的实践价值。③丰富了中国市场中股票截面收益影响因素的研究。基于目前最全的 96 个异象因子库,本文从机器学习的视角重新检视了股票截面收益的决定因素。除了采用经典的单组合分析或线性回归方法(Hsu et al.,2018;Jiang et al.,2019),本文首次运用非线性机器学习方法检验了异象因子与股票收益间的非线性关系。非线性机器学习算法习得的重要因子与单组合分析和线性方法分析所得的重要因子区别较大,实证上从非线性的视角丰富了中国市场股票截面收益影响因素的研究。

① 比如,美国金融学年会(AFA2017、AFA2019)和美国经济学年会(AEA2018)已经连续三年设置了机器学习和经济金融的专题;2019 年中国工业经济杂志社与厦门大学联合举办了《中国工业经济》"机器学习在经济学和管理学中的应用"专题研讨会。

二、文献综述

在过去数十年的研究中,学术界宣称发现了数以百计能够提供超额收益的异象(因子),前美国金融学会主席 Cochrane(2011)称之为“因子动物园”(Factor Zoo),但对于其中真正有效的因子一直存有争议。众多学者从各个角度通过组合排序(Portfolio Sort)和 Fama-MacBeth(FM)回归等现有方法系统地重新检验异象,发现仅有部分仍能持续地提供超额收益(Harvey et al.,2016; McLean and Pontiff,2016;Green et al.,2017;Linnainmaa and Roberts,2018;Hou et al.,2019)。

近年来,已经有学者开始探索运用机器学习方法来解决因大量因子涌现而给传统研究方法带来的技术挑战。主要研究思路有以下三种:①采用变量选择方法衡量因子对于资产定价的贡献。比如 Feng et al.(2017)采用 Lasso 方法来衡量因子对资产定价的贡献,发现盈利因子和投资因子比之前发现的数百种因子更具有统计上显著的解释力。②运用机器学习方法从因子中提取共同因素来解释截面收益。比如 Light et al.(2017)采用“偏最小二乘法”(Partial Least Square,PLS)来检验公司特征对期望收益的预测能力;Kozak et al.(2019)和 Kelly et al.(2019)分别运用 PCA 方法和 IPCA 方法(Instrumented PCA)提取因子中的共同因素,发现基于少数主成分的模型就可以独立地预测截面收益。③设计新的集成方法以获取更好的预测效果。比如 Lewellen(2015)发现通过 FM 回归方法综合 15 个因子能够很好地预测股票超额收益;DeMiguel et al.(2017)从投资者效用的角度预测截面收益的公司特征,发现 6 个公司特征可以独立地预测平均收益;Gu et al.(2018)检验了常见的机器学习算法在美国市场上的表现,发现机器学习模型可以有效地超越传统线性回归模型。

同美国市场的研究相比,中国股票市场上将机器学习与因子策略结合的研究则相对缺乏。潘莉和徐建国(2011)检验了六个因子与股票收益的关系,构建了适用于 A 股市场的因子模型。胡熠和顾明(2018)从安全性、便宜性和质量三个维度共选取 8 个异象因子构造综合性指标,并将其应用于中国 A 股市场,验证了巴菲特的价值投资策略在中国市场的适用性。Hsu et al.(2018)检验了美国市场的常见异象在中国市场上的有效性,发现中美市场因子的有效性存在着很大区别。Jiang et al.(2019)分别采用 FM 回归、PCA、PLS 和 FC 等线性方法整合 A 股市场中的 75 个异象因子,发现这些方法能够从因子中提取出有助于预测的信息。

除了上述相关研究以外,也有大量文献运用机器学习方法预测股票价格或收益。李斌等(2017)分别采用支持向量机、神经网络、Adaboost 等机器学习算法,利用 19 项技术指标预测股价涨跌方向,发现这些算法具有更高的预测准确率,而根据预测所构建的投资组合也取得了更好的投资绩效。Krauss et al.(2017)集成了深度神经网络 DNN、梯度提升树 GBDT 和随机森林 RF 策略,利用过去所有股票的收益来预测标普 500 的涨跌。Fischer and Krauss(2018)采用长短期记忆模型(Long Short-Term Memory,LSTM),利用日频收益率数据预测股票相对于其截面中值收益率的涨跌方向,而相应构建的投资组合绩效显著优于其他线性模型。

三、研究设计

1. 模型总体设计

图 1 展示了机器学习驱动的基本面量化投资模型的整体框架。在图中的“证券池的设计”模块,本文选择市场中所有的证券;“资产定价”模块运用机器学习模型集成异象因子来预测股票收益;“投资组合”模块则根据预测构建投资组合,包含多空组合、多头组合和空头组合等;“交易仿真”则根据投资组合模块所给出的头寸进行仿真交易。本文的主要创新在于图中虚线框中的资产定价模块。

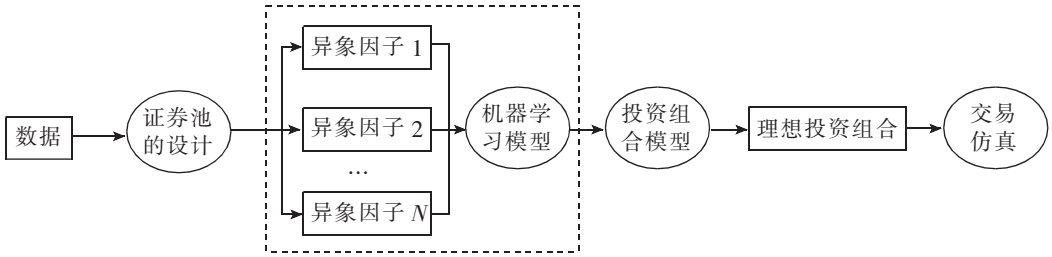

图 1　机器学习驱动的基本面量化投资框架

注：虚线框内的资产定价模块为本文的主要创新点所在。

资产定价模块的任务是一个标准的监督学习和回归任务，即发现如下的函数形式：

$$R_{t,i}=f(x_{t-1,i};\theta)+\epsilon_{t,i}$$

其中，$f(\cdot)$定义一个参数为 θ 的函数，在本文中为丰富的机器学习和深度学习方法中的函数形式，$R_{t,i}$为股票 i 第 t 期的超额收益，$x_{t-1,i}=(x_{t-1,i,1},x_{t-1,i,2},\cdots,x_{t-1,i,N})$为公司 i 在第 $t-1$ 期的异象因子向量，$\epsilon_{t,i}$为误差项。本文将在后面介绍采用的机器学习和深度学习算法。

在确定某一个具体的函数形式 $f(\cdot)$之后，本文将采用决策时点之前的数据进行拟合决定模型的参数。为了保证计算的有效性和投资的可行性，将采用图 2 所示的滑动窗口法划分训练和测试数据集(DeMiguel et al.，2009)。模型训练和测试的步骤如下：①假定目前处于 1998 年 1 月初，模型将决定该月的投资组合。本文采用过去 12 个月(即 1997 年)的异象因子—超额收益数据作为训练集，用于拟合机器学习模型得到模型参数。②将训练所得的模型用在 1997 年 12 月的异象因子数据之上，得到模型对 1998 年 1 月股票收益的预测。③基于步骤②的预测，在截面上对股票排序并构建多空组合，即做多预期收益最好的 10%股票，做空最差的 10%，多头和空头头寸均为等权重配置。④持有该组合一个月，得到投资组合在 1998 年 1 月的投资收益。⑤时间到 2 月初，本文重复步骤①—④。以此类推，直至数据期末。

