

基于机器学习的富时A50期指价格预测及其与A股的定价关联研究

左邵成

贵州大学经济学院, 贵州 贵阳

收稿日期: 2025年4月11日; 录用日期: 2025年4月24日; 发布日期: 2025年5月31日

摘要

本文将研究视角聚焦于富时中国A50股指期货, 深度挖掘其价格预测的有效方法, 并对其是否具备A股定价因子的属性展开全面探究。研究选取2014年10月31日至2024年12月13日的相关数据, 在经过严谨的数据预处理流程后, 运用随机森林、长短期记忆网络(LSTM)以及梯度提升树三种模型对其价格进行预测, 结果显示梯度提升树模型在预测精度上表现最为突出。在定价因子检验环节, 以沪深300成分股收盘价作为A股的代理变量, 通过Fama-Macbeth检验以及滚动回归稳健性检验进行深入分析, 最终证实富时中国A50期指对A股市场定价存在显著的正向影响, 这一研究成果为投资者的决策制定以及金融领域的学术研究提供了极具价值的参考依据。

关键词

机器学习, 富时A50期指, Fama-Macbeth

Research on Price Forecasting of FTSE China A50 Index Futures Based on Machine Learning and Its Pricing Correlation with A-Shares

Shaocheng Zuo

School of Economics, Guizhou University, Guiyang Guizhou

Received: Apr. 11th, 2025; accepted: Apr. 24th, 2025; published: May 31st, 2025

Abstract

This article will focus on the FTSE China A50 stock index futures, deeply explore its effective

文章引用: 左邵成. 基于机器学习的富时 A50 期指价格预测及其与 A 股的定价关联研究[J]. 电子商务评论, 2025, 14(5): 3137-3151. DOI: 10.12677/ecl.2025.1451627

methods for price prediction, and comprehensively explore whether it has the attributes of A-share pricing factors. The study selected relevant data from October 31, 2014 to December 13, 2024. After a rigorous data preprocessing process, three models including random forest, long short-term memory network (LSTM), and gradient boosting tree were used to predict their prices. The results showed that the gradient boosting tree model performed the most outstandingly in terms of prediction accuracy. In the pricing factor testing stage, the closing prices of the Shanghai and Shenzhen 300 constituent stocks were used as proxy variables for A-shares. Through Fama Macbeth test and rolling regression robustness test, in-depth analysis was conducted, and it was finally confirmed that the FTSE China A50 futures index has a significant positive impact on the pricing of the A-share market. This research result provides valuable reference for investors' decision-making and academic research in the financial field.

Keywords

Machine Learning, FTSE A50 Futures Index, Fama-Macbeth

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

富时 A50 股指期货(东方财富代码:CN00Y)以富时中国 A50 指数为标的,在新加坡交易所挂牌交易。富时中国 A50 指数由富时罗素指数公司精心筛选 A 股市场中市值居前 50 位的公司构成,其成分股总市值在 A 股总市值中占比约达三分之一,堪称 A 股市场的关键代表指数之一。在交易时间上, A 股每个交易日自上午 9:00 准时开盘,至下午 15:00 收盘,为 T + 1 交易模式。而富时 A50 股指期货则拥有长达 20 个小时的交易时长,从北京时间上午 9:00 持续至次日凌晨 5:15,其间仅在 16:30 至 17:00 有半小时休市时段,为 T + 0 交易模式。鉴于其交易模式的灵活性和交易时长都超过 A 股市场,当 A 股结束当日交易后,富时 A50 股指期货仍活跃于市场之中。一旦晚间或周末突发重大新闻或事件, A50 股指期货往往能够率先做出反应,其价格波动与交易态势随即成为国际市场对 A 股市场整体预期及投资情绪的直观映射。

对于境内投资者而言,富时中国 A50 股指期货的走势具备重要的参考价值,可作为预判次日 A 股涨跌的关键指标之一。因为 A50 期指可以更及时的对突发事件做出反应且标的资产也是 A 股龙头企业,所以 A 股的投资者据此能够提前谋划布局或灵活调整投资策略,有效规避潜在风险并精准捕捉盈利契机。从现实层面考量,若能深入探究富时中国 A50 股指期货的可预测性及其预测成效,无疑可为投资者的决策过程提供精准且可靠的量化依据。进一步的,鉴于 A50 期指不仅会对境内投资者的心理预期产生影响,还为境外投资者开辟了风险对冲、投机运作以及资产配置的重要渠道,在一定程度上有力推动了 A 股的价格发现进程。那么从理论视角出发, A50 期指究竟能否担当 A 股的定价因子呢?本文将借助机器学习技术深入剖析富时中国 A50 股指期货的预测性和预测效果,并进一步运用 Fama-Macbeth 检验判定其是否构成 A 股市场的定价因子,以期为金融市场研究与投资实践提供参考。

