doi: 10.19920/j.cnki.jmsc.2023.10.008

中国股票市场可预测性研究: 基于机器学习的视角①

李 斌[1,2], 龙 真[1,2*]
(1. 武汉大学经济与管理学院, 武汉 430072; 2. 武汉大学金融研究中心, 武汉 430072)

摘要: 股市风险溢价是金融学中的一个经典研究问题. 常见的线性模型存在着模型误设和参数不稳定的问题,难以有效预测风险溢价. 本研究从机器学习的视角重新检视了中国股票市场的可预测性. 基于 1996 年 1 月—2019 年 12 月的数据,构建提升回归树 (boosting regression trees, BRT) 模型对股市收益率与波动率进行样本外预测,并构建了最优风险资产配置模型. 实证结果显示: 1) 提升回归树方法能够对收益率、波动率和最优风险资产权重做出准确预测; 2) 在收益率预测中最重要的三个变量分别是净权益增加值、换手率和股价方差; 挖掘预测变量之间的非线性关系是 BRT 预测能力的来源; 3) 结合提升回归树预测构建的最优风险资产组合可以为投资者带来更高的收益和效用. 本研究将机器学习方法引入股票市场风险溢价的研究,为此问题的研究提供了全新的视角.

关键词: 权益风险溢价; 提升回归树; 样本外预测; 机器学习
中图分类号: F830.91 **文献标识码:** A **文章编号:** 1007-9807(2023)10-0138-21

0 引 言

股票市场收益率和波动率的准确预测是金融学术界和实务界均关注的重要问题. 在学术上,股票市场的收益率和波动率是众多金融学研究的核心变量; 在实务中,股票市场收益率和波动率的准确预测将影响投资者的资源配置和投资收益. 不准确的预测将导致相关研究的误判,导致投资者的资源错配,而金融市场的资源配置将最终影响实体经济的发展. 众多国内外研究表明,股票市场收益存在显著的样本内可预测性,但样本外可预测性欠佳[1,2],更进一步,现有研究也未从投资者角度充分考虑预测的经济效益.

中国股市超额收益样本外预测面临的挑战主要来自两个方面: 第一,现有研究往往预设变量间为线性关系,但采用线性模型做预测时,容易产生模型误设和参数不稳定性问题[3]. 第二,股市收益率预测涉及大量经济变量,在现有参数方法中需要对大量参数进行估计,容易产生过拟合问题,导致模型的样本外预测能力较弱. 第三,股市预测数据量较小,可能导致复杂模型不再适用.

为了解决上述挑战,本研究拟从机器学习的视角重新审视中国股票市场的可预测性问题,原因有二: 首先,金融市场具有显著的非参特点,金融数据的复杂性与非线性导致很多传统方法不适用于金融数据的研究[5]. 机器学习 (machine learning) 旨在从数据中挖掘出内在模式并更好地预测[6],天然适用于金融资产可预测性的研究[7-9]. 其次,机器学习提供了众多模型和算法,可以灵活地挖掘变量间的复杂关系,并通过正则化与模型选择等技术避免过拟合,从而实现更准确的样本外预测.

① 收稿日期: 2020-11-16; 修订日期: 2022-06-27.
基金项目: 国家自然科学基金资助项目 (71971164; 72371191); 科技创新 2030-"新一代人工智能"重大项目课题资助项目 (2020AAA0108505); 国家社会科学基金资助重大项目 (20&ZD105).
通讯作者: 龙 真 (1998—), 女, 湖北宜昌人, 博士生. Email: zhenlong@whu.edu.cn

具体来说,本研究拟采用提升回归树 BRT (boosting regression tree) 来研究股票市场可预测性问题. 首先,现有研究发现,树型机器学习能够很好地挖掘金融数据中的复杂关系[7,8]. 作为树型机器学习算法的一种,BRT 也是一种非参数方法,无需先验假设,能够避免模型误设的问题; 它的一些技术特性,包括收缩和袋装,使其稳健性进一步加强[6]. 其次,股市时序收益率预测的主要挑战之一是样本量过少,比如中国市场仅有 288 个月度样本. 训练数据的不足使得神经网络类等模型复杂度较高的方法在该问题上不再适用[10],而 BRT 算法隐含着变量选择的功能,在无需大量数据的条件下,仍具有较好的样本外预测效果. 最后,复杂的机器学习算法常常被看作黑箱,难以直观解释,而 BRT 方法在具有优异性能的同时相对容易解释[11].

基于上述分析,本研究采用提升回归树 BRT 研究了以下三个问题: 第一,提升回归树方法对股市市场超额收益、波动率和最优风险资产权重的预测能力是否优于传统方法,比如线性回归模型与历史均值估计? 第二,若提升回归树方法表现出优异的预测能力,如何解释? 第三,从投资者角度出发,更好的预测能否带来更高的投资收益与效用?

本研究收集了中国股票市场 1996 年 1 月—2019 年 12 月的数据构建了 12 个经济变量,运用提升回归树方法进行中国股市收益率和波动率的样本外预测,并与传统方法进行了对比; 继而分析提升回归树预测能力的来源; 最后构建了最优风险资产配置策略,研究其预测是否给投资者带来收益与效用.

实证分析发现: 第一,提升回归树方法的收益率和波动率预测效果都优于传统方法. 提升回归树预测超额收益的样本外 R^2 达到 3.72%,远高于线性模型的 0.25%. 第二,分析变量的依赖关系发现,提升回归树能够挖掘变量中的非线性关系,从而做出更准确的预测. 与美国市场相异,净权益增加值是收益预测中最重要的变量,股价方差是波动率预测中最重要的变量. 这两个变量均与我国的市场结构相关. 第三,结合提升回归树预测构建的最优风险资产组合可以为投资者带来更高的收益和效用. 提升回归树方法可得到的月度超额收益为 1.18%,阿尔法为 0.55%,夏普比率为 0.13.

本研究的贡献主要体现在以下两点: 第一,从机器学习的视角重新审视了中国股市预测的问题,丰富了中国股票市场风险溢价预测的相关文献[12,13]. 就作者所知,本研究是第一篇系统性地运用非线性机器学习方法预测中国股市风险溢价的研究. 第二,从小数据的视角丰富了机器学习方法在金融预测问题上的研究. 与运用机器学习方法检验股票截面收益预测的大数据量不一样,股市风险溢价的预测是一个典型的小数据问题②,而 BRT 方法仍有较好的表现. 第三,借助于 BRT 的部分依赖方法(partial dependence),挖掘了预测变量和目标变量间非线性关系的证据,验证了机器学习提升股市风险溢价预测的来源,有助于进一步推进金融与机器学习的交叉融合研究[14].

1　文献综述

在过去数十年的研究中,众多学者主要从两个角度对资本市场风险溢价预测问题进行了研究. 第一个研究角度是挖掘具有市场收益率预测能力的变量. 除了 Welch 和 Goyal[2] 归纳的股息支付率等多种基本面指标外,学者们还发现产出缺口[15]、企业盈利总和[16]、短期利率[17]、工业材料价格[18]、石油价格[19]、股市政策或货币政策[20,21]、股市收益率动量[22]、和市场分歧指数[23]等指标与市场风险溢价有关. 姜富伟等[12]和张琳等[24]都指出通货膨胀率、换手率等宏观基本面指标对中国股市有较强预测能力. Goh 等[25]研究结果表明美国经济变量对中国资本市场有显著影响

② 比如 Gu 等[8]的美国股票截面收益数据多达 370 万条公司 - 月,李斌等[7]的中国股票截面收益数据超过 38 万条公司 - 月; 而本研究仅有 288 条股市 - 月样本.

与预测能力. Chen 等[26]研究指出我国宏观经济政策的不确定性与股票市场收益存在反向关系. 此外,基于非理性的市场情绪指标,如投资者情绪[27,28]、经理人情绪[29]、网络平台支持倾向[30];基于市场风险的波动率指标,如投资组合收益波动率[31]、时变尾部风险[32]、期权隐含尾部风险[33]、方差风险[34]等变量也被发现对股市收益率有预测作用. 但有学者研究发现,具有样本内预测能力的变量并不能有效预测样本外股市收益率[2],可能是由于许多研究忽略了预测中模型的不确定性和参数的不稳定性[3].

第二个研究角度是从提出新的模型方法出发,提高已有变量对市场收益率的预测能力. Welch 和 Goyal[2]和姜富伟等[12]均采用传统的简单线性回归方法,前者发现通过线性回归并不能有效预测美国股市收益,而后者研究发现中国市场收益率的样本内和样本外均有可预测性. 还有许多学者提出了创新模型方法研究市场收益率预测问题. 例如 Pettenuzzo 等[35]和 Rapach 和 Zhou[36]在最小二乘回归中加入经济机制约束以提高预测的准确性. Zhang 等[3]针对模型的不确定性和参数的不稳定性在线性模型中引入了收缩算法和窗口平均法,使模型预测能力显著提高. Ferreira 等[37]把市场收益率分解成市盈率对数增长率、企业盈利的对数增长率、对数化的股利价格比与当期股价变化,分别估计后加总,发现这种局部求和法得到了较好的预测结果. Neely 等[38]从 28 个经济指标和技术面指标中构造一个潜在因子预测美国股市超额收益. 蒋志强等[1]通过可行拟广义最小二乘法研究认为中国股市收益率是可预测的,并发现模型在样本内、样本外、熊市、牛市有不同的预测效果.

机器学习方法在资产定价领域的应用也在慢慢兴起. Baillie 等[39]指出金融和经济变量与股票收益率之间具有非线性关系. 同时资产定价中出现的大量因子使得传统的组合排序和回归研究方法不再适用[40],因此近年来众多学者运用机器学习挖掘金融和经济变量与资产收益率之间的非线性关系. Gu 等[8]和 Bianchi 等[9]分别对股票和债券的截面收益率进行预测,发现非线性的决策树方法和浅层神经网络表现最好. Tobek 等[41]采用了多种机器学习方法对国际市场上的个股截面收益率进行预测,发现非线性树型和神经网络的预测效果与经济效益比线性方法更好; Freyberger 等[42]在 Lasso 回归中引入非线性函数用于股票收益率预测,发现引入非线性函数后的模型预测表现更好. Krauss 等[43]集成了随机森林、梯度提升树和神经网络等模型,利用过去的股票收益来预测市场收益率的涨跌,表现出了显著的预测效果. 在中国股市上,Leippold 等[44]和李斌等[7]运用机器学习研究了股票截面收益预测,同样发现非线性方法能够获得较好的收益. 姜富伟等[45]构建了大数据集应用机器学习方法开发了智能动态 CAPM 模型,检验了时变系统性风险对股市收益的解释能力. 姜富伟等[46]采用 Lasso 等机器学习模型充分挖掘我国上市公司财务基本面大数据信息,也发现了机器学习方法在定价模型中的优异表现。这些研究均显示了在资产收益预测问题中挖掘预测变量与收益率之间非线性关系的重要性,非线性机器学习方法可以获得比传统方法更优的预测效果.