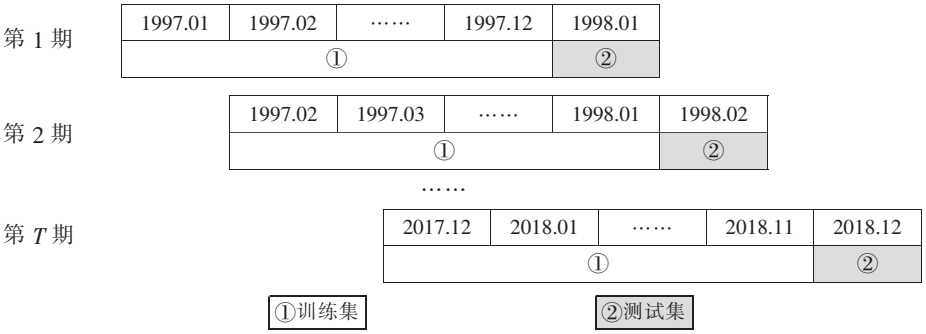

图 2　滑动窗口法示意

相较于常见的训练集和测试集划分方式(如留出法、交叉验证等)，滑动窗口法与现实中的投资决策过程一致，保留了数据中的时间序列特征。将训练集控制在一个固定区间内，可以减少模型的训练时间。与月度交易频率一致，测试集长度固定为 1 个月。训练集的默认区间长度为 12 个月，可得到 250 个训练和测试数据集。同时，为验证模型稳健性，本文也检验了训练集长度在 3 个月、24

个月和36个月时的模型表现。

所有机器学习模型参数的选择均采用网格调参(Grid Search)的方式。首先设定一个初始参数池,然后在训练集上针对每个参数训练得到多空组合的收益,筛选得到最优参数[①];同时,将最优参数应用于测试集上,由此得到最终的投资绩效。理论上,随着窗口的滑动,每个模型的最优参数也会随之改变。由于每一期网格搜索的计算成本较高,本文仅在第一个滑动窗口的训练集中进行调参。在此后的窗口滑动过程中,模型参数保持不变。所以,在本研究中,不同时期的模型最优参数是固定的,即为第一个滑动窗口训练所得的最优参数。

为了评估模型的绩效,本文将依据模型预测构建多空组合的月度收益率和风险调节收益(包括Fama-French三因子调节的阿尔法、Fama-French五因子调节的阿尔法和夏普比率等)作为绩效衡量指标。由于A股市场做空机制不够灵活,本文还分别构建了多头组合和空头组合,其中多头组合等权重持有预测收益前10%的股票,空头权重等权重做空预测收益后10%的股票。

本文进一步通过计算去除某个因子所导致的收益损失来衡量异象因子在预测模型中的重要性。具体来说,本文将某个特定因子从全因子集合中剔除,以剩余因子构建预测模型及投资组合,并计算该组合与基于全因子构建的组合月度收益差值。去除某个因子所导致的收益损失越大,则该因子越重要。作为比较基准,单因子筛选过程则采用因子构建多空组合月度收益的显著性作为因子重要性的度量标准,即因子组合的收益越显著,因子越重要。

本文将采用"跑马比赛"的方式研究三大类共12种机器学习算法的实证表现。与其他研究仅聚焦于单一算法不同,本文系统性地检验多种算法的动机如下:①按照机器学习领域"没有免费午餐定理"(Wolpert,1996),本文无法预知哪个算法在资产收益预测任务上会取得最好的效果。②选择多种不同机制的算法有助于全面理解机器学习算法在资产收益预测这一特定金融问题上的运用与效果,最终为经济、金融和管理学提供机器学习这一新的研究工具。③在机器学习领域,算法在数据上的实证表现也是一个重要的研究问题(Fernández-Delgado et al.,2014)。在金融领域,也有众多研究聚焦于系统性地检验一种方法在不同异象因子的实证表现(Mclean and Pontiff,2016;Harvey et al.,2016;Green et al.,2017;Hou et al.,2019)。与这两类研究相区别,本文检验了多种机器学习算法聚合多个异象因子的实证表现。

同时,笔者注意到券商研报也开展了类似的研究[②]。尽管检验方法上类似,但本文的研究与券商研报之间有着明显的区别:①最主要的区别在于,本文采用学术研究中验证过的异象因子,均以现有文献的理论解释为基础,更多地着眼于因子对于股票截面收益的可预测性,在学术研究上具有重要的理论价值与意义,而券商研报选取的因子并非全是学术研究中验证过的异象因子,相当一部分因子缺乏理论支撑。比如华泰证券股份有限公司的研报所选取的61个因子中有19个因子缺乏文献或理论支撑。②本文系统性地检验了12种机器学习算法,并通过实证检验揭示了预测资产收益时算法挖掘线性与非线性模式的不同效果,而券商研报则孤立地分析单一算法。缺乏系统性的检验和观察使得券商研报略显肤浅,难以深入剖析机器学习这一新的研究工具在经济学和管理学研究中的作用与边界。

2. 机器学习预测算法

机器学习是众多预测函数形式$f(\cdot)$及其各种算法的集合。如前所述,股票收益预测是一项监督

① 由于本文算法较多,算法的具体说明、初始参数池和不同滑动窗口时的最优参数可访问《中国工业经济》网站(http://www.ciejournal.org)附件获取。

② 比如华泰证券股份有限公司金融工程部门于2017年至2019年6月发布了近21篇人工智能专题报告。

学习的回归任务，而适用于回归任务的机器学习和深度学习算法均可以被用来建立股票收益预测模型。参照机器学习算法在其他预测研究中的表现，本文拟通过系统性地检验机器学习驱动的基本面量化投资模型来验证如下三个现象：①第一个观察：如果预测模型$f(\cdot)$采用线性函数形式，那线性机器学习算法的绩效能否超越传统计量模型中的线性回归模型？②第二个观察：如果预测模型$f(\cdot)$采用非线性函数形式，那非线性机器学习算法的绩效能否超越第一个观察中线性机器学习算法及传统线性回归模型？③第三个观察：如果预测模型$f(\cdot)$采用非线性函数形式，那么深度学习算法的绩效能否超越传统非线性机器学习算法？

为了验证上述三个观察，本文选择传统的线性回归模型作为基准，并选取三大类共 12 种代表性的机器学习算法。为了验证第一个观察，本文选取了 5 种线性机器学习模型，包括基于线性回归的预测组合模型（Forecast Combination，FC）、岭回归（Ridge）、Lasso 回归（Lasso）、弹性网络回归（Elastic Net）和偏最小二乘回归（Partial Least Square，PLS）。选取预测组合模型主要是因为其在金融预测问题被运用并取得了较好的效果（Rapach et al.，2010）。选取岭回归、Lasso 回归和弹性网络是因为它们是三种极具代表性的线性模型（Hastie et al.，2009）。选取偏最小二乘回归同样是因为其在金融学术研究中被运用并取得了较好的预测效果（Light et al.，2017）。

为了验证第二个观察，本文选取了 7 种机器学习和深度学习算法，其中 4 种传统机器学习算法包括支持向量机（Support Vector Machines，SVM）、梯度提升树（Gradient Boosting Decision Tree，GBDT）、极端梯度提升树（Extreme Gradient Boost Tree，XGBoost）、集成神经网络（Ensemble Artificial Neural Network，EN-ANN），3 种深度学习算法包括深度前馈网络（Deep Feedforward Neural Network，DFN）、循环神经网络（Recurrent Neural Network，RNN）和长短期记忆网络（Long Short-Term Memory，LSTM）。选择 SVM 主要是因为在深度学习出现之前，SVM 一直是机器学习理论的核心算法之一，在很多任务上取得了较好的结果。本文选取了两种随机森林类的算法，即 GBDT 和 XGBoost，其原因是 Fernández-Delgado et al.（2014）检验了 179 种分类算法的表现，得出结论是随机森林类算法在绝大多数分类任务中可以取得理想的结果。