2. 文献综述

国内学术界对富时中国 A50 期指的研究主要聚焦于其与 A 股市场的互动关系。已有研究表明,富时中国 A50 期指对在岸股指期货存在显著的价格传导效应,且这种影响发生的频率较高;与此同时,在岸

股指期货对离岸的富时中国 A50 期指的影响力度更为突出。特别值得注意的是,富时中国 A50 期指对沪深 300 指数存在显著的极端风险溢出效应[1]。从长期视角来看,富时中国 A50 股指期货与股指现货之间存在稳定的均衡关系,二者不仅存在双向价格引导机制,还表现出明显的波动溢出效应,且 A50 股指期货的价格发现功能存在时效性限制[2]。在市场竞争层面,富时中国 A50 股指期货与境内股指期货形成了一定程度的竞争格局。这种竞争关系对境内期货及现货市场的稳定健康发展产生了一定的负面影响[3],同时二者之间存在着相互影响的动态关联[4]。

已有研究发现富时中国 A50 期指对于 A 股市场有重要投资参考意义。研究指出,A50 期指的夜盘和盘前涨跌能有效预测沪深 300 指数开盘走势,且在岸市场暂停无法显著制约离岸市场的价格发现能力。当在岸市场出现发展迟滞或过度管制等问题时,竞争性离岸市场就会利用契机快速发展。此外,从 A50 期指与 A 股交易制度的差异,也可以侧面佐证 A50 期指对 A 股投资有较大参考价值[5]。针对 A 股市场风险收益关系与交易时段的研究表明,在整日时段和日内时段,A 股市场系统性风险与收益呈正相关;而在隔夜时段,这一正相关关系消失,甚至出现高风险-低收益的现象[6]。这说明交易时长确实对 A 股市场的收益存在影响。同时,研究认为 A 股卖空约束是导致 A 股存在定价泡沫的成因之一[7],与之形成对比的是,A50 期指拥有每个交易日二十小时的交易时间,且不存在卖空约束。

机器学习在股指期货价格预测领域展现出显著的应用潜力。研究显示,在采用随机森林、极端梯度提升树(XGBoost)、支持向量回归(SVR)和长短期记忆网络(LSTM)等模型对沪深 300 股指期货价格进行预测时,随机森林和极端梯度提升树模型表现出更为优异的预测效果[8]。此外,提升回归树方法能够对收益率、波动率以及最优风险资产权重进行精准预测[9]。从方法对比来看,非线性树型模型与神经网络在预测准确性和经济效益方面,相比传统线性方法具有显著优势[10]。在模型优化方面,通过在 Lasso 回归模型中引入非线性函数,股票收益率预测的性能得到有效提升[11]。而在模型集成领域,将随机森林、梯度提升树和神经网络等模型进行整合,基于历史股票收益数据对市场收益率涨跌趋势进行预测,能够取得极为出色的预测效果[12]。

在 A 股市场定价机制研究中,由市场因子、规模因子和盈利因子构成的三因子模型被证实对我国股票价格变动具有最优解释力,这一发现表明 A 股市场能够对企业盈利能力进行较为理性的价值评估[13]。进一步研究显示,通过考察五因子模型在 A 股市场不同发展阶段的应用效果发现,该模型同样展现出强大的价格解释能力[14]。此外,学界围绕 A 股定价因子开展了多维度的深入研究。相关成果涵盖行为金融学视角下投资者非理性行为对定价的影响[15][16],宏观经济指标波动与股票定价的内在关联[17][18],以及股票市场技术指标对价格趋势的预测作用[19][20]等多个领域。鉴于该领域研究成果丰富,在此不再逐一详述。

在机器学习应用于股指期货预测领域,虽有针对在岸股指期货的相关研究,但针对富时 A50 股指期货的机器学习预测研究近乎空白。相比已有研究,本文的贡献主要体现在以下几个方面:一是首次针对富时 A50 股指期货,运用多种前沿机器学习模型进行系统的价格预测研究,弥补了该领域在这方面的空白;二是在定价因子检验中,不仅采用了经典的 Fama-Macbeth 检验,还通过滚动回归进行稳健性检验,全面深入地分析富时 A50 期指对 A 股市场定价的影响,增强了研究结论的可靠性和普适性;三是在模型选择和应用过程中,充分考虑了 A50 期指数据的特点,对模型进行了合理的调整和优化,提高了预测和分析的准确性。