除了对收益率预测进行研究外,也有众多文献研究了股市波动率预测问题. 第一,现有研究发现有多种指标都和股市整体波动性有关. 夏婷和闻跃春[47]把股价波动性分为宏观经济的不确定性和经济政策的不确定性. 在经济指标方面,Paye[48]和 Engle 等[49]研究发现多种经济变量对波动率具有预测能力. Adrian 和 Rosenberg[50]、Kim 和 Nelson[51]都认为影响股市波动最重要的因素为经济周期. 还有研究发现通货膨胀[52]、投资者情绪[53]也与股市波动率有显著相关性,陈其安和雷小燕[53]利用中国市场数据研究了政策不确定性与股市波动的关系. 第二,许多研究从方法角度展开研究. 目前对波动性预测的模型方法以 GARCH 系列模型为主[54],陈强和叶阿忠[55]采用 EGARCH-M 模型研究了居民消费倾向对收益波动率的关系. 也有学者在 GARCH 模型基础上进行改进,Tong 等[56]、郑挺国和尚玉皇[57]和 As-

gharian 等[58] 构建 GARCH-MIDAS 模型对股市波动率进行预测分析,检验了基本面变量对股市波动率的预测力. Stefan 等[59] 将梯度提升方法应用于预测股票市场波动率,Christiansen [60] 采用贝叶斯模型平均 BMA(Bayesian model average) 方法对资产波动率进行研究,发现多种经济变量与金融变量都对资产价格波动率具有显著的样本外预测能力.

综上所述,本研究发现: 第一,对股票市场收益率有预测能力的变量主要集中在宏观经济变量上,现有研究已经较为充分. 但对预测模型的创新较为缺乏,针对中国市场的研究仍集中在线性方法上,关注于预测变量与被预测变量间的线性关系,而这些变量与收益率和波动率的关系很可能并非线性[39,61],李斌等[7]、Gu 等[8] 和 Leippold 等[44] 采用机器学习方法挖掘了公司特征与截面收益的非线性关系,说明忽略预测变量与目标变量之间的非线性关系可能不利于模型的预测. 第二,大多数研究仅孤立地讨论收益率或波动率,而鲜有文献将两者结合起来研究,未进一步研究预测的经济价值. 基于以上两点观察,本研究从机器学习方法的视角对上述问题展开了研究.

2　研究设计

2.1　模型框架

本研究从经济变量出发,构建 BRT 模型和多种对比模型分别预测中国股市的超额收益率与波动率,然后将两者结合得到最优风险资产配置权重,最终分析最优风险资产配置产生的效用.

超额收益预测模型表示为如下的函数形式

$$r_t = f_1(x_{t-1};\theta_1) + \varepsilon_t \tag{1}$$

$$\hat{r}_{t+1} = f_1(x_t;\theta_1) \tag{2}$$

其中 r_t 为第 t 期的市场超额收益率,$f_1(\cdot)$ 是参数为 θ_1 的预测模型,x_{t-1} 为第 $t-1$ 期的经济变量. 将拟合训练数据所得的模型运用于第 t 期的经济变量 x_t,可以得到第 $t+1$ 期的预期收益率.

波动率预测模型与收益率预测模型类似,区别在于目标变量由超额收益率换成了市场波动率

$$\sigma_t^2 = f_2(x_{t-1};\theta_2) + \varepsilon_t \tag{3}$$

$$\hat{\sigma}_{t+1}^2 = f_2(x_t;\theta_2) \tag{4}$$

其中 σ_t^2 为第 t 期的市场波动率,$f_2(\cdot)$ 为参数为 θ_2 的预测模型.

获得收益率和波动率的预测之后,本研究进一步构建风险资产与无风险资产的资产配置组合. 以流通市值加权的市场组合作为风险资产,则均值方差投资者的最优风险资产权重 $\hat{w}_{t+1}$ 为

$$\hat{w}_{t+1} = \frac{\hat{r}_{t+1}}{\gamma\,\hat{\sigma}_{t+1}^2} \tag{5}$$

其中 γ 为投资者的风险厌恶系数,$\hat{r}_{t+1}$ 和 $\hat{\sigma}_{t+1}^2$ 分别为式(2) 和式(4) 的预测值. 投资者的实现收益率为市场组合和无风险资产的加权收益率③.

2.2　数据与变量

股票市场收益采用 A 股市场上市公司(剔除了“ST”和“PT”公司) 的月度流通市值加权超额收益率作为收益率预测的因变量,以日度流通市值加权收益率平方和作为波动率预测的因变量. 以整存整取定期年利率作为无风险利率,按复利方法计算得到月度无风险利率. 参照现有研究[2,3,12,35],本研究选取了 12 个月度经济变量作为中国股市超额收益率的预测变量④. 由于数据可得性限制,时间区间为 1996 年 1 月—2019 年 12 月⑤. 参考我国股权分置改革完成时间在 2006 年末,以 2007 年 1 月为样本外估计起点,即 1996 年

③　本研究所采用最优权重的计算方法相对直接,主要是为了从投资者的角度衡量收益率和波动率两个预测的综合效果. 现有的文献中通常固定波动率预测来衡量投资者效用. 实际应用中可以针对收益率和波动率选择不同的预测变量和预测模型,或针对两者的组合直接进行优化.

④　本研究选用 12 个预测变量主要依据该研究的主要文献,具体列表联系作者邮箱获取. 除了经济变量以外,也有学者尝试运用经济变量以外的数据,比如 Dong 等[4] 首次运用 100 个异象因子的多空收益来预测整体股票收益. 本研究的重点并非在于变量的选择,因此仅选用了常见的 12 个经济变量. 有关提出新的预测变量的文献可以参考 Dong 等[4]、Adämmer 和 Schüssler[62]、Jacobsen 等[18]、Jiang 等[29] 等的文献. 本研究采用与和现有文献相同的变量也是为了聚焦于从机器学习方法论的视角针对该问题有所贡献.

⑤　12 个变量中最晚可得的数据为变量 *M2G*(M2 货币量增长率) ,从 1996 年 1 月开始.

1 月—2006 年 12 月作为样本内估计区间,2007 年 1 月—2019 年 12 月作为样本外预测区间. 在滚动窗口情形下,以 1996 年 1 月—2006 年 12 月的时间跨度 132 个月为窗口长度.

表 1 预测变量说明

Table 1 Predictive variable description

变量名	缩写	定义
股利支付率	*D/E*	总股利除以总盈利的对数值
股息价格比	*D/P*	总股利除以总市值的对数值
股息率	*D/Y*	总股利除以滞后一期总市值的对数值
盈利价格比	*E/P*	总盈利除以总市值的对数值
账面市值比	*B/M*	总账面价值除以总市值
股价方差	*SVR*	市值加权市场日收益率的平方和
通货膨胀率	*INFL*	根据国家统计局公布的 *CPI* 算出,由于每期通货膨胀率滞后一期公布,故 *INFL* 数据也滞后一期
净权益增加值	*NTIS*	过去 12 个月沪市与深市新股发行量的移动加总除以当月总市值
换手率	*TO*	A 股市场总交易量除以总市值
*M*0 增长率	*M*0*G*	当月流通现金总量的增长率
*M*1 增长率变动量	*M*1*G*	当月 *M*1 增长率与上月 *M*1 增长率的差值,其中 *M*1 定义为当月 *M*0 加上企事业单位的活期存款
*M*2 增长率	*M*2*G*	当月 *M*1 加上居民存折储蓄与定期存款总额的增长率

表 1 展示了 12 个月度预测变量的定义⑥. 其中,总股利为 A 股市场过去 12 个月的股利移动加总,总盈利为过去 12 个月的市场盈利移动加总. 账面价值数据参考公司年报、中报和季报. 由于公司报表往往滞后公布,数据可得性有所限制,每年 1 月—3 月的账面市值比为上年 6 月的账面价值除以当月市值,4 月—6 月的账面市值比为一季度季数据得到,7 月—9 月账面市值比为半年报数据得到,10 月—12 月为三季度季报数据得到. 由于 2002 年以前的企业季报数据缺失较多质量不高,2002 年前的数据填充均只根据年报与半年报数据,2002 年及以后的数据填充根据年报、半年报与季报. 其中,股票回报率、市值、股利、交易量与 *M*0、*M*1、*M*2 数据来自国泰安数据库(CSMAR),盈利、账面价值、新股发行数据与 CPI 数据来自万得数据库(Wind). ⑦

表 2 展示了单变量线性模型的收益率和波动率预测的样本内拟合优度、样本外拟合优度和方向准确率[2]. 从中可以有如下发现. 首先,单变量预测能力最强的变量为股利支付率(*DE*),其扩展窗口和滚动窗口情况下的 R^2_{OOS} 分别达到 1.83% 和 1.13%. 而其他变量均不能同时在扩展窗口和滚动窗口得到为正的 R^2_{OOS}. 其次,尽管变量在样本内表现出一定的预测能力,比如换手率(*TO*)变量的样本内拟合优度达到了 3.18%,但其样本外拟合优度大幅度下降至扩展窗口下的 0.21% 和滚动窗口下的 -1.29%. 这两个观察显示了样本外预测相对于样本内预测具有更高的难度. 最后,波动率预测的样本外拟合优度较高. 各单变量对波动率预测的 R^2_{OOS} 几乎都超过了 40%,其中波动率滞后项的预测效果最好,说明波动率序列本身具有较强的自相关性,波动率预测的 R^2_{IS} 结果则均小于 R^2_{OOS}.

2.3 研究设计

2.3.1 研究方法

为了实现收益率和波动率预测的模型 $f_1(\cdot)$ 和 $f_2(\cdot)$,本研究拟运用提升回归树 BRT 构建股票市场预测的模型,本节简要介绍选择 BRT 作为主要研究方法的经济机理. 由于篇幅

⑥ 部分变量采用了对数值,包括股利支付率、股息价格比、股息率、盈利价格比等. 主要是为了和现有文献一致;同时将其对数化也减少异方差和趋势性,使得序列平稳.

⑦ 如需解释变量的描述性统计,见在线附录 https://github.com/Zhen1236/ML_ChinaStockMarket

所限,关于 BRT 方法及对比方法的详细介绍请参见⑧.

金融市场是一个具有非参特点的动态系统,具有复杂且非线性的特征. 这一特征也导致了很多传统方法难以适用于金融的研究,而机器学习方法以挖掘数据中的复杂模式为目标,由此众多学者在金融问题中引入了机器学习方法,试图解决金融研究面临的非线性挑战,如近年来将机器学习运用在预测资产截面收益[7-9,44],挖掘大数据生成变量[29,62]等. 这一挑战同样体现在股市风险溢价预测难题上,因此本研究从机器学习的视角来重新审视金融市场风险溢价预测这一具体的金融问题,预期挖掘其中的非线性关系,从而做出更好的预测.