近年来，深度学习算法在各类任务中表现卓越（Goodfellow et al.，2016），为了进一步验证第三个观察，本文选取 4 种神经网络类的算法：集成神经网络（EN-ANN）是多个神经网络的集成版本，从理论上预期该算法能够取得较单神经网络更好的效果。深度前馈网络（DFN）是一种易于使用的深度学习算法（Goodfellow et al.，2016）。循环神经网络（RNN）和长短期记忆网络（LSTM）这两种深度学习算法均考虑了数据中时序的影响（Goodfellow et al.，2016）。同时 LSTM 能够学习长距离依赖关系，在很多问题上能够取得比 RNN 更好的预测效果。

值得说明的是，本文选取的算法并非机器学习回归算法的完整集合。尽管不能穷尽机器学习算法，但选取的几种代表性算法在其他领域已经取得了较好的预测表现。可以预期，选取的算法能够很好地支撑本文在股票收益预测和基本面量化投资中拟观察的三个现象。

在模型设计中，本文将基本面量化投资的资产定价模块建模为监督式学习和回归任务。因此，本文仅选取了 12 种监督式学习模型。无监督学习在金融中也有应用，比如运用聚类算法提取异象因子中的共同因素等，但就本研究而言，股票收益的存在从本质上决定了监督学习方法天然适用于股票收益预测问题，本文也无需为了适应无监督学习和半监督学习而丢弃掉部分或全部的股票收益标签。

同时，本文的模型是基于方程视角（Hastie et al.，2009），即学习得到回归方程$f(\cdot)$后再进行预测，而机器学习的另一重要研究方向是基于概率视角（Murphy，2012），即学习数据的概率分布。从概

率视角出发去理解机器学习,可以有效地提高机器学习模型的可解释性。但在本文研究中,从概率视角求解存在三个难点:①概率视角要求数据量充分大,否则参数估计会不稳定,而目前每个滑动窗口内的数据量有限,在有限样本条件下,从方程视角求解机器学习也获得了很多成功;②从概率视角理解机器学习,需要了解或假设参数和异象因子的先验,但在本文的研究问题中,这些先验难以获得,并且也很难验证先验的准确性;③从概率角度理解和推导机器学习模型,往往需要引入大量额外参数(如高斯回归中的协方差矩阵等),优化求解比较困难,这些参数对于模型分析有帮助,但在预测中不是必需的。因此,本文仅从方程视角出发,而未考虑概率视角。

本文运用已有的机器学习算法来实现基本面量化投资,因此,采用了现成的算法包。所有数据处理与机器学习实现均基于 Python 语言, 其中传统机器学习算法基于 scikit-learn 算法包实现(Pedregosa et al.,2011),深度学习算法基于 MXNet 软件包实现(Chen et al.,2015)。深度学习算法均调用了 GPU 建模实现,部分运算量较大的环节(如计算因子重要性和调参等)基于 40 台机器的计算机集群实现①。

3. 数据来源与样本选取

本文选取 1997 年 1 月至 2018 年 10 月中国 A 股市场所有上市公司为研究样本,数据为月度频率。由于 1996 年 12 月 16 日起上海证券交易所和深圳证券交易所对上市的股票和基金的交易实行 10%的涨跌幅限制,为避免这一重大交易机制的变化对研究结果的影响,样本从 1997 年 1 月开始。选择月度频率是为了同相关研究保持一致(Jiang et al.,2019;胡熠和顾明,2018)。

借鉴 Green et al.(2017),本文选取了 96 个公司特征变量代理异象因子,并按照因子属性分为交易摩擦因子、动量因子、价值因子、成长因子、盈利因子、财务流动因子共六大类②。财报数据大部分为季度公布,本文采用季度数据进行了月度填充。由于上市公司财报披露时间存在延迟③,填充数据的基本原则是仅在规定的报表全部可用后再进行填充。数据均来自 CSMAR 数据库。为了更好地理解所构建的异象因子数据库,本文采用标准的资产定价方法检验了所有异象因子④。

本文选取的输出变量为考虑现金股利再投资的股票月度收益率。比如,用于预测 1997 年 1 月股票收益率的输入变量为 1996 年 12 月的 96 个异象因子。将 $t-1$ 期的公司特征和 t 期的月度收益率配对后即可获得“公司—月”数据。

初始股票池为 A 股市场所有股票。由于市场中 ST 股票业绩亏损,存在退市风险,为避免对研究产生额外的影响,本文剔除了 ST 股票;金融行业股票部分指标计量方式有别于其他上市公司,本文也剔除了金融类股票;由于 IPO 抑价效应的存在,股票上市第一年的股价可能存在异常波动,本文也剔除了股票上市第一年的数据。同时,数据库中存在着一定比例的缺失值,其处理过程分为两步:

① 感谢武汉大学金融科技与量化投资实验室提供的计算设备支持, 运行环境为 Core i7-6850K+32G+GTX1080Ti(12G)。

② 具体异象因子的构建方式可访问《中国工业经济》网站(http://www.ciejournal.org)附件获取。

③ 中国证券监督管理委员会《上市公司信息披露管理办法》规定,上市公司年报的披露时间为每个会计年度结束之日起 4 个月内;季报的披露时间为每个会计年度第 3 个月、第 9 个月结束后的 1 个月内编制完成并披露。

④ 感谢匿名评审专家的建议。系统性地重新检验异象因子是金融学的研究热点之一。比如 Hou et al. (2019) 采用统一的组合排序方法重新检验了 447 个文献中报告的异象因子。由于本文的核心并非单因子的显著性,且篇幅有限,因此,本文将实证检验结果放置于《中国工业经济》网站(http://www.ciejournal.org)的附件,供有兴趣的读者参考。

①若某只股票在第 t 月收益数据存在缺失(通常由股票连续停牌造成,共包含 3730 条缺失值),则剔除该股票在月份 t 上的所有数据;②若某只股票的因子值存在缺失,则以 0 填充。

在剔除 ST 股票、金融股、股票上市首年数据并处理完缺失值后,1997 年 1 月至 2018 年 10 月的有效样本为 381062 条。图 3 中展示了 1997 年 1 月至 2018 年 10 月月度有效样本量。总体来看,月度样本量随年份呈现上升趋势,由 1997 年 1 月的 307 条有效样本增至 2018 年 10 月的 3211 条。