探索其机器学习预测方法,对境内投资者而言,能助力更精准把握市场趋势、优化投资组合、降低风险并提高收益,现实投资参考意义重大。从金融市场理论层面看,研究其是否为 A 股市场定价因子,有助于完善 A 股定价理论体系,拓展对市场间相互作用与价值关联的认知,为金融理论发展添砖加瓦。

3. 预测模型与研究方法

3.1. 富时中国 A50 股指期货价格预测模型

随机森林是基于决策树的集成模型，通过构建多个决策树并综合结果进行预测，能处理高维数据且不易过拟合；LSTM 是专门用于处理序列数据的循环神经网络，可利用历史信息挖掘长期依赖关系；梯度提升树是基于树模型的迭代式提升算法，通过不断拟合残差来优化预测结果，能有效处理特征间复杂关系且对异常值和噪声有一定鲁棒性。现有文献表明他们在机器学习的细分领域具有一定的代表性，本文利用这三种截然不同的机器学习方法对富时中国 A50 股指期货价格进行预测，以期获得最好的预测效果。

3.1.1. 随机森林

随机森林是一种集成学习算法，它由多个决策树组成。在训练过程中，每个决策树是基于原始训练数据集的一个自助采样样本构建的。通过综合多个决策树的预测结果来进行最终的预测，在回归问题中通常采用平均的方式，在分类问题中采用投票的方式。这种集成方式能够减少单个决策树的过拟合风险，提高模型的泛化能力。随机森林有预测速度快，可以处理高维数据，并且在训练过程中能够自动评估各个特征的重要性，对数据中的异常值和噪声有一定的容忍度。

在股价预测方面，随机森林可以捕捉到股市数据键复杂的非线性关系，处理特征交互性，无需对数据进行平稳化处理，适应数据变化的灵活性。使用随机森林算法对标普成分股的价格定向变动进行预测，结果现实模型的表现良好[12]。由于股价数据存在大量的短期波动，随机森林如果没有适当调整超参数，可能会过度拟合这些短期波动，而忽略了更本质的长期趋势和市场结构变化。但是如果配合网格搜索，随机森林相比传统的单一随机森林、支持向量机和决策树效果更好[21][22]。在长期依赖关系的捕捉能力上，随机森林相较于长短期记忆网络等模型存在先天的劣势。

3.1.2. 长短期记忆网络

LSTM 是一种特殊的循环神经网络(RNN)，专门设计用于处理序列数据。它能够对序列中的每个元素进行顺序处理，并通过内部的循环结构记住之前的信息。对于时间序列股价数据，它可以自然地按照时间顺序来处理每日的股价、成交量等数据，充分利用数据的时间顺序特性。LSTM 具有独特的门控机制，包括输入门、遗忘门和输出门，这些门可以控制信息的流动和更新。遗忘门决定了要从细胞状态中丢弃哪些信息，输入门决定了要将哪些新信息添加到细胞状态中，输出门则控制细胞状态中的哪些信息作为输出。通过这种方式，LSTM 能够有效地捕捉时间序列数据中的长期依赖关系、自动学习数据中的复杂模式和动态变化。

LSTM 在对股指期货价格进行预测时，能够充分利用股价数据的历史信息，无论是短期的波动还是长期的趋势。对于非平稳的股价数据，它可以学习到价格的动态变化规律，如市场短期内价格倾向于延续之前的走势和长期超买或超卖后的价格反转。股价数据的非平稳性表现为均值、方差等统计特性随时间变化。LSTM 的自适应学习能力使其能够自动适应这种变化，不会像一些基于平稳性假设的模型(如 ARIMA)那样在非平稳数据面前失效。它可以随着股价数据的波动性变化而调整自身的预测策略，更好地拟合数据中的复杂波动情况。LSTM 是一种动态模型，能够实时更新预测结果。通过横向对比分析人工神经网络、SVR、随机森林以及 LSTM 在股票收盘价预测上的精度，认为 LSTM 在预测精度方面表现最优[23]。但是与一些传统的机器学习模型相比，LSTM 的内部结构和参数数量较多，导致其解释性较差。而且 LSTM 有多个参数，导致其对于超参数敏感，若设置不当会导致性能不佳甚至会出现无法收敛或出现梯度消失等问题。LSTM 由于其强大的拟合能力，又容易出现过拟合的问题。