表 2　单变量线性模型的超额收益率预测结果/%

Table 2 Excess return prediction results of univariate linear models/%

被预测变量	收益率预测					波动率预测				
变量	扩展窗口		滚动窗口		样本内	扩展窗口		滚动窗口		样本内
	R^2_{OOS}	*Acc*	R^2_{OOS}	*Acc*	R^2_{IS}	R^2_{OOS}	*Acc*	R^2_{OOS}	*Acc*	R^2_{IS}
DE	1.83	57.05	1.13	58.97	1.74	42.59	56.41	40.37	53.85	0.94
DP	-2.00	55.13	-0.05	55.13	0.57	44.37	64.10	45.87	57.69	2.54
DY	-1.81	55.13	0.56	54.49	0.84	44.18	62.82	46.67	57.69	2.42
EP	-4.40	55.13	-2.47	52.56	0.07	43.10	61.54	43.20	52.56	1.46
BM	-2.04	55.13	-1.00	55.13	0.16	47.72	73.08	48.03	67.95	6.39
SVR	-3.52	55.13	0.04	55.13	0.01	65.27	75.64	66.87	74.36	27.80
INFL	-1.28	52.56	0.05	53.85	1.73	41.61	53.85	39.02	46.15	0.37
NTIS	-0.22	47.44	-1.66	48.72	1.01	43.30	66.67	42.26	57.69	3.24
TO	0.21	53.21	-1.29	51.92	3.18	47.14	58.97	45.55	60.26	10.22
M0G	-1.77	55.13	0.27	60.90	0.03	41.88	56.41	40.39	44.87	0.00
M1G	-1.95	57.69	-0.67	55.77	0.73	41.82	57.69	40.44	47.44	0.00
M2G	-2.20	55.13	-3.37	51.28	0.01	42.45	60.26	40.79	46.15	1.13

注: 此表报告了 12 个单变量回归模型对市场超额收益率和市场波动率的样本外与样本内预测表现. 包括 R^2_{OOS}、方向精度(*Acc*)和样本内 R^2(R^2_{IS}). 表中结果均表示百分数.

针对股市的收益率和波动率预测这一金融问题,本研究运用 BRT 方法挖掘其中非线性关系的原因主要包括: 1) 现有机器学习应用在金融和非金融领域的研究均发现,挖掘非线性表现最好的算法主要包括树型算法和神经网络类算法⑨. 从设计机制上讲,BRT 方法本身基于回归树,具有较好的挖掘变量间非线性关系的能力. 同时,它是一种非参数方法,无需先验假设,无需假定变量服从一定的概率分布. 在股市收益率和波动率预测问题中,经济变量的概率分布往往难以被准确估计[3],因此非参数方法在设定上避免了假设误设的问题; 2) 不同于股票截面收益的预测,股市收益率预测的主要问题之一是可用样本量过少,比如中国市场的市场-月样本量为 288,远小于机器学习应用中常见的大样本. 缺乏足够的训练数据使得神经网络类等模型复杂度较高的方法在该

⑧ 见在线附录 https://github.com/Zhen1236/ML_ChinaStockMarket

⑨ 根据机器学习中的"没有免费午餐定理"(no free lunch theorem),没有一种模型可以在所有应用上超越其他模型,其在应用中的具体表现是一个实证问题. 运用机器学习研究资产截面收益预测的文献显示,表现优异的模型主要是树型模型与神经网络类模型,具体可参考 Gu 等[8],Bianchi 等[9],Leippold 等[44] 和李斌等[7] 等. 非金融类的应用中,BRT 的一种实现软件包 LightGBM 在其网站列出了其在各类应用中的优异表现,详见https://github.com/microsoft/LightGBM

问题上不再适用[10]. 而 BRT 算法隐含着变量选择的功能,在每个节点选择一个特征与临界值将所有样本分为两类子样本,只有最相关的变量会被选择,冗余变量则会被忽略; 同时相比于 Lasso 等方法 BRT 能够更有效地提取变量间的交互关系[6]. 因此,BRT 在不需要大量数据的条件下,仍具有较好的预测效果; 3) BRT 对异常值和缺失值具有相当好的鲁棒性,收缩和袋装机制又加强了 BRT 方法的稳健性; 4) 复杂的机器学习算法常常被看作黑箱,难以直观解释,而 BRT 方法在具有优异性能的同时相对容易解释. 比如,BRT 可以绘制树形结构直观地展示其决策过程,衡量决策过程中变量的相对重要性,并可通过部分依赖图展示预测变量与目标变量之间的关系.

具体来说,提升回归树 BRT 以回归树为基本模型,加入提升算法,通过迭代计算减少每次模型拟合的残差,再加总所有迭代得到的模型实现最优拟合. 同时,BRT 通过收缩(shrinkage) 和袋装(bagging) 两项技术来提升模型的拟合和样本外预测的效果. 收缩技术是一种正则化技术,通过降低学习的速率避免模型过拟合,增强模型样本外预测的能力. 袋装技术是按均匀概率分布从数据集中重复有放回抽样的技术,该技术通过降低基本模型的方差改善泛化误差,即降低训练数据的随机波动导致的误差[10].

基于上述分析,由于 BRT 方法可以更好地挖掘变量之间的非线性关系,更有利于做出准确预测,所以本研究针对第一个研究问题做出假设: BRT 方法预测收益率与波动率的效果将优于其他对比模型. 同时,本研究预期通过结合 BRT 方法预测的收益率与波动率所得的最优风险资产权重的效果也将优于对比模型.

2.3.2 模型选择

本研究的模型选择采用了扩展窗口(expanding window) 与滚动窗口(rolling window) 两种方式[3]. 在扩展窗口情形下,本研究以前 t 期的所有样本为训练集拟合得到模型,对第 $t+1$ 期的市场进行预测; 在滚动窗口情形下,本研究固定滚动窗口长度为 132 个月,将前 t 期固定窗口作为训练集得到模型,对第 $t+1$ 期的市场进行预测. 以此类推按月向前滚动直到期末.

在超参数选择上,本研究将提升回归树的迭代次数设置为 1 000,学习速率设置为 0.001[11]. 结果显示超参数的选择对结论影响不大.

2.3.3 模型评估

为了度量和比较模型的样本外预测能力,本研究采用了 R^2_{OOS}、方向精度和经济效益三种指标[12]. R^2_{OOS} 的计算方式为[13]

$$R^2_{oos} = 1 - \frac{\sum_{t=T_0+1}^{T}(r_t - \hat{r}_t)^2}{\sum_{t=T_0+1}^{T} r_t^2} \tag{6}$$

其中 T_0 为样本内预测区间,T 为全部样本区间长度,r_t 为超额收益率真实值,$\hat{r}_t$ 表示预测值. R^2_{oos} 越大则说明模型预测能力越强,反之说明模型预测能力较弱. 为了度量 R^2_{oos} 的稳健性,避免异常值对结果产生较大影响,本研究还计算了去除 10% 相对误差最大的样本后得到的 R^2_{oos},记为 $R^2_{oos\text{-}robust}$.

收益的方向精度 Acc 即预测收益率与实际收益率正负方向一致的样本占比,方向精度越高的模型预测能力越好. 由于波动率无负值,定义波动率的方向为波动率与全样本波动率中位数相比的大小方向. 最优风险权重的方向精度定义与收益率的类似.

为了进一步从投资者角度衡量不同预测模型的经济效益,本研究还计算了基于不同预测模型的最优风险资产策略在不同风险厌恶系数条件下

[10] 如需模型的具体描述及与其他方法,见在线附录 https://github.com/Zhen1236/ML_ChinaStockMarket

[11] 如需不同超参数的样本外预测效果,见在线附录 https://github.com/Zhen1236/ML_ChinaStockMarket

[12] 文献中主要采用这三种样本外绩效评价指标[1-3,12,35]. 除此之外,蒋志强等[1] 还采用了相对均方根误差($RRMSE$) 等,但该指标可以转换为与 R^2_{OOS} 等价换算的指标,即 $R^2_{OOS} = 1 - RRMSE^2$. 如需具体选择依据,可联系作者邮箱获取.

[13] Welch 和 Goyal[2] 计算 R^2_{oos} 是以历史均值为比较基准,因此以历史均值估计误差平方和为分母. 本研究将 BRT 同时与线性模型、历史均值比较,因此参考 Gu 等[3] 以零估计为比较基准,故分母为收益率平方和.

的实现效用，表示为

$$U = \bar{r}_p - \frac{1}{2}\gamma Var\{r_{p,t+1}\} \quad (7)$$

其中 $\bar{r}_p$ 表示投资组合所得收益率序列的均值，$r_{p,t+1}$ 为 $t+1$ 期组合得到的收益率，$Var\{r_{p,t+1}\}$ 表示收益率序列方差，U 表示投资者得到的效用.

除了采用三个指标研究 BRT 模型的预测能力，本研究还分析了 BRT 模型预测能力的来源. BRT 模型中提供了特征重要性工具分析了变量在模型中的重要性，并提供了部分依赖图 PDP 来刻画预测变量与目标变量之间的线性或非线性关系. 特征重要性是每个变量作为内部节点时累计减少的误差平方和，这一累计增益越大则说明变量越重要. 部分依赖图则是指固定其他变量的取值，计算某一变量对目标变量的影响，图像中曲线越曲折说明该变量与目标变量非线性越强⑭.

3　股票市场样本外预测：基于 BRT 方法的研究

本节首先研究了 BRT 方法在股票市场收益率、波动率和最优风险权重三方面的预测能力，其次通过统计检验证实了 BRT 方法相对于现有主流方法的优越性，之后通过分析 BRT 中变量的相对重要性和变量对目标变量的边际效应研究了 BRT 优越于主流方法的原因所在.

3.1　BRT 模型的样本外预测能力

为了更好地预测股票市场的超额收益率、波动率和最优风险资产权重，本研究采用机器学习方法中的 BRT 算法.

表 3 展示了股票市场样本外预测的结果，分别包括了超额收益率、波动率和最优风险资产权重的预测. 作为对比，本研究也实现现有文献的主要基准方法，包括收益率预测中的历史均值估计(prevailing mean)、线性模型(linear model，亦被称为 kitchen sink 模型). 同时也参照相关文献在收益率预测上实现了三种线性机器学习模型(包括 Ridge、Lasso 和 ENet)、一种降维模型 PLS 和两种树型机器学习模型(包括随机森林 RF 和 Adaboost). 在波动率预测中，本研究实现了主流的 GARCH 和 MIDAS 模型，并实现了随机森林 RF 和 Adaboost⑮. 此外，由于 GARCH 模型与 MIDAS 模型的输入变量分别为市场超额收益率与历史高频波动率数据，难以直接与 BRT 模型比较，本研究还加入了 BRT-MIDAS 模型，即 BRT 方法采用 *MIDAS* 的历史高频数据为输入变量. 由于该研究的数据样本量较少，全样本仅有 288 个，难以满足神经网络类算法的数据量要求[10]，表 3 中未列出神经网络作为对比算法⑯.

从面板 A 的收益率预测结果可以发现，BRT 模型的样本外拟合优度达到了 3.72%(滚动窗口模式下为 3.96%)，优于所有对比模型. BRT 模型的预测方向准确率达到了 58.97%(扩展窗口)和 62.18%(滚动窗口)，均优于其他对比模型. 同时，树类算法(包括 BRT 方法、RF 方法和 Adaboost 方法)表现显著优于基准和线性类算法. 在扩展窗口下，非线性树类算法的平均拟合优度达到了 3.44%，大幅优于基准模型、线性机器学习模型和维度压缩模型. 这一观察与现有研究结论类似⑰. $R^2_{oos-robust}$ 的对比结果也与 R^2_{oos} 相似，说明这一结果对异常值稳健. 市场收益率微小的预测提升就能够转化为显著的经济效益[63]，在收益率预测任务中，非线性机器学习算法的表现均大幅度优于线性算法，显示了挖掘经济变量非线性关系的重要性，及其从机器学习视角重新审视该问题的必要性.

---

⑭ 如需特征重要性与 PDP 的具体计算方法，见在线附录 https://github.com/Zhen1236/ML_ChinaStockMarket

⑮ PLS 和 Lasso 回归等方法在波动率预测的研究中并不常见. 因此，在波动率预测中本研究仅加入了随机森林 RF 和 Adaboost 作为非线性机器学习模型的补充.