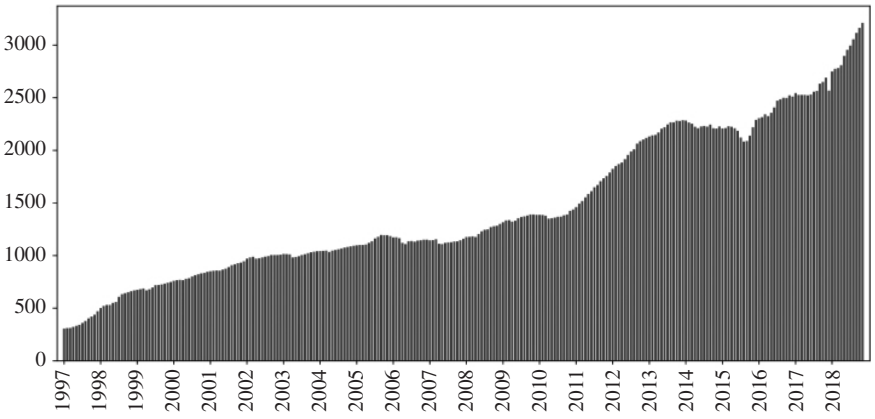

图 3　数据集中的月度有效样本量

因子的描述性统计显示,不同因子的取值在数量级及分布上存在显著差异,可能导致预测偏差,比如:①量级较大的特征在预测时占据主导地位;②数量级的差异会引起部分机器学习算法迭代收敛速度减慢。由此,本文将训练集数据标准化,假设这些样本来自某一均值为 0,方差为 1 的随机变量。标准化方式为:

$$X_{scale}=\frac{X-\bar{X}}{\sigma_X}$$

其中,$\bar{X}$和 σ_X 分别为变量 X 的均值和标准差。

四、实证结果与分析

1. 机器学习驱动的基本面量化投资模型在 A 股市场的绩效

本文检验了机器学习驱动的基本面量化投资模型在 A 股市场的实证绩效。表 1 展示了 12 种机器学习方法在 12 个月滑动窗口时多空组合、多头组合和空头组合的风险收益情形。作为对比,表中列出了基于线性回归模型预测的组合收益(OLS)、单因子检验中平均收益最高的市值因子构建的组合收益(SIZE)和市场指数的收益(MKT,仅多头)。

观察 Panel A 可以发现:①从多空组合的收益可以看出,最好的单因子模型超过市场指数收益(MKT),显示了因子投资研究的有效性;而线性回归模型能够取得超越最好的单因子的绩效,显示了基本面量化投资的重要性。②同为线性模型,线性机器学习算法(FC,Ridge,Lasso,ElasticNet,PLS)均能够获得较基准 OLS 回归更高的多空组合收益,表明机器学习算法能够更好地识别因子间的线性关系从而提升投资绩效。③非线性机器学习算法总体而言能够获得比线性算法更好的绩效,显示了异象因子间非线性模式的存在。其中,GBDT、XGBoost 和 DFN 三种非线性算法带来的投资

表 1　　机器学习驱动的基本面量化投资模型在 A 股市场的投资绩效

Panel A:组合投资绩效

	多头组合			空头组合			多空组合		
	Mean (%)	FF5-α (%)	夏普比率	mean (%)	FF5-α (%)	夏普比率	mean (%)	FF5-α (%)	夏普比率
OLS	2.35	0.82	0.7108	0.34	-0.97	0.0460	2.01	1.58	1.5088
FC	2.62	1.00	0.7853	0.34	-0.76	0.0463	2.28	1.55	1.2931
Ridge	2.41	0.88	0.7334	0.33	-0.98	0.0422	2.08	1.65	1.5469
Lasso	2.44	0.87	0.7409	0.36	-0.98	0.0527	2.08	1.64	1.5300
Elastic	2.46	0.90	0.7489	0.34	-1.00	0.0473	2.12	1.70	1.5694
PLS	2.48	1.00	0.7549	0.19	-1.13	-0.0084	2.30	1.92	1.5709
SVM	2.52	1.00	0.7770	0.27	-1.06	0.0199	2.25	1.86	1.7378
EN-ANN	2.59	1.01	0.7910	0.25	-1.07	0.0124	2.34	2.34	1.8082
XGBoost	2.72	1.04	0.8310	-0.01	-1.32	-0.0782	2.73	2.15	2.0066
GBDT	2.67	1.03	0.8143	-0.01	-1.31	-0.0779	2.68	2.13	1.9264
DFN	2.86	1.25	0.8595	0.08	-1.10	-0.0470	2.78	2.14	2.0150
RNN	2.52	1.10	0.7871	0.00	-0.01	0.0763	2.10	1.79	1.9794
LSTM	2.86	1.18	0.8584	0.29	-0.01	0.0298	2.57	2.01	1.9670
SIZE	2.45	0.51	0.7068	0.72	0.02	0.2034	1.73	0.27	0.6823
MKT	0.61	-0.01	0.1795						

Panel B:基本面量化投资策略收益的 Newey and West (1987) t 值

	多头组合		空头组合		多空组合	
	mean	FF5-α	mean	FF5-α	mean	FF5-α
OLS	3.2755	3.3728	0.5299	-5.1562	6.6004	5.7030
FC	3.8991	3.6078	0.5548	-3.4657	6.5055	4.3407
Ridge	3.3631	3.6109	0.5136	-5.1797	6.8460	5.8722
Lasso	3.3815	3.4890	0.5543	-5.0803	6.6450	5.7436
Elastic	3.4190	3.6474	0.5301	-5.1316	6.8489	5.9722
PLS	3.3824	3.8011	0.2884	-5.5474	6.8837	6.2067
SVM	3.5469	4.0976	0.4103	-5.5382	7.5761	6.7214
EN-ANN	3.5553	4.2140	0.3784	-5.7212	8.1185	6.7538
XGBoost	3.7138	4.3610	-0.0138	-6.4265	9.0991	7.3924
GBDT	3.6357	4.2110	-0.0139	-6.3163	8.6613	7.1469
DFN	3.8534	5.0756	0.1255	-5.7653	8.5983	7.6142
RNN	3.5811	5.4974	0.6519	-6.5062	8.4566	8.9661
LSTM	3.8501	5.9988	0.4477	-6.2635	8.0456	7.9333
SIZE	3.3326	3.2667	1.2327	0.1332	3.6258	1.6174
MKT	1.1757	-2.7593				

注:结果基于全变量和 12 个月滑动窗口,样本区间为 1997 年 1 月至 2018 年 10 月。

绩效提升最为显著,多空组合月度收益率较传统OLS 回归分别提升了 33.33%,35.82%和 38.30%,夏普比率则分别提升了 27.68%,32.99%和 33.55%。④深度学习中的 DFN 算法明显超过了线性算法和其他非线性算法,表明深度学习算法能够更有效地识别非线性模式。此外,Panel B

显示,所有机器学习算法投资收益的 Newey and West (1987)t 值均为 1%显著。总之,所有机器学习驱动的基本面量化投资模型均能超越单因子模型的收益(12 个月滑动窗口情形下,单因子表现最好的 SIZE 多空组合月度收益为 1.73%,对应夏普比率 0.6823),表明机器学习算法能够提升基本面量化投资模型的绩效表现。

由于中国股票市场缺乏有效的做空机制,单独做多预测收益前 10%的股票组合也应予以关注。Panel A 的结果显示,多头组合的收益远高于空头组合的收益,表明多空组合的收益主要来源于多头头寸。根据 XGBoost、LSTM 和 DFN 构建的做多组合能够获得相对较高的投资绩效,其中表现最好的是基于 DFN 和 LSTM 预测的做多组合,较 OLS 而言月度收益率可提升 21.70%。同时,Panel A 中最后一行(MKT)显示,同期上证综指的平均月度收益仅为 0.61%,且不显著。基于 FF5-α 和夏普比率的观察也一样。由此可见,各种基本面量化投资策略的多头组合均能超越市场指数的平均收益和风险调节收益,显示其在中国市场的适用性。