3.1.3. 梯度提升树

梯度提升树是一种基于 Boosting 思想的集成学习算法。它通过迭代地构建决策树来逐步优化模型。在每次迭代中,新的决策树会学习之前决策树的残差(在回归问题中)或错误分类样本(在分类问题中),以此来不断减少模型的预测误差。这种迭代优化的方式使得模型能够逐渐聚焦于难以拟合的数据部分,从而提高整体的预测性能。

梯度提升树在对股票市场价格预测时,因为其非参的特征可以无需验证假设,避免模型误设的问题;其收缩和袋装等技术特性增强了稳健性。同时,股市时序收益率预测常因样本量少(如中国市场仅 288 个月度样本)使复杂度高的神经网络类模型不适用,而 BRT 隐含变量选择功能,在数据量不大时也有较好样本外预测效果,且它性能优异又相对容易解释[24]。

但是尽管 BRT 对异常值和缺失值具有一定的鲁棒性,但在股价数据中,如果存在大量异常值(例如由突发事件导致的股价大幅波动)或噪声(如高频交易数据中的微小波动干扰),这些异常数据仍可能影响模型的决策树分裂规则和节点选择,进而降低预测的准确性。

3.2. A 股定价因子的检验方法

Fama-MacBeth 检验结合了时间序列回归和横截面回归。对于富时中国 A50 期指和 A 股市场股票的关系研究,它可以同时考虑期指在不同时间点上对 A 股股票价格的影响(时间序列维度),以及这种影响在不同 A 股股票之间的差异(横截面维度)。这有助于全面地分析富时中国 A50 期指是否是一个普遍有效的定价因子,而不仅仅是从单一维度进行判断。通过两步回归的方式,该检验不仅可以检验富时中国 A50 期指与 A 股价格之间的相关性(通过时间序列回归得到系数估计值),还可以进一步检验这种相关性是否受到其他因素的影响(通过横截面回归),以及其是否稳定且显著。其具体步骤如下:

第一步,估计富时中国 A50 期指的贝塔系数。A 股市场的股票收盘价为被解释变量,富时中国 A50 期指价格为解释变量,采用滚动窗口的方法进行时间序列回归,由此得到每个股票对不同风险因子的敏感度。如公式(1)所示:

$$C_{it} = \alpha_i + \beta_i P_t + \varepsilon_{it}, t=1, 2, \dots, T, \forall i \quad (1)$$

其中, C_{it} 表示 A 股的收盘价; P_t 是富时中国 A50 股指期货收盘价在 t 时期的取值; β_i 是股票 i 对 P_t 的回归系数; α_i 是截距项, ε_{it} 是随机误差项。

第二步,将第一步得到的时间序列回归系数 $\hat{\beta}_t$ (t 为时间序列的长度)作为新的横截面数据。然后对这些横截面数据进行回归,回归方程为:

$$\hat{\beta}_t = \gamma_0 + \gamma \bar{C}_i + \eta_t, t=1, 2, \dots, T, \forall i \# \quad (2)$$

γ_0 是截距项; γ 是因子敏感度的系数; $\bar{C}_i$ 股票 i 的平均收盘价格; η_t 是误差项。

在 Fama-MacBeth 检验中,主要关注的是第二步横截面回归中的系数估计值及其统计显著性。通过计算系数估计值的标准误,可以构造 t 统计量来检验系数是否显著不为零。如果富时中国 A50 期指价格相关系数(在上述横截面回归方程中)在统计上显著不为零,那么就可以认为富时中国 A50 期指价格对 A 股价格有显著的影响,即它可能是一个定价因子。

4. 富时中国 A50 股指期货价格预测模型构建与结果比较分析

4.1. 数据来源及预处理

使用 2014 年 10 月 31 日至 2024 年 12 月 13 日富时中国 A50 股指期货,去掉有误和有缺失的数据,得到 2480 个交易日每日 open、close、high、low、volume、pct_chg、MA-5、MA-60、MACD-DIFF、MACD-

DEA、MACD、RSI-6、RSI-12、RSI-24、BOLL-MID、BOLL-UPPER、BOLL-LOWER 等数据。变量详见表 1，来源 Wind 数据库。