⑯ 作者参照 Gu 等[8] 和李斌等[7] 实现了 1 层～5 层隐藏层的前馈神经网络和长短期记忆模型 LSTM，但由于数据量不足导致的训练不足，其样本外预测拟合优度较差. 具体的预测结果可联系作者邮箱获取.

⑰ 现有文献发现，运用机器学习研究截面收益预测的文献发现，树型算法和神经网络表现最好(Gu 等[8]、Dong 等[4]、Bianchi 等[9]、Leippold 等[44]、李斌等[7])，其中神经网络甚至略优于树型算法. 本研究在市场股权收益溢价上的这一发现与现有文献中一致，树型算法表现最优. 但神经网络类算法在股市风险溢价预测的结果较差，与文献相悖.

表 3 对市场收益率、波动率和最优风险资产权重的样本外预测结果/%

Table 3 Out-of-sample prediction results of the market returns, volatilities, and optimal risky asset's weight/%

变量	扩展窗口			滚动窗口		
	R^2_{OOS}	*Acc*	$R^2_{OOS\text{-}robust}$	R^2_{OOS}	*Acc*	$R^2_{OOS\text{-}robust}$
Panel A: 收益率预测						
Prevailing Mean	-0.57	55.13	-0.14	0.18	53.85	0.70
Linear Model	0.25	50.00	2.76	-5.10	54.49	-2.43
Ridge	-1.37	50.00	0.43	0.62	55.77	2.59
Lasso	-0.62	55.13	-0.17	0.41	54.49	0.81
ENet	-0.62	55.13	-0.17	0.41	54.49	0.81
PLS	-2.75	55.77	-0.70	0.13	54.49	1.86
RF	3.17	57.69	5.56	3.82	61.54	6.38
Adaboost	3.43	55.77	4.51	2.42	58.33	4.05
BRT	3.72	58.97	6.83	3.96	62.18	6.04
Panel B: 波动率预测						
GARCH	56.47	73.08	52.99	53.44	71.79	49.97
MIDAS	63.58	75.64	59.75	63.54	75.64	59.09
RF	63.23	74.36	60.91	59.83	75.64	57.68
Adaboost	62.09	74.36	60.04	51.47	76.92	50.19
BRT	64.59	73.08	62.42	61.65	75.64	59.69
BRT-MIDAS	66.89	73.08	64.98	62.46	73.08	59.62
Panel C: 最优风险权重预测						
Prevailing Mean	3.88	55.13	4.35	5.11	54.49	6.22
Linear + GARCH	6.15	50.00	8.71	3.81	54.49	8.17
Linear + MIDAS	6.38	50.00	10.95	3.09	54.49	9.49
RF	5.78	57.69	7.39	7.15	61.54	9.04
Adaboost	6.08	55.77	8.08	6.62	58.33	9.42
BRT	6.51	58.97	8.49	6.73	62.18	8.32
BRT-MIDAS	7.34	58.97	9.79	6.05	62.18	11.05

注: 此表报告了在扩展窗口和滚动窗口下,不同模型对市场收益率、波动率与最优风险投资权重的样本外预测表现,包括各种模型的 R^2_{OOS}、方向预测精度 *Acc* 与稳健性调整后的 R^2_{OOS}($R^2_{OOS\text{-}robust}$). 表中数据均表示百分数. 其中 $R^2_{OOS\text{-}robust}$ 为去掉相对误差最大的 10% 样本后计算得到的 R^2_{OOS}.

其次,面板 B 中波动率预测的结果显示. 首先,传统方法 GARCH 和 MIDAS 表现较好,显示了波动率预测的方法已经较为成熟. 传统方法可以很好地挖掘波动率的高自相关性与异方差性,是目前预测波动率的主流方法. 其次,树类机器学习算法的表现与 MIDAS 不相上下. 比如,在扩展窗口下 MIDAS 方法的方向精度优于其他方法,但 BRT 与 BRT-MIDAS 方法 R^2_{OOS} 更高; 在滚动窗口下 MIDAS 方法的 R^2_{OOS} 高于其他方法,但方向精度不如 Adaboost 方法. 此外,$R^2_{oos\text{-}robust}$ 结果显示,在去掉 10% 极端样本后的,MIDAS 方法效果有较大下滑,BRT 方法与 BRT-MIDAS 方法的 $R^2_{oos\text{-}robust}$ 在扩展与滚动窗口下都高于 MIDAS,说明 BRT 类方法稳健性更强. 总体而言,在波动率预测上,尽管不能全面优于所有模型,BRT 模型仍可实现较好的预测效果.

再次,面板 C 显示 BRT 与 BRT-MIDAS 模型估计最优风险资产权重的能力整体优于其他模型,除了在扩展窗口下 Linear-MIDAS 模型的 $R^2_{oos\text{-}robust}$ 高于 BRT 方法外,其他方法的各项指标均不如 BRT 与 BRT-MIDAS 方法. 两种窗口模式下,BRT 方法的权重估计 R^2_{OOS} 为 6.51% 和 6.73%,方向精度达到 58.97% 和 62.18%,优于多数对比模型. BRT-MIDAS 方法的权重估计 R^2_{OOS} 为 7.34% 和 6.05%,方向精度达到 58.97% 和 62.18%,和 BRT 方法的效果整体不相上下,这显示了从 BRT 方法的视角研究该问题的重要性.

综上所述,在中国市场的实证发现,提升回归树 BRT 方法在市场风险溢价、波动率与最优风险资产权重的预测上,表现大概率优于现有的基准方法和线性方法,具有明显优势. 通过对比树型算法与线性算法的表现,可以发现提升回归树方法能够有效地挖掘预测变量与目标变量间的线性和非线性关系,实现更好的预测效果.

为了更直观地对比 BRT 方法和线性方法的预测效果,图 1 绘制了 BRT 模型的预测值与实际值的散点图. 从图中可以看出线性模型对真实预测值没有明显相关关系,其预测值对真实值的拟合线接近水平直线; 而 BRT 模型的预测值与真实值有正向的相关关系,直观地展示出 BRT 预测相对于线性模型的预测优势.

现实中经济变量对市场可能存在滞后的影响,因此本研究考虑了预测变量的滞后对市场风险溢价的预测能力. 具体来说,本研究将预测变量分别滞后 3 个月、6 个月、9 个月、12 个月,并衡量其预测能力. 表 4 对比了历史均值预测、线性模型预测和 BRT 模型预测的样本外拟合优度. 可以看出,随着滞后期的增加,模型的预测效果均呈现出下降的趋势. 线性模型在扩展窗口1 个月滞后期的情况下有微弱的正向 R^2,但在 3 个月滞后期的情况下变负. BRT 方法的预测能力可以持续到 3 个月以后,但在 6 个月滞后期情况下 R^2_{OOS} 变负. 横向对比来看,在各滞后期下,BRT 方法均优于历史均值预测和线性模型预测.

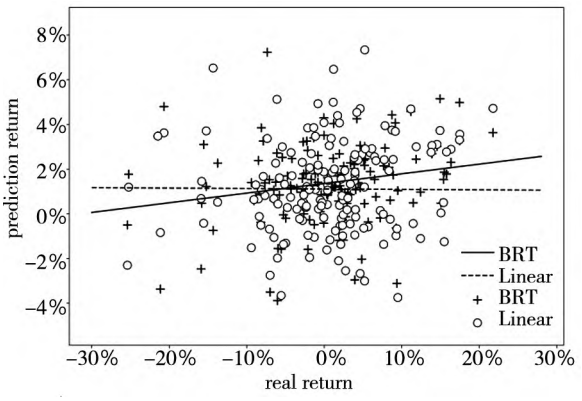


图 1　BRT 模型的预测值与实现值的散点图

Fig. 1 Scatter plot of BRT's return predictions and real returns

表 4　考虑预测变量不同滞后期的收益率预测拟合优度/%

Table 4 Return prediction R^2_{OOS} results of different lag periods of predictors/%

滞后期	扩展窗口			滚动窗口		
	Prevailing Mean	*Linear*	*BRT*	*Prevailing Mean*	*Linear*	*BRT*
1	−0.57	0.25	3.72	0.18	−5.10	3.96
3	−1.20	−1.22	1.64	−0.14	−3.15	2.36
6	−2.73	−1.10	−0.76	−1.80	−6.71	−1.05
9	−3.52	−15.34	−3.02	−3.12	−14.63	−1.90
12	−2.98	−19.10	−1.89	−2.89	−20.83	−1.25

3.2　模型预测的统计性检验

除了预测的结果,本研究进一步对上述算法的样本外预测结果进行了统计检验,进一步验证 BRT 方法是否能够有效地预测股票市场. 具体来说,本研究针对上述算法进行了 HM 检验[64]、CM 检验[65] 和 BH 检验[66]. 以收益率预测为例,式(8)、式(9)、式(10)列出了三个模型的检验方程,其中 $\hat{\mu}_{t+1}$ 和 r_{t+1} 分别表示 $t+1$ 期超额收益率的预测值和实际值,三个模型原假设均为 $\alpha_1 = 0$,备择假设为 α_1 显著大于 0. 其中,HM 检验验证模型的预测值能

否有效地预测方向(涨跌);CM 检验验证模型能否在实际值的绝对值更大时做出更准确的预测;BH 检验验证模型的预测值与实际值是否有显著线性关系.

$$I_{\{\hat{\mu}_{t+1}>0\}} = \alpha_0 + \alpha_1 I_{\{r_{t+1}>0\}} + \varepsilon_{t+1} \quad (8)$$

$$r_{t+1} = \alpha_0 + \alpha_1 I_{\{\hat{\mu}_{t+1}>0\}} + \varepsilon_{t+1} \quad (9)$$

$$r_{t+1} = \alpha_0 + \alpha_1 \hat{\mu}_{t+1} + \varepsilon_{t+1} \quad (10)$$

对最优风险资产权重的检验与超额收益预测检验相似;由于波动率无负方向,本研究将波动率相对其中位数的方向进行类似检验. 结合 HM 检验、CM 检验和 BM 检验可以更为全面地反映出模型对超额收益、波动率和最优风险资产权重的方向、大小的预测精度.

表 5 展示了三种统计检验的系数值(α_1)及其 t 值. 首先,面板 A 显示,针对超额收益率预测的三项检验中,在所有检验中 BRT 方法都在 5% 水平下显著;相对而言,线性模型则几乎都不显著. 其次,面板 B 显示,针对波动率预测的三项检验中,四种方法均在 1% 水平下显著,说明四种方法从预测的方向和幅度上都可以较好地预测波动率. 再次,面板 C 同样显示在最优风险资产权重的预测中,BRT 方法和 BRT-MIDAS 方法具有显著预测力,而其他方法则并不显著. 总的来说,只有 BRT 类的方法对超额收益率、波动率和最优风险资产权重的三项预测都能通过 BM、CM 和 BH 检验,在预测值的方向和幅度上均优于其他方法.