为验证机器学习算法相对于传统线性回归模型的优越性, 表 2 展示了各种机器学习算法与 OLS 回归所构建投资组合收益率差异显著性的检验结果,以及表现最好的 DFN 模型相对于其他传统机器学习算法的收益差异检验结果。数据显示,几乎所有非线性机器学习算法都较线性回归投资绩效有明显提升,证明机器学习算法能够通过识别数据间的非线性模式有效提升预测效果。对 DFN 与其他模型预测效果的检验结果显示,DFN 能够显著超越 Ridge、Lasso、ElasticNet、PLS 和 SVM 获得更高的投资绩效,证明了深度学习算法挖掘非线性模式的有效性。

表 2　基于 Newey and West(1987)的不同策略间收益差异性检验(12 个月滚动窗口)

	OLS 与其他算法				DFN 与其他算法		
	多头组合	空头组合	多空组合		多头组合	空头组合	多空组合
FC	1.8044	1.8044	0.0201	FC	1.5921	-1.6284	1.8126
Ridge	1.4412	1.4412	-0.8061	Ridge	4.0901	-1.8952	3.4790
Lasso	1.4247	0.4539	0.7950	Lasso	3.8110	-2.1672	3.5617
ElasticNet	2.1662	0.1431	1.5268	ElasticNet	3.5755	-2.0009	3.2768
PLS	1.1332	-1.3566	1.4788	PLS	3.1101	-0.8950	2.3957
SVM	2.5459	-1.1565	2.4288	SVM	2.9808	-1.5177	2.6016
EN-ANN	1.6926	-0.7538	1.4916	EN-ANN	1.8888	-1.2340	1.8226
XGBoost	2.6377	-2.7590	3.2382	XGBoost	1.0217	0.6688	0.2266
GBDT	2.3413	-2.6646	3.0821	GBDT	1.3841	0.6528	0.4485
DFN	4.2601	-1.9629	3.6477				
RNN	1.1670	0.6897	0.3742				
LSTM	3.5156	-0.3451	2.5548				

2. 机器学习算法集成后的基本面量化投资模型在 A 股市场的绩效

为了直观地说明机器学习算法相对于线性回归模型的绩效提升, 本文集成了 Ridge、Lasso、ElasticNet、PLS、EN-ANN、XGBoost、GBDT、SVM、DFN、LSTM、RNN 共 11 种机器学习算法 (由于 FC 模型仅是 OLS 回归的组合,集成模型中未包含 FC),在集成算法预测基础之上构建投资组合。具体而言,集成预测(Ensemble)为 11 种机器学习算法预测的算术平均值,股票 i 在 t 月的预测收益为:

$$R_{t,i}^{ensemble}=\frac{1}{11}\sum_{j=1}^{11}R_{t,i}^{j}$$

表3展示了3个月、12个月、24个月和36个月滑动窗口时集成算法的投资绩效，表4则对集成算法与线性回归模型的差异进行显著性检验。结果显示,除36个月滑动窗口外,基于集成算法的基本面量化投资策略在多空组合和单独做多/空组合上均能够获得显著优于OLS算法的收益和风险调节收益。

表3　基于11种机器学习算法集成的基本面量化投资模型在A股市场的投资绩效

	3个月滑动窗口			12个月滑动窗口		
	多空组合	多头组合	空头组合	多空组合	多头组合	空头组合
Mean(%)	2.56***	2.60***	0.04	2.98***	2.89***	-0.09
	(7.0439)	(3.5751)	(0.0683)	(9.2611)	(3.8840)	(-0.1451)
FF3-α(%)	2.38***	1.34***	-1.20***	2.67***	1.41***	-1.46***
	(6.7331)	(4.7965)	(-6.8497)	(8.8577)	(5.8175)	(-8.1355)
FF5-α(%)	2.18***	1.20***	-1.21***	2.55***	1.28***	-1.48***
	(6.2674)	(4.7571)	(-6.2301)	(9.2451)	(6.1019)	(-7.7384)
夏普比率	1.4698	0.7715	-0.0651	2.1736	0.8720	-0.1085
	24个月滑动窗口			36个月滑动窗口		
	多空组合	多头组合	空头组合	多空组合	多头组合	空头组合
Mean(%)	2.69***	2.78***	0.09	1.96***	2.41***	-0.45
	(8.0896)	(3.5986)	(0.1321)	(5.8373)	(3.0867)	(0.6592)
FF3-α(%)	2.37***	1.33***	-1.24***	1.70***	1.06***	-0.84***
	(7.6746)	(5.1849)	(-7.3478)	(6.2174)	(4.4701)	(-4.6277)
FF5-α(%)	2.17***	1.13***	-1.24***	1.48***	0.85***	-0.83***
	(8.3591)	(5.4178)	(-6.7986)	(5.1328)	(3.3757)	(-4.0735)
夏普比率	1.9439	0.8296	-0.0398	1.4439	0.7216	-0.0912

注:样本区间为1997年1月至2018年10月。

表4　基于11种机器学习算法集成的基本面量化投资模型与基准线性回归模型收益差异的显著性检验

滑动窗口(月)	多空组合	多头组合	空头组合
3	2.1306	2.2565	-1.4317
12	6.5091	5.2342	-4.8208
24	5.1849	4.3609	-4.2635
36	0.6321	1.4356	0.2487

3. 考虑交易成本时的实证绩效

实际投资中,每一笔交易都有交易成本,包含显性的券商佣金和隐性的买卖价差等。本文进一步检验了机器学习驱动的基本面量化投资策略在不同交易成本下的表现。本文考虑了交易成本(Transaction Cost,*transcost*)为单边0.50%、0.75%和1.00%的三种情形,表5展示了三种交易成本时投资组合的风险收益情形[①]。当交易成本为0.50%时,各种基本面量化投资策略仍能获得显著的超

① 这里和后面实证检验中的投资收益统计性检验结果可访问《中国工业经济》网站(http://www.ciejournal.org)附件获取。

额收益,DFN 算法仍获得最好的表现。当交易成本为 0.75%时,OLS、Ridge 和 Lasso 算法构建的多空组合收益不再显著,而其他算法多空组合仍能保持显著超额收益。而当交易成本上升至 1.00%时,XGBoost、GBDT 和 DFN 这三种算法的平均收益仍显著,而所有算法经 Fama-French 五因子调整后的超额收益均不再显著。考虑到 A 股市场上投资者的交易成本通常小于 0.5%,本文所提出的基本面量化投资策略在承担合理交易成本时仍能获得显著的收益与风险调节收益。

表 5　机器学习驱动的基本面量化投资策略在三种交易成本时的多空组合投资绩效

	transcost=0.50%			*transcost*=0.75%			*transcost*=1.00%		
	Mean (%)	FF5-α (%)	夏普比率	Mean (%)	FF5-α (%)	夏普比率	Mean (%)	FF5-α (%)	夏普比率
OLS	0.80	0.58	0.6710	0.30	0.08	0.2521	-0.20	-0.42	-0.1668
FC	1.07	0.55	0.6675	0.57	0.05	0.3547	0.07	-0.45	0.0419
Ridge	0.87	0.65	0.7209	0.37	0.15	0.3078	-0.13	-0.35	-0.1052
Lasso	0.87	0.64	0.7109	0.37	0.14	0.3014	-0.13	-0.36	-0.1081
Elastic	0.91	0.70	0.7474	0.41	0.20	0.3364	-0.09	-0.30	-0.0746
PLS	1.09	0.92	0.8176	0.59	0.42	0.4409	0.09	-0.08	0.0643
SVM	1.04	0.86	0.8846	0.54	0.36	0.4580	0.04	-0.14	0.0314
EN-ANN	1.13	0.87	0.9586	0.63	0.37	0.5338	0.13	-0.13	0.1090
XGBoost	1.52	1.15	1.2099	1.02	0.65	0.8115	0.52	0.15	0.4131
GBDT	1.47	1.13	1.1465	0.97	0.63	0.7565	0.47	0.13	0.3665
DFN	1.57	1.14	1.2311	1.07	0.64	0.8392	0.57	0.14	0.4473
RNN	0.89	0.79	0.9306	0.39	0.29	0.4062	-0.11	-0.21	-0.1182
LSTM	1.36	1.01	1.1327	0.86	0.51	0.7156	0.36	0.01	0.2984

注:这里基于 12 个月的滑动窗口,样本区间为 1997 年 1 月至 2018 年 10 月。

4. 剔除市值因子后的模型绩效

在单因子检验过程中,本文发现市值因子(*size*)构建的多空组合年化收益率为 20.76%(或月度收益率为 1.73%),显著高于其他因子。为了检验模型是否仅受到 *size* 因子的驱动而并非多因子带来的信息聚合效果,本文从全部 96 项因子集合中剔除了市值因子(*size*)和行业调整市值因子(*size_ia*),用剩余 94 项因子作为输入变量重新构建了基本面量化投资模型。实证结果显示①,在剔除市值因子后,各类模型的投资绩效较全变量(96 项因子)而言不存在明显差异,而 OLS、Ridge、GBDT、LSTM 和 DFN 五种算法的预测绩效小幅提升,可能的原因是 *size* 因子与其他因子存在较强的相关关系,剔除之后反而消除了部分共线性影响而提升预测效果。总之,实证结果显示机器学习驱动的基本面量化投资模型的绩效并非由 *size* 因子主导,而是多因子共同作用的结果。

5. 变动滑动窗口时的绩效

除采用 12 个月滑动窗口外,本文也分别采用了 3 个月、24 个月和 36 个月的滑动窗口。实证结果显示,除 36 个月滑动窗口外,DFN 算法均优于其他机器学习算法和 OLS 的预测效果(包括多空组合和做多组合),验证了 12 个月滑动窗口时所得结果的一般性。从各种滑动窗口的对比结果来看,采用较长的训练区间(如 36 个月)时的绩效相对较弱,而 3 个月、12 个月和 24 个月滑动窗口时的绩效不存在明显差异。

① 这部分和下一部分的实证结果可访问《中国工业经济》网站(http://www.ciejournal.org)附件获取。

五、进一步分析:异象因子的重要性

回答第二个研究问题涉及到检验异象因子在基本面量化投资模型中的重要性,以确定中国股票市场上预测能力最强的异象因子集合,从而从机器学习的视角进一步审视中国股票截面收益的影响因素。本文拟采用14种检验方法,包括单因子检验、OLS、FC、PLS、Lasso、Ridge、ElasticNet、SVM、EN-ANN、XGBoost、GBDT、DFN、RNN和LSTM等,其中单因子检验是传统资产定价研究中所采用的检验方法(Bali et al.,2016)。对于其他方法,本文计算去除某一因子后的收益损失来衡量该因子的重要性。将收益损失最大的因子重要性记为100%,剩余因子的重要性数值则按照收益损失最大值折算得到。本文将筛选出重要性数值位于前20的因子作为重要因子。

表6展示了单因子检验中显著的因子以及各项因子在线性和非线性算法所得重要因子中出现的次数。其中线性算法包括:OLS、FC、PLS、Lasso、Ridge、ElasticNet和SVM[①]等7种;非线性算法包括:EN-ANN、GBDT、XGBoost、DFN、RNN和LSTM等6种。结果显示,三类检验方式均对交易摩擦

表6　单因子检验中显著的因子以及分别在线性和非线性算法所得重要因子中出现次数不少于4次和3次的异象因子

单因子		线性(次数N最大为7)			非线性(次数N最大为6)		
因子	因子类别	因子	因子类别	$N\geq4$	因子	因子类别	$N\geq3$
stdvold	交易摩擦因子	*aeavol*	交易摩擦因子	7	*beta*	交易摩擦因子	4
LM	交易摩擦因子	*egr*	成长因子	7	*lagretn*	动量因子	4
turnsd	交易摩擦因子	*LM*	交易摩擦因子	7	*retvol*	交易摩擦因子	4
rd_mve	成长因子	*turnsd*	交易摩擦因子	7	*abacc*	成长因子	3
vold	交易摩擦因子	*CFdebt*	财务流动性因子	6	*aeavol*	交易摩擦因子	3
chfeps	盈利因子	*illq*	交易摩擦因子	6	*betasq*	交易摩擦因子	3
retmax	交易摩擦因子	*skewness*	交易摩擦因子	6	*egr*	成长因子	3
aeavol	交易摩擦因子	*chfeps*	盈利因子	5	*idskewness*	交易摩擦因子	3
lagretn	动量因子	*retvol*	交易摩擦因子	5	*idvol*	交易摩擦因子	3
illq	交易摩擦因子	*tang*	财务流动性因子	5	*PMG*	成长因子	3
skew	交易摩擦因子	*vold*	交易摩擦因子	5	*ROE*	盈利因子	3
size	交易摩擦因子	*idvol*	交易摩擦因子	4	*skewness*	交易摩擦因子	3
idvol	交易摩擦因子	*retmax*	交易摩擦因子	4	*stdvold*	交易摩擦因子	3
SGINVG	成长因子	*sharechg*	交易摩擦因子	4	*turnsd*	交易摩擦因子	3
pchsaleinv	财务流动性因子				*vold*	交易摩擦因子	3
momchg	动量因子						
size_ia	交易摩擦因子						
mom_36	动量因子						
retvol	交易摩擦因子						
SP	价值因子						
CRG	财务流动性因子						
depr	成长因子						

① 将SVM列入线性算法是因为12个月滑动窗口时,网格调参选择了线性核,其决策函数为线性。

因子存在明显偏好，但不同类型算法筛选出的重要因子差异明显。计算三类算法所得重要因子的 Spearman 秩相关系数也显示相互之间的相关程度较低，其中单因子和线性算法的重要因子相关系数为 0.28，单因子和非线性算法的因子相关系数为 0.08，线性和非线性算法的因子相关系数为 0.13。

本文进一步统计了各个因子被 13 种算法筛选为重要因子的次数，表 7 展示了 16 个被选中次数超过 5 次的重要因子，其中交易摩擦类因子 11 项，证明此类因子在 A 股市场具有极强的预测能力。另一方面，尽管超过 2/3 的异象因子利用上市公司财务报表数据构建（价值、成长、盈利和财务流动性因子），但结果显示此类因子在中国市场的预测能力相对较弱，筛选出的重要因子多为采用股票交易数据计算所得的异象因子。此外，收益公告异常交易量（*aeavol*）、股东权益变化（*egr*）和换手率的波动率（*turnsd*）因子被选为重要因子的次数最多，说明这三项因子在中国市场具有很强的预测能力。表 8 展示了各类因子中重要因子的占比。在全部 6 类因子中，交易摩擦类因子被选为重要因子的比例为 52%，证明交易摩擦因子相对更强的预测能力并非是由于交易摩擦因子本身占比较大所致。同时，价值因子未被选为重要因子，显示其在 A 股市场上的预测能力相对较弱。