表 5 样本外预测的统计检验

Table 5 Statistical test results of out-of-sample prediction

统计检验	扩展窗口			滚动窗口		
	BM test	CM test	BH test	BM test	CM test	BH test
Panel A: 超额收益率预测						
BRT	0.112 3*	0.042 2**	1.016 9***	0.181 1***	0.048 1***	1.201 4**
	(1.88)	(2.41)	(2.66)	(2.95)	(2.89)	(2.53)
Linear Model	−0.099 7	0.002 4	0.390 9	0.053 8	0.012 6	0.459 8**
	(−1.40)	(0.16)	(1.23)	(0.68)	(0.92)	(1.96)
Panel B: 波动率预测						
BRT	0.461 5***	0.006 1***	1.243 6***	0.512 8***	0.006 3***	0.981 1***
	(6.46)	(5.61)	(10.13)	(7.41)	(5.93)	(8.78)
BRT-MIDAS	0.461 5***	0.007 0***	1.170 0***	0.461 5***	0.006 6***	0.920 5***
	(6.46)	(6.69)	(10.86)	(6.46)	(6.18)	(9.17)
GARCH	0.461 5***	0.006 7***	1.033 6***	0.435 9***	0.006 0***	0.699 4***
	(6.46)	(6.31)	(7.10)	(6.01)	(5.57)	(7.09)
MIDAS	0.512 8***	0.007 3***	0.785 5***	0.512 8***	0.007 3***	0.785 5***
	(7.41)	(7.15)	(10.22)	(7.41)	(7.15)	(10.22)
Panel C: 最优风险资产权重预测						
BRT	0.112 3*	1.787 7*	0.994 8*	0.181 1***	2.155 7**	1.455 7**
	(1.88)	(1.79)	(1.75)	(2.95)	(2.27)	(2.16)
BRT – MIDAS	0.112 3*	1.787 7*	0.880 0**	0.181 1***	2.155 7**	0.695 1*
	(1.88)	(1.79)	(2.00)	(2.95)	(2.27)	(1.82)
Linear + GARCH	−0.099 7	−0.174 9	0.408 4	0.053 8	0.600 2	0.412 4
	(−1.40)	(−0.21)	(1.03)	(0.68)	(0.78)	(1.55)
Linear + MIDAS	−0.099 7	−0.174 9	0.363 8*	0.053 8	0.600 2	0.334 2*
	(−1.40)	(−0.21)	(1.38)	(0.68)	(0.78)	(1.63)

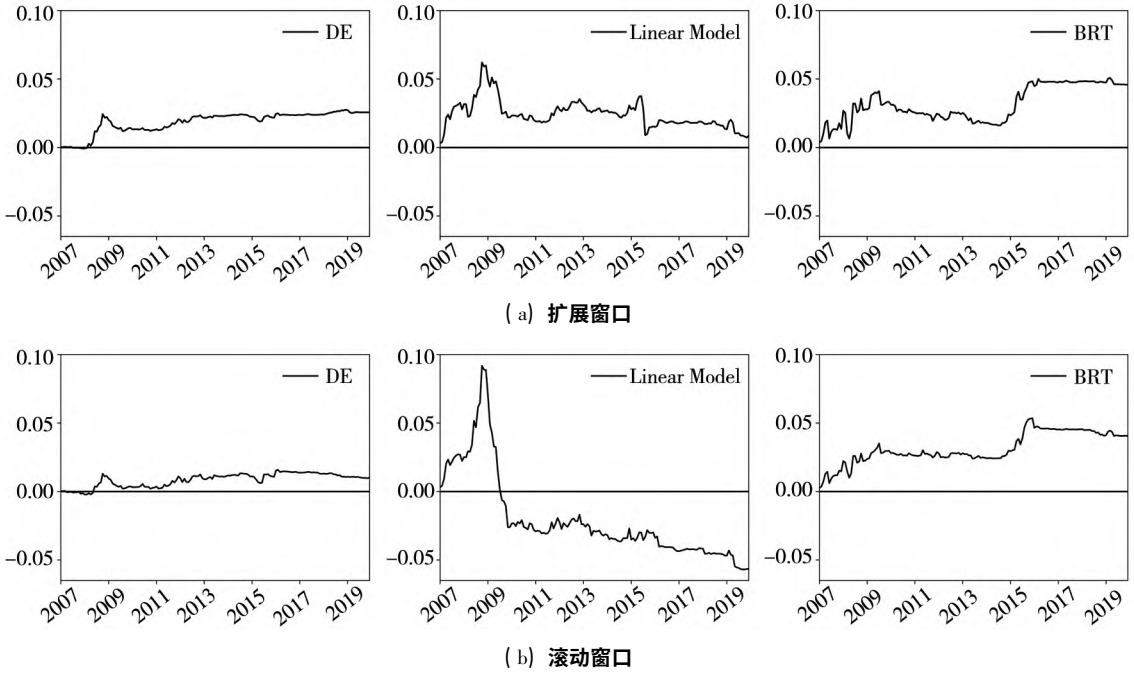
注:此表展示了扩展窗口和滚动窗口情况下三项检验的结果,即 HM 检验、CM 检验与 BH 检验对应的系数 α_1,括号中的数据表示相应检验的系数 t 值. 面板 A 展示了对超额收益率预测结果的检验,面板 B 展示了对市场波动率预测结果的检验,面板 C 展示了对最优风险资产配置权重预测结果的检验. *、**、*** 分别表示在 10%、5% 和 1% 水平下显著.

为了更直观地比较不同方法在时间序列上的预测效果,本研究绘制了基于历史均值估计的累积平方误差序列(*CSED*) [2]

$$CSED_t=\sum_{k=1}^{t}\left[\left(r_k-\hat{r}_{k,PreMean}\right)^2-\left(r_k-\hat{r}_{k,M}\right)^2\right] \quad (11)$$

其中 $CSED_t$ (cumulative squared error difference) 表示基于历史均值估计的累积平方误差, r_k 表示第 k 期实际市场收益率, $\hat{r}_{k,PreMean}$ 表示历史均值估计, $\hat{r}_{k,M}$ 表示其他方法得到的预测值. 当 $CSED_t$ 上升,说明该模型预测误差小于历史均值估计,反之则说明该模型的预测误差大于历史均值.

图 2 展示了扩展窗口和滚动窗口下表现最优的单变量(DE) 线性回归模型、多变量线性方法(linear model) 和 BRT 方法的 $CSED_t$ 序列图像,滚动窗口情况相似,不再赘述. 从 BRT 方法得到图线和其他图线的对比可以发现,在两种情形下,BRT 的图线都较为稳定地呈现上升趋势,而单变量线性方法(DE) 的图线更加平缓,最终得到为正的累积误差平方和,多变量线性模型(linear model) 的预测精度在 2008 年显著上升后在 2009 年迅速恶化,最终只在扩展窗口情形得到了为正的累积误差平方和,说明这些线性方法并不能得到比历史均值估计更小的预测误差,预测精度不如 BRT 方法.



(a) 扩展窗口

(b) 滚动窗口

图 2　基于历史均值的样本外累积平方误差

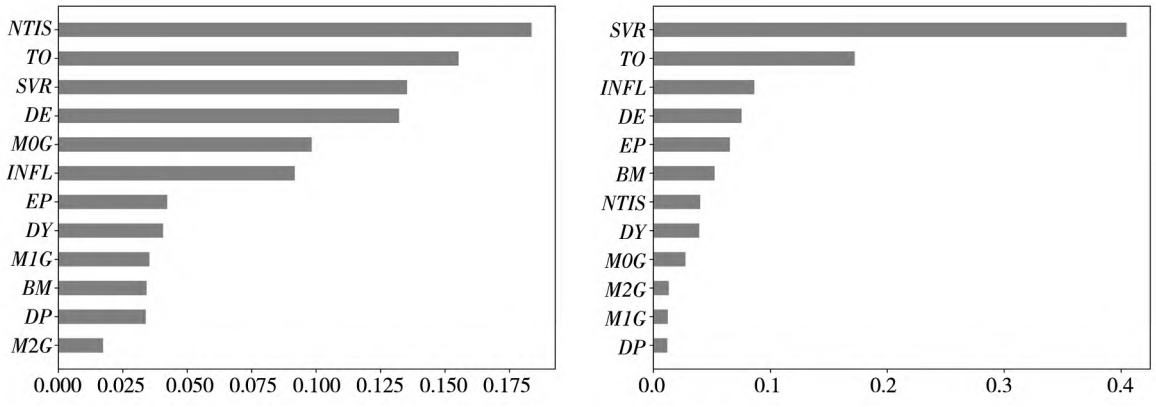
Fig. 2 Out-of-sample cumulative squared error differences relative to historical means

3.3　为什么 BRT 可以超越现有方法?

从上述分析可以判断出提升回归树方法相较于其他方法的显著优势,本节进一步分析 BRT 方法超越其他方法的原因. 传统的线性回归方法可以通过系数和显著性体现出不同变量对因变量影响的大小,而多数机器学习方法被认为是黑箱,无法观察变量对预测的影响. 提升回归树提供了变量重要性指标和部分依赖图对模型进行解释[6],其中变量重要性指标衡量了单个变量的变动对样本误差平方和的影响,而部分依赖图则能够直观地反映预测变量对目标变量的边际影响.

图 3 分别展示了 BRT 模型在收益率和波动率预测中的预测变量由高到低排列的变量相对重要性. 首先,从面板 A 可以看出,对于收益率预测来讲,重要性最高的三个变量分别是净权益增加值(*NTIS*) (相对重要性为 18.36%) 、换手率(*TO*) (相对重要性为 15.53%) 和股价方差(*SVR*) (相对重要性为 13.53%) . 这一结果与现有文献的解释一致. 比如,韩立岩和伍燕然[57]研究发现中国

市场新股发行数量与市场情绪具有较强相关性，高涨的市场情绪往往能吸引更多新股发行，形成“热销市场”(hot issue)，而市场情绪与市场收益率又存在着双向反馈的关系，从理论上印证了净权益增加值(*NTIS*)变量和市场收益率存在内在联系. 在市场微观结构和行为金融学研究中，换手率常被用来衡量股票的不确定性[68]，股票的不确定性越高，投资者持有风险资产的风险越高，也就会要求更高的风险补偿，因而换手率与预测股票收益间存在正相关关系. 徐浩峰和朱松[69]也发现高的换手率会使股票价格短期内上涨. 胡昌生和池阳春[70]研究发现市场波动率也与投资者情绪显著相关，从而也对市场收益率具有预测能力.



(a) 收益率预测 (b) 波动率预测

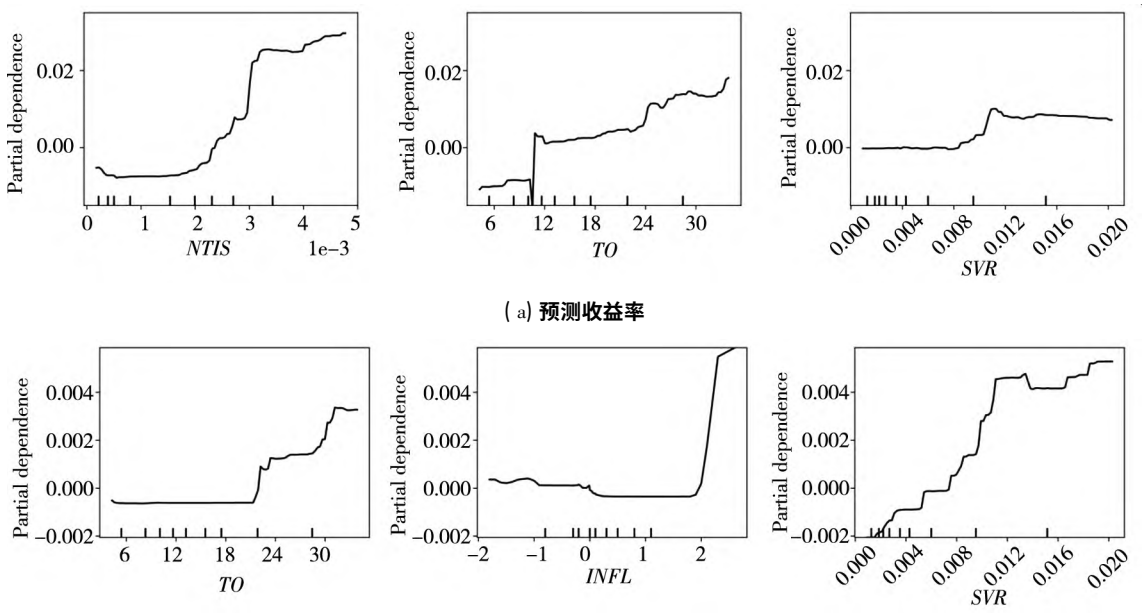
图 3 提升回归树模型预测收益率和波动率的变量相对重要性

Fig. 3 Relative feature importance of BRT prediction models for return and volatility

其次，从面板 B 可以看出，对波动率预测最重要的三个变量分别是股价方差(*SVR*)(相对重要性为 40.47%)、换手率(*TO*)(相对重要性为 17.21%)和通货膨胀率(*INFL*)(相对重要性为 8.63%)，其中股价方差的变量重要性远超其他变量. 杨科和陈浪南[71]等学者已经发现我国股市波动率具有较强的正自相关性，所以前一期股价方差对波动率具有较强预测能力符合预期. 换手率是度量市场情绪的重要变量，王美今和孙建军[72]研究发现市场波动与市场情绪关系密切，所以换手率变量可以一定程度上解释市场波动的变化. 刘凤根等[73]的研究表明通货膨胀等宏观经济变量对股市整体波动率具有显著影响.

图 4 展示了提升回归模型预测收益率与波动率时相对重要性最高三个变量的部分依赖图. 首先，图 4 中的六张图都很直观地显示了预测变量与目标变量间的非线性关系. 比如，从面板 A 中可以看出净权益增加值变量(*NTIS*)与市场收益率呈现出明显的非线性正相关关系，极差约达到 0.04，高于其他预测变量. 换手率变量(*TO*)部分依赖曲线中部呈现线性关系，头部和尾部则同样呈现非线性关系. 股价方差变量(*SVR*)的部分依赖图线则平缓很多，显示股价方差对模型的影响主要来自少数波动大的样本点. 从面板 B 可以看出，换手率(*TO*)、通货膨胀率(*INFL*)与股价方差变量(*SVR*)变量与市场波动率的关系均大致呈现非线性正相关关系，其中股价方差与波动率的关系最明显，在整个区间呈现阶梯上升状. 换手率变量部分依赖的前半部分呈水平状，在换手率大于 22 的区间与波动率呈现非线性正相关关系. 通货膨胀变量则只在尾部呈现快速上升，说明通货膨胀率对波动率的影响主要来自少数极大值.

综合来看，BRT 方法中最重要的变量与收益率和波动率之间均存在明显的非线性关系，这佐证了有效地挖掘变量间的非线性关系正是 BRT 区别于线性方法的关键所在.



(a) 预测收益率

(b) 预测波动率

图 4　提升回归树模型预测收益率和波动率的部分依赖图

Fig. 4 Partial dependence plot of BRT prediction of return and volatility

3.4　极端市场情况下的预测表现

从上述结果的分析中可以发现 BRT 方法对股市整体风险溢价具有显著的预测能力，那么该方法在极端市场情况下的预测能力如何？能否表现出对市场尾部风险防范的作用？为了比较不同模型在极端市场情况下的表现，本研究参考 Jacobsen 等[74]将市场实际收益率序列标准化为均值为 0，方差为 1 的序列，并按照标准化后的收益率将全部样本划分为不同时间区间，计算各种模型 R^2_{OOS}，结果如表 6 所示.

表 6　不同市场收益率情况下各方法 R^2_{OOS} /%

Table 6 R^2_{OOS} results of models under different market conditions/%

区间标准		扩展窗口			滚动窗口		
	Obs	*Prevailing Mean*	*Linear*	*BRT*	*Prevailing Mean*	*Linear*	*BRT*
all	156	−0.57	0.25	3.72	0.18	−5.10	3.96
Panel A: 极端市场时期							
$\lvert r\rvert \geqslant 0.5\%$	79	−1.10	2.46	4.92	−0.30	−0.14	4.28
$\lvert r\rvert \geqslant 1.0\%$	40	−1.06	2.50	5.40	−0.59	2.55	4.17
$\lvert r\rvert \geqslant 1.5\%$	23	−1.93	2.76	7.35	−1.46	5.16	4.83
Panel B: 极端上行时期							
$r \geqslant 0.5\%$	40	18.50	24.21	30.59	13.91	−4.00	21.72
$r \geqslant 1.0\%$	21	15.95	23.00	28.47	12.29	1.88	20.69
$r \geqslant 1.5\%$	12	14.13	27.65	31.26	11.29	8.80	21.17
Panel C: 极端下行时期							
$r \leqslant -0.5\%$	39	−19.54	−18.01	−19.24	−13.66	3.49	−12.13
$r \leqslant -1.0\%$	19	−16.30	−15.88	−15.28	−12.13	3.15	−10.63
$r \leqslant -1.5\%$	11	−14.50	−16.71	−11.36	−11.44	2.31	−7.96

注：此表展示了在不同市场收益率水平下各模型的预测 R^2_{OOS}.

面板 A 显示在不区分方向的极端市场现下，BRT 的预测效果总体上优于其他方法，其中扩展窗口模式下，R^2_{OOS} 比线性方法高出 2% 以上. 面板 B 和面板 C 分别展示了 BRT 在极端上行和极端下行时期的样本外预测能力. 其中，面板 B 显示 BRT 在股市情况上行的时期依然可以做出较好的样本外预测，R^2_{OOS} 分别达到了 28% 和 20% 以上. 面板 C 显示了在市场极端下行时期，BRT 模型的表现相对较弱. 总体而言，BRT 模型具备一定的尾部风险防范能力，尤其是在市场极端上行时期.

4 市场预测的经济价值

预测准确率的微小提升可以转化为盈利的投资策略[63]. 上一节的分析已经发现 BRT 可以大幅度提升股票市场预测的样本外拟合优度，一个自然的问题是，预测的提升究竟能够带来多大的经济价值?

本研究按上文所述的最优风险资产权重估计方法得到风险资产权重，并以此构建投资组合. 风险资产是指按流通市值加权的市场组合，风险资产超额收益率记为 r_t，无风险收益率记为 $r_{f,t}$，最优投资组合由权重为 $\hat{w}_t$ 的风险资产与权重为 $(1-\hat{w}_t)$ 的无风险资产构成，投资组合实现的收益率表示为

$$r_{p,t} = r_{f,t} + \hat{w}_t r_t \tag{12}$$

为了全面比较预测模型带来的经济价值，本研究对比了两种投资策略. 第一种策略依据超额收益预测，当预测超额收益率为正是则全部买进市场组合($\hat{w}_t = 1$)，否则仅获得无风险收益率($\hat{w}_t = 0$). 第二种策略以最优风险资产权重估计值为依据，由于中国股票市场中，投资者具有借贷与卖空的限制，即风险投资权重被限制在 $[0,1]$ 区间内，当 $\hat{w}_t > 1$ 时取 $\hat{w}_t = 1$，当 $\hat{w}_t < 0$ 时取 $\hat{w}_t = 0$. 主要通过比较超额收益率均值、夏普比率和 *CAPM* α 判断预测模型是否能够为投资者带来显著的经济效益.

表 7 的展示了不同预测模型对应策略的表现，整体上看，BRT 方法优于其他模型. 首先，本研究发现全部买入持有市场投资组合得到的月度收益率均值仅为 0.67%，只略高于面板 A 中的线性方法表现，低于其他策略的收益率，夏普比率为 0.057 0，低于所有模型策略. 其次，面板 A 表明在只依据收益率预测的策略下，BRT 方法获得了显著高于线性方法的收益. BRT 方法夏普比率达到 0.133 0 和 0.155 3(年化夏普率为 0.460 7 和 0.537 9)，是线性模型夏普比率的两倍左右，而且得到了 5% 水平上显著的 *CAPM* α，年化 α 达到 6.60% 和 8.52%. 最后，面板 B 表明在中国市场依据最优风险资产权重估计值构建投资组合时，BRT 与 BRT-MIDAS 方法相差不大，表现均优于其他线性方法，在扩展窗口下 BRT 方法的收益率均值(月度 1.06%，年化 12.72%)与夏普比率(月度 0.129 1，年化 0.447 2)略高于 BRT-MIDAS 方法，但 *CAPM* α 低于 BRT-MIDAS 的 0.53%(年化 6.36%)，在滚动窗口下 BRT-MIDAS 方法的各个指标均高于其他方法，收益均值达到 1.07%(年化 12.84%)，夏普比达到 0.140 4(年化 0.486 4)，*CAPM* α 达到 0.63%(年化 7.56). 在面板 A 和面板 B 中，BRT 方法的 *Market* β 均高于其他对比方法，且 *Market* β 值均小于 1，说明 BRT 方法的市场敏感性总体上强于对比方法.

为了直观地体现 BRT 等方法的经济效益，并与线性模型进行对比，图 5 展示了不同模型的累积收益率表现. 从图 5 中可以看出 BRT 等非线性机器学习的累积收益率比线性模型高，从 2007 年—2019 年，线性模型的累积收益率约为 50%，而 BRT 模型的累积收益率超过 200%，RF 模型和 Adaboost 模型的表现也都超过了 100%. 说明非线性机器学习方法对应的资产配置策略可以为投资者带来可观的收益率.

表 7　投资组合收益表现

Table 7 Portfolio return performance

变量	扩展窗口				滚动窗口			
	Mean	*Sharpe*	*CAPM* α	*Market* β	*Mean*	*Sharpe*	*CAPM* α	*Market* β
Panel A：依据超额收益预测								
BRT	1.18%*	0.133 0	0.55%**	0.79***	1.31%*	0.155 3	0.71%**	0.75***
	(1.70)		(2.06)	(9.43)	(1.89)		(2.49)	(9.03)
Linear Model	0.66%	0.065 4	0.08%	0.71***	0.74%	0.084 9	0.23%	0.58***
	(1.03)		(0.27)	(6.84)	(1.20)		(0.69)	(5.36)
Panel B：最优风险资产权重								
BRT	1.06%*	0.129 1	0.49%*	0.62***	1.05%*	0.134 6	0.51%*	0.54***
	(1.64)		(1.78)	(8.04)	(1.66)		(1.79)	(7.07)
BRT – MIDAS	1.02%	0.109 7	0.53%*	0.55***	1.07%	0.140 4	0.63%**	0.44***
	(1.36)		(1.78)	(7.35)	(1.62)		(2.20)	(6.12)
Linear + GARCH	0.86%	0.111 7	0.36%	0.50***	0.74%	0.111 7	0.32%	0.34***
	(1.53)		(1.22)	(5.71)	(1.50)		(1.15)	(4.29)
Linear + MIDAS	0.90%	0.119 9	0.40%	0.44***	0.84%*	0.133 4	0.43%	0.32***
	(1.62)		(1.38)	(5.74)	(1.64)		(1.52)	(4.34)
Panel C：被动策略								
Market	0.67%	0.057 0						
	(0.83)							

注：此表报告了扩展窗口和滚动窗口不同投资限制条件下各模型实际收益率的月度平均超额收益、月度夏普比率、*CAPM* α 值指标和 *Market* β. 括号中数据表示平均超额收益率和 *CAPM* α 值的 Newey-West 调整 t 值. 面板 A 表示仅依据超额收益率是否为正设置风险资产权重为 1 或 0 的策略，面板 B 表示将风险资产权重限制在 $[0,1]$ 区间内的投资组合策略，面板 C 表示全部买入持有风险资产不变的策略. *、**、*** 分别表示在 10%、5% 和 1% 水平下显著.