表 7　　线性和非线性算法中累计被选中次数超过 5 次的重要因子

序号	因子	因子名称	因子类别	*N*
1	*aeavol*	收益公告异常交易量	交易摩擦因子	10
2	*egr*	股东权益变化	成长因子	10
3	*turnsd*	换手率的波动率	交易摩擦因子	10
4	*LM*	标准化的换手率	交易摩擦因子	9
5	*retvol*	总波动率	交易摩擦因子	9
6	*skewness*	总偏态	交易摩擦因子	9
7	*vold*	交易额	交易摩擦因子	8
8	*CFdebt*	现金流负债比	财务流动性因子	7
9	*idvol*	异质波动率	交易摩擦因子	7
10	*illq*	非流动性风险	交易摩擦因子	7
11	*tang*	偿债能力/总资产	财务流动性因子	7
12	*chfeps*	预期每股收益的变化	盈利因子	6
13	*lagretn*	短期反转	动量因子	6
14	*retmax*	最大日收益率	交易摩擦因子	5
15	*sharechg*	股本增长率	交易摩擦因子	5
16	*stdvold*	交易额的波动率	交易摩擦因子	5

表 8　　各类因子中被选中超过 5 次的重要因子占比

因子类别	因子总数	重要因子数	占比（%）
交易摩擦因子	21	11	52
财务流动性因子	10	2	20
动量因子	6	1	17
盈利因子	14	1	7
成长因子	35	1	3
价值因子	10	0	0

交易摩擦因子更强的预测能力可能有以下两点原因:①交易摩擦因子主要是依据交易数据构造的月频数据,及时性更高并包含了更多的市场信息(而根据财报数据构建的异象因子为季度数据,人工填充为月度),因此能更好地反映整体市场的短期预期。②交易摩擦因子稳定性相对更高。以 *aeavol* 因子为例,无论经济周期如何变化,市场对公司收益公告的反应不会大幅波动。

最后,表 9 中展示了以这 16 项重要因子作为模型输入构建投资组合的绩效。结果显示,所有算法较全变量均存在不同幅度的提升,其中 DFN 具有最好的绩效表现,多空组合获得了 3.41%的平均月度收益,夏普比率为 2.0182。

表 9　　机器学习驱动的基本面量化投资模型的投资绩效

	多头组合			空头组合			多空组合		
	Mean (%)	FF5-α (%)	夏普比率	Mean (%)	FF5-α (%)	夏普比率	Mean (%)	FF5-α (%)	夏普比率
OLS	2.65	1.08	0.8192	0.10	-1.19	-0.0405	2.55	2.06	1.6652
FC	2.84	1.18	0.8612	-0.15	-1.31	-0.1283	2.98	2.28	1.7573
Ridge	2.65	1.08	0.8192	0.10	-1.19	-0.0405	2.55	2.06	1.6652
Lasso	2.68	1.10	0.8292	0.06	-1.20	-0.0527	2.62	2.09	1.6744
Elastic	2.67	1.09	0.8240	0.07	-1.20	-0.0483	2.59	2.07	1.6652
PLS	2.73	1.16	0.8462	0.04	-1.16	-0.0619	2.69	2.10	1.7739
SVM	2.51	1.00	0.7811	0.12	-1.18	-0.0335	2.39	1.97	1.5760
EN-ANN	2.46	0.85	0.7570	0.30	-0.98	0.0322	2.16	1.61	1.7334
XGBoost	2.85	1.21	0.8789	-0.01	-1.29	-0.0783	2.86	2.30	2.0182
GBDT	2.81	1.15	0.8610	0.07	-1.22	-0.0519	2.74	2.16	1.8468
DFN	3.24	1.44	0.9533	-0.17	-1.40	-0.1394	3.41	2.63	2.0182
RNN	2.42	0.88	0.7547	0.46	-0.90	0.0882	1.95	1.57	1.5986
LSTM	2.91	1.22	0.8651	0.06	-1.21	-0.0538	2.85	2.22	2.0861

注:结果基于表 7 的重要因子集合,滑动窗口设置为 12 个月,样本区间为 1997 年 1 月至 2018 年 10 月。

六、结论与启示

1. 研究结论

本文收集了 1997 年 1 月至 2018 年 10 月 A 股市场的 96 个异象因子,构建了 12 种机器学习算法驱动的基本面量化投资模型,系统性地对比了机器学习驱动的模型与基于线性回归的模型在中国市场取得的实证绩效,并从机器学习的视角检验了异象因子在模型中的重要性。实证结果发现:对比不同类型算法的实证绩效,可以发现线性机器学习算法表现优于单因子和线性回归模型,而非线性机器学习算法表现总体优于线性机器学习算法,深度学习算法在该问题上能够取得最好的绩效。这一结果显示异象因子间存在着传统线性回归难以识别的非线性模式,而机器学习算法能够自动识别其中的非线性模式从而获得更好的预测效果及组合收益。稳健性检验显示,机器学习驱动的基本面量化投资模型在考虑卖空限制、交易成本、不同参数等情形时均表现稳健。从机器学习视角能够发现与传统单因子检验不一样的重要因子。从机器学习的视角看,以收益公告异常交易量为代表的交易摩擦类因子对股票截面收益有较强的预测能力。同时,机器学习驱动的基本面量化投资策略在重要因子数据上表现更好。

2. 启示与建议

(1)本文对机器学习在经济学和管理学中的应用研究具有重要的启示。机器学习在经济学和管理学中的应用主要有两个视角:①运用机器学习处理非结构化数据并提取代理变量,比如运用机器学习算法从文本中提取投资者情绪指标等;②在经济学和管理学中的预测问题上运用机器学习方法以提升其预测能力,尤其是样本外预测的效果。本文系统性地对比了 12 种机器学习算法在股票收益预测问题上的实证表现,主要在第二个视角上提供了一个典型范例,以启发其他经济学和管理学的研究。研究过程显示,在相关预测问题的建模中,可以采用现有研究中具有理论基础的影响因素作为模型的输入变量,同时也需要对于相关的机器学习预测方法进行有效的选择。这两方面的选择均可能影响到预测模型的实证绩效。

(2)本文的研究结论对于理解中国股票的截面收益有重要的启示。区别于传统研究聚焦于个别异象,本文通过变量选择的思想从聚合多个异象因子的角度考虑其对股票截面收益的影响。研究发现,交易摩擦因子对于中国股票截面收益有着重要的影响。据作者所知,该结论并未在相关研究中出现,从机器学习的视角能够发现与传统研究不一样的股票截面收益影响因素。这一研究方式亦可扩展至其他类似研究问题之中。

(3)本文的研究结论对资产管理行业实践有丰富的启示。智能化是近年来资产管理行业的一大潮流,而对于智能化在资产管理中的作用仍在探索中。本文研究发现,机器学习能够切实地提升资产管理的效率和效益,提供了一个潜在的新工具。同时,本文从基本面量化投资的角度为人工智能(含机器学习)在资产管理中的运用提供了一个视角。在实际运用时,资产管理公司也可以直接部署或改造本文的量化投资模型,将有助于资产管理实践的开展。

(4)本文的研究对于推进国家人工智能战略有一定的启示。人工智能作为一项通用性的技术,可以在金融行业的各个方面发挥作用,能够有效地节省成本和控制风险。本文的研究通过研究机器学习来助力资产管理行业,是人工智能在金融领域的典型运用。研究表明,现阶段智能金融模型仍离不开金融专家的经验,需要积极地引导机器学习在金融专家经验之上构建更加高效的智能金融模型。从金融专家的经验出发有助于控制智能金融或智能投资模型的风险,提升其稳健性。本文的研究结论将进一步启发国家人工智能战略在金融行业的应用与落地。

〔参考文献〕

〔1〕胡熠,顾明. 巴菲特的阿尔法:来自中国股票市场的实证研究[J]. 管理世界,2018,(8):41-54.