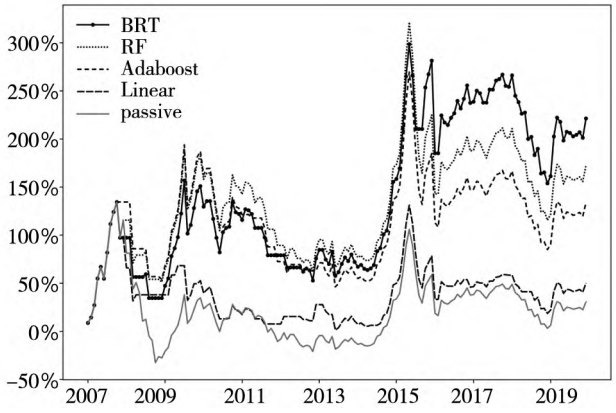


图 5　不同策略累计收益率变化图

Fig. 5 Line chart of cumulative returns of different strategies

5　结束语

本文基于 1996 年 1 月—2019 年 12 月的 12 种变量数据构建了提升回归树模型用于预测中国资本市场超额收益率，并从投资者角度出发，对市场波动率、风险资产最优权重进行了预测，然后将提升回归树模型与线性模型等其他方法进行了系统对比，此外，研究了提升回归树模型中的变量重要性，依据中国股市特点得到解释. 实证结果表明提升回归树方法对市场收益率与波动率均有显著的样本外预测能力，表现显著强于其他模型. 净权益增加值和股价方差分别是提升回归树模型中对股市超额收益和波动率预测最重要的变量，这与我国股市的市场结构有关. 从投资者的角度看，提升回归树方法对应的策略能为投资者带来更高的收益率与实际效用. 相比于线性方法，机器学习方法可以更好地捕捉到公司层面变量与宏观经济变量的

信息，避免样本外预测的模型不确定性和参数不稳定性的问题，在模型涉及变量较多的情况下避免了过拟合问题，可以稳健有效地对我国资本市场风险溢价做出预测.

本研究对机器学习方法在投资领域的研究具有重要意义. 相比于传统线性方法，机器学习方法的优势主要体现在两点: 第一，机器学习方法可以灵活处理变量较多、关系较为复杂的问题，挖掘出数据与变量之间线性或非线性的关系，提取信息的能力比线性方法更强. 第二，机器学习方法在经济金融领域的预测问题上相比线性模型更有优势. 线性方法往往在过拟合与丢失信息之间两难，而机器学习方法的内在机制决定它可以在充分挖掘信息的同时避免过拟合问题，尤其适用于样本外预测相关问题的研究. 未来学者们可从如下方面对这一问题继续研究，例如运用机器学习方法从大数据中挖掘新的预测变量，探索更优化的预测算法等.

参 考 文 献:

[1] 蒋志强，田婧雯，周炜星. 中国股票市场收益率的可预测性研究 [J]. 管理科学学报，2019，22(4)：92 – 109.
Jiang Zhiqiang，Tian Jingwen，Zhou Weixing. Return predictability in the Chinese stock markets [J]. Journal of Management Sciences in China，2019，22(4)：92 – 109. (in Chinese)

[2] Welch I，Goyal A. A comprehensive look at the empirical performance of equity premium prediction [J]. Review of Financial Studies，2008，21(4)：1455 – 1508.

[3] Zhang H，He Q，Jacobsen B，et al. Forecasting stock returns with model uncertainty and parameter instability [J]. Journal of Applied Econometrics，2020，35(5)：629 – 644.

[4] Dong X，Li Y，Rapach D，et al. Anomalies and the expected market return [J]. Journal of Finance，2021，77(1)：639 – 681.

[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.
Su Zhi，Lu Man，Li Dexuan. Deep learning in financial empirical applications: Dynamics，contributions and prospects [J]. Journal of Financial Research，2017，(5)：111 – 126. (in Chinese)

[6] Hastie T，Tibshirani R，Friedman J. The Elements of Statistical Learning: Data Mining，Inference，and Prediction [M]. Berlin: Springer，2009.

[7] 李 斌，邵新月，李玥阳. 机器学习驱动的基本面量化投资研究 [J]. 中国工业经济，2019，(8)：61 – 79.
Li Bin，Shao Xinyue，Li Yueyang. Research onmachine learning driven quantamental investing [J]. China Industrial Economics，2019，(8)：61 – 79. (in Chinese)

[8] Gu S，Kelly B，Xiu D. Empirical asset pricing via machine learning [J]. Review of Financial Studies，2020，33(5)：2223 – 2273.

[9] Bianchi D，Büchner M，Tamoni A. Bond risk premiums with machine learning [J]. Review of Financial Studies，2021，34(2)：1046 – 1089.

[10] Schmidhuber J U R. Deep learning in neural networks: An overview [J]. Neural Networks，2015，(61)：85 – 117.

[11] Du M，Liu N，Hu X. Techniques for interpretable machine learning [J]. Communications of the ACM，2019，63(1)：68 – 77.

[12] 姜富伟，涂 俊，周国富，等. 中国股票市场可预测性的实证研究 [J]. 金融研究，2011，(9)：107 – 121.
Jiang Fuwei，Tu Jun，Zhou Guofu，et al. An empirical study on the predictability of China's stock market [J]. Journal of Financial Research，2011，(9)：107 – 121. (in Chinese)

[13] 张春玲，姜富伟，唐国豪. 资本市场收益可预测性研究进展 [J]. 经济学动态，2019，(2)：133 – 148.
Zhang Chunling，Jiang Fuwei，Tang Guohao. The progress of research on return predictability in the capital market [J]. Economic Perspectives，2019，(2)：133 – 148. (in Chinese)

[14] 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.

Huang Naijing, Yu Mingzhe. The progress of research on the influence of machine learning on economic research [J]. Economic Perspectives, 2018, (7): 115 – 129. (in Chinese)

[15] Cooper I, Priestley R. Time-varying risk premiums and the output gap [J]. Review of Financial Studies, 2009, 22(7): 2801 – 2833.

[16] He W, Hu M R. Aggregate earnings and market returns: International evidence [J]. Journal of Financial and Quantitative Analysis, 2014, 49(4): 879 – 901.

[17] Rapach D E, Ringgenberg M C, Zhou G. Short interest and aggregate stock returns [J]. Journal of Financial Economics, 2016, 121(1): 46 – 65.

[18] Jacobsen B, Marshall B R, Visaltanachoti N. Stock market predictability and industrial metal returns [J]. Management Science, 2019, 65(7): 3026 – 3042.

[19] 朱小能, 袁经发. 去伪存真: 油价趋势与股票市场——来自"一带一路"35 国的经验证据 [J]. 金融研究, 2019, 471(9): 131 – 150.
Zhu Xiaoneng, Yuan Jingfa. Oil price trends and the stock market: Empirical evidence from 35 countries along "the Belt and Road" [J]. 金融研究, 2019, 471(9): 131 – 150. (in Chinese)

[20] 杨晓兰, 王伟超, 高　媚. 股市政策对股票市场的影响——基于投资者社会互动的视角 [J]. 管理科学学报, 2020, 23(1): 15 – 32.
Yang Xiaolan, Wang Weichao, Gao Mei. The impact of stock market policies on stock market: From the perspective of investor social interaction [J]. Journal of Management Sciences in China, 2020, 23(1): 15 – 32. (in Chinese)

[21] 姜富伟, 胡逸驰, 黄　楠. 央行货币政策报告文本信息、宏观经济与股票市场 [J]. 金融研究, 2021, 492(6): 95 – 113.
Jiang Fuwei, Hu Yichi, Huang Nan. Textual information of central bank monetary policy report, macroeconomy and stock market performance [J]. Journal of Financial Research, 2021, 492(6): 95 – 113. (in Chinese)

[22] 王若昕, 马　锋. 日内收益率预测: 基于日内跳跃和动量研究 [J]. 系统工程理论与实践, 2021, 41(8): 2004 – 2014.
Wang Ruoxin, Ma Feng. Intraday return predictability: Based on intraday jumps and momentum [J]. Systems Engineering: Theory & Practice, 2021, 41(8): 2004 – 2014. (in Chinese)

[23] Huang D, Li J, Wang L. Are disagreements agreeable? Evidence from information aggregation [J]. Journal of Financial Economics, 2021, 141(1): 83 – 101.

[24] 张　琳, 张　军, 王　擎. 宏观经济信息发布对股票市场收益率及其波动的影响 [J]. 系统工程理论与实践, 2020, 40(6): 1439 – 1451.
Zhang Lin, Zhang Jun, Wang Qing. The effect of macroeconomic announcements on stock market return and volatility [J]. Systems Engineering: Theory & Practice, 2020, 40(6): 1439 – 1451. (in Chinese)

[25] Goh J C, Jiang F, Tu J, et al. Can US economic variables predict the Chinese stock market? [J]. Pacific-Basin Finance Journal, 2013, 22(1): 69 – 87.

[26] Chen J, Jiang F, Tong G. Economic policy uncertainty in China and stock market expected returns [J]. Accounting & Finance, 2017, 57(5): 1265 – 1286.

[27] Huang D, Jiang F, Tu J. Investor sentiment aligned: A powerful predictor of stock returns [J]. Review of Financial Studies, 2015, 28(3): 791 – 837.

[28] 尹海员, 吴兴颖. 投资者高频情绪对股票日内收益率的预测作用 [J]. 中国工业经济, 2019, 377(8): 80 – 98.
Yin Haiyuan, Wu Xingying. Predictive effect of high-frequency investor sentiment on the intraday stocks return [J]. China Industrial Economics, 2019, 377(8): 80 – 98. (in Chinese)

[29] Jiang F, Lee J, Martin X, et al. Manager sentiment and stock returns [J]. Journal of Financial Economics, 2019, 132(1): 126 – 149.

[30] 钱　宇, 李子饶, 李　强, 等. 在线社区支持倾向对股市收益和波动的影响 [J]. 管理科学学报, 2020, 23(2): 141 – 155.

Qian Yu, Li Zirao, Li Qiang, et al. Impact of online community support tendencies on returns and volatility in Chinese stock market [J]. Journal of Management Sciences in China, 2020, 23(2): 141 – 155. (in Chinese)

[31] Atilgan Y, Bali T G, Demirtas K O. Implied volatility spreads and expected market returns [J]. Journal of Business & Economic Statistics, 2015, 33(1): 87 – 101.

[32] Kelly B, Jiang H. Tail risk and asset prices [J]. Review of Financial Studies, 2014, 27(10): 2841 – 2871.