〔2〕黄乃静,于明哲. 机器学习对经济学研究的影响研究进展[J]. 经济学动态,2018,(7):115-129.

〔3〕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[J]. 系统工程理论与实践,2017,37(5):1089-1100.

〔4〕潘莉,徐建国. A 股市场的风险与特征因子[J]. 金融研究,2011,(10):140-154.

〔5〕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[J]. 金融研究,2017,(5):111-126.

〔6〕伊志宏,杨圣之,陈钦源. 分析师能降低股价同步性吗——基于研究报告文本分析的实证研究[J]. 中国工业经济,2019,(1):156-173.

〔7〕张然,汪荣飞. 基本面量化投资:运用财务分析和量化策略获取超额收益[M]. 北京:北京大学出版社,2017.

〔8〕周志华. 机器学习[M]. 北京:清华大学出版社,2016.

〔9〕Athey, S., and G. W. Imbens. Machine Learning Methods that Economists Should Know about[J]. Annual Review of Economics, 2019, https://doi.org/10.1146/annurev-economics-080217-053433.

〔10〕Bali, T. G., R. F. Engle, and S. Murray. Empirical Asset Pricing: The Cross Section of Stock Returns[M]. Hoboken, New Jersey: Wiley, 2016.

〔11〕Chen, T., M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang, and Z. Mxnet. A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems [R]. Neural Information Processing Systems, Workshop on Machine Learning Systems, 2015.

〔12〕Cochrane, J. H. Presidential Address: Discount Rates[J]. The Journal of Finance, 2011,66(4):1047-1108.

〔13〕DeMiguel, V., A. Martin-Utrera, and F. J. Nogales. A Transaction-Cost Perspective on the Multitude of Firm Characteristics[R]. LBS Working Paper, 2017.

〔14〕DeMiguel, V., L. Garlappi, and R. Uppal. Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy[J]. The Review of Financial Studies, 2009,22(5):1915-1953.

〔15〕Feng, G., S. Giglio, S., and D. Xiu. Taming the Factor Zoo: A Test of New Factors [R]. NBER Working Paper, 2017.

〔16〕Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems[J]. Journal of Machine Learning Research, 2014,15:3133-3181.

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

〔18〕Goodfellow, I., Y. Bengio, and A. Courville. Deep Learning[M]. Cambridge, Massachusetts: The MIT Press, 2016.

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

〔20〕Gu, S., B. Kelly, and D. Xiu. Empirical Asset Pricing via Machine Learning[R]. NBER Working Paper, 2018.

〔21〕Harvey, C. R., Y. Liu, H. Zhu. … and the Cross-Section of Expected Returns[J]. The Review of Financial Studies, 2016,29(1):5-68.

〔22〕Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. New York: Springer, 2009.

〔23〕Hou, K., C. Xue, and L. Zhang. Replicating Anomalies [EB/OL]. The Review of Financial Studies, 2019, https://doi.org/10.1093/rfs/hhy131.

〔24〕Hsu, J., V. Viswanathan, M. Wang, and P. Wool. Anomalies in Chinese A-Shares[J]. The Journal of Portfolio Management, 2018,44(7):108-23.

〔25〕Jiang, F., G. Tang, and G. Zhou. Firm Characteristics and Chinese Stocks [J]. Journal of Management Science and Engineering, 2019,3(4):259-84.

〔26〕Kelly, B. T., S. Pruitt, and Y. Su. Characteristics Are Covariances: A Unified Model of Risk and Return[EB/OL]. Journal of Financial Economics, https://doi.org/10.1016/j.jfineco.2019.05.001.

〔27〕Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. Human Decisions and Machine Predictions[J]. Quarterly Journal of Economics, 2018,133(1):237-293.

〔28〕Kozak, S., S. Nagel, and S. Santosh. Shrinking the Cross Section [EB/OL]. Journal of Financial Economics, https://doi.org/10.1016/j.jfineco.2019.06.008

〔29〕Krauss, C., X. A. Do, and N. Huck. Deep Neural Networks, Gradient-boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500[J]. European Journal of Operational Research, 2017,259(2):689-702.

〔30〕Lee, C. M., and E. C. So. Alphanomics:The Informational Underpinnings of Market Efficiency [J]. Foundations and Trends in Accounting, 2015,9(2-3):59-258.

〔31〕Lewellen, J. The Cross-Section of Expected Stock Returns[J]. Critical Finance Review, 2015,4(2):1-44.

〔32〕Light, N., D. Maslov, and O. Rytchkov. Aggregation of Information about the Cross Section of Stock Returns: A Latent Variable Approach[J]. The Review of Financial Studies, 2017,30(4):1339-1381.

〔33〕Linnainmaa, J. T., and M. R. Roberts. The History of the Cross-Section of Stock Returns [J]. The Review of Financial Studies, 2018,31(7):2606-2649.

〔34〕McLean, R. D., and J. Pontiff. Does Academic Research Destroy Stock Return Predictability[J]. The Journal of Finance, 2016,71(1):5-32.

〔35〕Mullainathan, S., and J. Spiess. Machine Learning: An Applied Econometric Approach [J]. The Journal of Economic Perspectives, 2017,31(2):87-106.

〔36〕Murphy, K. P. Machine Learning: A Probabilistic Perspective [M]. Cambridge, Massachusetts: The MIT press. 2012.

〔37〕Newey, W., and K. West. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix[J]. Econometrica, 1987,55(3):703-708.

〔38〕Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2011,(12):2825-2830.

〔39〕Rapach, D. E., J. K. Strauss, and G. Zhou. Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy[J]. The Review of Financial Studies, 2010,23(2):821-862.

〔40〕Wolpert, D. H. The Lack of a Priori Distinctions Between Learning Algorithms [J]. Neural Computation, 1996, (8):1341-1390.

Research on Machine Learning Driven Quantamental Investing

LI Bin, SHAO Xin-yue, LI Yue-yang

(Economics and Management School of Wuhan University, Wuhan 430072, China)

Abstract Quantamental investing is an emerging hot topic in financial technology and quantitative investments. As a representative technique in Artificial Intelligence (AI), machine learning can significantly improve the prediction task in economics and management. This paper investigates the application of machine learning in quantamental investing. Based on 96 anomaly factors in the Chinese stock markets ranging from January 1997 to October 2018, we adopt Forecast Combination, LASSO regression, Ridge regression, Elastic Net, Partial Least Square, Support Vector Machine, Gradient Boosting Decision Tree, Extreme Gradient Boosting Tree, Ensemble Artificial Neural Network, Deep Feedforward Network, Recurrent Neural Network, and Long-Short Term Memory to build stock return prediction model and construct portfolios. Empirical evidence shows that machine learning algorithms can efficiently identify complex patterns hidden in the anomaly factor and excess return, the quantamental investment strategy can deliver better performance than the traditional linear model and all factors. Long-short portfolios based on the forecast of Deep Feedforward Network can obtain a monthly return of 2.78%. We further evaluate factors' importance in the prediction model, and find that trading friction factors demonstrate better predictive ability in the Chinese stock markets. Deep Feedforward Network driven quantamental investing models running on the selected feature set provide the best performance of 3.41% per month. This study introduces the machine learning toolbox to the research on quantamental investing, which will further facilitate the joint research on AI, machine learning and economics and management and finally will boost the national strategy of AI.

Key Words: quantamental investing; market anomaly factors; machine learning; deep learning

JEL Classification: C80 G00 O11

〔责任编辑:覃毅〕