[33] 陈 坚, 张轶凡, 洪集民. 期权隐含尾部风险及其对股票收益率的预测 [J]. 管理科学学报, 2019, 22(10): 72 – 81.

Chen Jian, Zhang Yifan, Hong Jimin. Option implied tail risk and predictability of stock return [J]. Journal of Management Sciences in China, 2019, 22(10): 72 – 81. (in Chinese)

[34] Pyun S. Variance risk in aggregate stock returns and time-varying return predictability [J]. Journal of Financial Economics, 2019, 132(1): 150 – 174.

[35] Pettenuzzo D, Timmermann A, Valkanov R. Forecasting stock returns under economic constraints [J]. Journal of Financial Economics, 2014, 114(3): 517 – 553.

[36] Rapach D, Zhou G. Forecasting stock returns [J]. Handbook of Economic Forecasting, 2013, 2(1): 328 – 383.

[37] Ferreira M A, Santa-Clara P. Forecasting stock market returns: The sum of the parts is more than the whole [J]. Journal of Financial Economics, 2011, 100(3): 514 – 537.

[38] Neely C J, Rapach D, Tu J, et al. Forecasting the equity risk premium: The role of technical indicators [J]. Management Science, 2014, 60(7): 1772 – 1791.

[39] Baillie R T, Kapetanios G. Nonlinear models for strongly dependent processes with financial applications [J]. Journal of Econometrics, 2008, 147(1): 60 – 71.

[40] Cochrane J H. Presidential address: Discount rates [J]. The Journal of Finance, 2011, 66(4): 1047 – 1108.

[41] Tobek O, Hronec M. Does it pay to follow anomalies research? Machine learning approach with international evidence [J]. Journal of Financial Markets, 2020, 1(1): 100588.

[42] Freyberger J, Neuhierl A, Weber M, et al. Dissecting characteristics nonparametrically [J]. Review of Financial Studies, 2020, 33(5): 2326 – 2377.

[43] Krauss C, Do X A, Huck N. 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.

[44] Leippold M, Wang Q, Zhou W. Machine learning in the Chinese stock market [J]. Journal of Financial Economics, 2021, 145(2): 64 – 82.

[45] 姜富伟, 马 甜, 张宏伟. 高风险低收益? 基于机器学习的动态 CAPM 模型解释 [J]. 管理科学学报, 2021, 24(1): 109 – 126.

Jiang Fuwei, Ma Tian, Zhang Hongwei. High risk low return? Explanation from machine learning based conditional CAPM model [J]. Journal of Management Sciences in China, 2021, 24(1): 109 – 126. (in Chinese)

[46] 姜富伟, 薛 浩, 周 明. 大数据提升了多因子模型定价能力吗? ——基于机器学习方法对我国 A 股市场的探究 [J]. 系统工程理论与实践, 2022, 42(8): 1 – 16.

Jiang Fuwei, Xue Hao, Zhou Ming. Does big data improve multi-factor pricing models? Exploration of China's A-share market [J]. Systems Engineering: Theory & Practice, 2022, 42(8): 1 – 16. (in Chinese)

[47] 夏 婷, 闻岳春. 经济不确定性是股市波动的因子吗? ——基于 garch-midas 模型的分析 [J]. 中国管理科学, 2018, 26(12): 1 – 11.

Xia Ting, Wen Yuechun. Dose economic uncertainty matter for stock market volatility? An analysis based on GARCH-MIDAS [J]. Chinese Journal of Management Science, 2018, 26(12): 1 – 11. (in Chinese)

[48] Paye B S. 'Deja vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables [J]. Journal of Financial Economics, 2012, 106(3): 527 – 546.

[49] Engle R F, Ghysels E, Sohn B. Stock market volatility and macroeconomic fundamentals [J]. Review of Economics and Sta-

tistics, 2013, 95(3): 776 – 797.

[50] Adrian T, Rosenberg J. Stock returns and volatility: Pricing the short-run and long-run components of market risk [J]. Journal of Finance, 2008, 63(6): 2997 – 3030.

[51] Kim Y, Nelson C. Pricing stock market volatility: Does it matter whether the volatility is related to the business cycle? [J]. Journal of Financial Econometrics, 2014, 12(2): 307 – 328.

[52] 董直庆, 王林辉. 我国通货膨胀和证券市场周期波动关系——基于小波变换频带分析方法的实证检验 [J]. 中国工业经济, 2008, (11): 35 – 44.
Dong Zhiqing, Wang Linhui. Cycle fluctuation relationship of inflation and stock market: An empirical study based oil frequency band approach of wavelet [J]. China Industrial Economics, 2008, (11): 35 – 44. (in Chinese)

[53] 陈其安, 雷小燕. 货币政策、投资者情绪与中国股票市场波动性: 理论与实证 [J]. 中国管理科学, 2017, 25(11): 1 – 11.
Chen Qi'an, Lei Xiaoyan. Monetary policy, investor sentiment and volatility of chinese stock market: Theory and evidence [J]. Chinese Journal of Management Science, 2017, 25(11): 1 – 11. (in Chinese)

[54] Pan Z, Wang Y, Liu L, et al. Improving volatility prediction and option valuation using VIX information: A volatility spillover GARCH model [J]. Journal of Futures Markets, 2019, 39(6): 744 – 776.

[55] 陈　强, 叶阿忠. 股市收益、收益波动与中国城镇居民消费行为 [J]. 经济学(季刊), 2009, 8(3): 995 – 1012.
Chen Qiang, Ye Azhong. Stock market returns, volatility and urban residential consumption behavior in China [J]. China Economic Quarterly, 2009, 8(3): 995 – 1012. (in Chinese)

[56] Fang T, Lee T, Su Z. Predicting the long-term stock market volatility: A GARCH-MIDAS model with variable selection [J]. Journal of Empirical Finance, 2020, 58: 36 – 49.

[57] 郑挺国, 尚玉皇. 基于宏观基本面的股市波动度量与预测 [J]. 世界经济, 2014, 37(12): 118 – 139.
Zheng Tingguo, Shang Yuhuang. Measurement and prediction of stock market volatility basedon macro fundamentals [J]. The Journal of World Economy, 2014, 37(12): 118 – 139. (in Chinese)

[58] Asgharian H, Hou A J, Javed F. The importance of the macroeconomic variables in forecasting stock return variance: A GARCH-MIDAS approach [J]. Journal of Forecasting, 2013, 32(7): 600 – 612.

[59] Mittnik S, Robinzonov N, Spindler M. Stock market volatility: Identifying major drivers and the nature of their impact [J]. Journal of Banking and Finance, 2015, 58: 1 – 14.

[60] Christiansen C, Schmeling M, Schrimpf A. A comprehensive look at financial volatility prediction by economic variables [J]. Journal of Applied Econometrics, 2012, 27(6): 956 – 977.

[61] Naifar N, Al Dohaiman M S. Nonlinear analysis among crude oil prices, stock markets' return and macroeconomic variables [J]. International Review of Economics & Finance, 2013, 27(1): 416 – 431.

[62] Adämmer P, Schüssler R A. Forecasting the equity premium: Mind the news [J]. Review of Finance, 2020, 24(6): 1313 – 1355.

[63] Campbell J, Thompson S P. Predicting excess stock returns out of sample: Can anything beat the historical average? [J]. Review of Financial Studies, 2008, 21(1): 1509 – 1531.

[64] Henriksson R D, Merton R C. On market timing and investment performance II. Statistical procedures for evaluating forecasting skills [J]. Journal of Business, 1981, 54(4): 513.

[65] Cumby R E, Modest D M. Testing for market timing ability: A framework for forecast evaluation [J]. Journal of Financial Economics, 1987, 19(1): 169 – 189.

[66] Bossaerts P, Hillion P. Implementing statistical criteria to select return forecasting models: What do we learn? [J]. Review of Financial Studies, 1999, 12(2): 405 – 428.

[67] 韩立岩, 伍燕然. 投资者情绪与 IPOs 之谜——抑价或者溢价 [J]. 管理世界, 2007, (3): 51 – 61.
Han Liyan, Wu Yanran. Investors' sentiment and the puzzle about IPO-underpricing or overpricing? [J]. Journal of Management World, 2007, (3): 51 – 61. (in Chinese)

[68] Jiang G, Lee C M C, Zhang Y. Information uncertainty and expected returns [J]. Review of Accounting Studies, 2005, 10(2-3): 185-221.

[69] 徐浩峰，朱 松. 机构投资者与股市泡沫的形成 [J]. 中国管理科学，2012，20(4)：18-26.
Xu Haofeng, Zhu Song. Does the trade from institutional investors induce stock bubble [J]. Chinese Journal of Management Science, 2012, 20(4): 18-26. (in Chinese)

[70] 胡昌生，池阳春. 投资者情绪、资产估值与股票市场波动 [J]. 金融研究，2013，(10)：181-193.
Hu Changsheng, Chi Yangchun. Investor sentiment, asset valuation, and the volatility of stock market [J]. Journal of Financial Research, 2013, (10): 181-193. (in Chinese)

[71] 杨 科，陈浪南. 中国股市高频波动率跳跃的特征分析 [J]. 系统工程学报，2012，27(4)：492-497.
Yang Ke, Chen Langnan. Jump dynamics of high-frequency volatility in Chinese stock markets [J]. Journal of Systems Engineering, 2012, 27(4): 492-497. (in Chinese)

[72] 王美今，孙建军. 中国股市收益、收益波动与投资者情绪 [J]. 经济研究，2004，(10)：75-83.
Wang Meijin, Sun Jianjun. Stock market returns, volatility and the role of investor sentiment in China [J]. Economic Research Journal, 2004, (10): 75-83. (in Chinese)

[73] 刘凤根，吴军传，杨希特，等. 基于混频数据模型的宏观经济对股票市场波动的长期动态影响研究 [J]. 中国管理科学，2020，28(10)：65-76.
Liu Fenggen, Wu Junchuan, Yang Xite, et al. Long-run dynamic effect of macro-economy on stock market volatility based on mixed frequency data model [J]. Chinese Journal of Management Science, 2020, 28(10): 65-76. (in Chinese)

[74] Jacobsen B, Jiang F, Zhang H. Ensemble Machine Learning and Stock Return Predictability [C]. International Workshop on Financial Markets and Nonlinear Dynamics, 2019.

Return predictability in the Chinese stock markets: A machine learning perspective

LI Bin[1,2], *LONG Zhen*[1,2*]

1. School of Economics and Management, Wuhan University, Wuhan 430072, China;
2. The Centre of Finance Research, Wuhan University, Wuhan 430072, China

Abstract: Equity premium prediction is one crucial research problem in finance. The traditional linear model suffers from model misspecification and parameter instability, weakening its out-of-sample prediction performance. This paper re-examines equity premium predictability in the Chinese stock markets. Based on the data from January 1996 to December 2019, we adopt Boost Regression Trees (BRT) to predict market return and volatility. Empirical results show that BRT outperforms traditional methods in predicting return, volatility, and optimal portfolio allocation. The three most important variables in the prediction model include Net Issuance, Turnover, and Stock Variance. The BRT's predictability is driven by its ability to capture nonlinearity among variables. Finally, the optimal portfolios constructed by BRT result in higher returns and utility to investors. Our work contributes to the literature by leveraging the nonlinear machine learning method for the equity premium study, thus providing new research perspective.

Key words: equity risk premium; boosting regression tree; out-of-sample prediction; machine learning