这些选定的指标能从不同维度反映市场状况，是预测其收盘价的关键要素。日期可捕捉时间序列特性，挖掘季节性、周期性规律；开盘价、最高价、最低价与涨跌幅%呈现日内价格波动与短期趋势，直观展现价格变化动态。移动平均线从短期和中期视角揭示价格走向，其交叉等关系蕴含买卖信号。MACD 指标通过快线、慢线与柱状线变化洞察趋势转换与市场动力强弱。RSI 指标依不同周期衡量买卖双方力量对比，探测超买超卖状态以预判价格转折点。布林线指标凭借中轨、上下轨反映价格波动性与相对位置，助力预测价格在波动区间的运动趋向。成交量体现市场活跃度与资金参与度，为理解价格变动背后的市场力量提供依据。并且这些数据作为公开数据，可得性好，可靠性高，适合作为模型的特征变量。

Table 1. Explanation of variables
表 1. 变量说明

变量	名称	变量类型
close	收盘价	目标变量
date	日期	特征变量
open	开盘价	特征变量
high	最高价	特征变量
low	最低价	特征变量
volume	成交量	特征变量
pct_chg	涨跌幅%	特征变量
MA-5	5 日移动平均	特征变量
MA-60	60 日移动平均	特征变量
MACD-DEA	慢线	特征变量
MACD	平滑异同移动平均线	特征变量
MACD-DIFF	快线	特征变量
RSI-6	6 日相对强弱指标	特征变量
RSI-12	12 日相对强弱指标	特征变量
RSI-24	24 日相对强弱指标	特征变量
BOLL-MID	布林线中轨	特征变量
BOLL-UPPER	布林线上轨	特征变量
BOLL-LOWER	布林线下轨	特征变量

4.1.1. 描述性统计

如表 2 对富时 A50 股指期货相关变量的描述性统计显示，各变量特征各异。价格相关变量中，开盘价、最高价、最低价与收盘价均值相近但波动区间宽且标准差较大，表明价格离散程度明显；成交量均值 284715.57，极值差极大，标准差 162858.61，波动极为剧烈。涨跌幅均值 0.02，然极值差异大，波动离散。移动平均线方面，MA-5 与 MA-60 均值接近收盘价等，MA-60 波动稍缓。MACD 指标各变量均值有正有负，取值范围与标准差均大，变化幅度剧烈。RSI 指标不同周期均值约 50 附近，周期越长波动越小、离散越低。布林线指标各变量均值有别，波动范围与标准差均大，中轨、上下轨变化显著，综合体现出富时 A50 股指期货各指标数据复杂多变的特性。

Table 2. Descriptive statistics of FTSE China A50 futures
表 2. 富时 A50 期货描述性统计

变量	样本量	均值	最小	最大	标准差
open	2480	12782.67	8540	20514	2159.22
high	2480	12919.36	8750	20606	2180.29
low	2480	12651.80	8310	20314	2137.67
close	2480	12783.86	8540	20508	2159.57
volume	2480	284715.57	105	1546320	162858.61
pct_chg(%)	2480	0.02	-14.77	17.48	1.64
MA-5	2480	12779.69	7855	20276.20	2157.50
MA-60	2480	12716.52	8004	18459.78	2145.63
MACD-DIFF	2480	15.74	-676.34	989.01	206.78
MACD-DEA	2480	15.48	-587.11	836.06	192.23
MACD	2480	0.52	-562.18	1110.40	136.86
RSI-6	2480	51.41	7.99	97.36	17.41
RSI-12	2480	51.36	16.10	93.39	12.17
RSI-24	2480	51.19	23.99	86.24	8.54
BOLL-MID	2479	12650.85	491.15	19167.12	2401.20
BOLL-UPPER	2479	13302.61	1290.18	21052.94	2541.60
BOLL-LOWER	2479	11999.09	-307.88	17932.03	2330.63

4.1.2. 数据归一化

为了消除不同变量量纲影响,以及加速模型的收敛,本文采用 Max-Min 方法对数据进行归一化处理,计算公式见式(3)

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

4.1.3. 特征变量相关性检验

相关性检验结果如图 1 和表 3 所示,各指标与收盘价(close)的相关性存在明显差异。其中,最低价(low)、最高价(high)、开盘价(open)以及 5 日移动平均(MA-5)与收盘价的相关性极高,相关系数均接近或超过 0.99; 60 日移动平均(MA-60)、布林线中轨(BOLL-MID)、布林线上轨(BOLL-UPPER)、布林线下轨(BOLL-LOWER)与收盘价也有较高相关性,相关系数在 0.87 至 0.93 之间; 日期(Date)与收盘价的相关性处于中等水平,相关系数为 0.458076。而 MACD-DEA、MACD-DIFF、RSI-24、成交量(volume)、RSI-12、RSI-6、涨跌幅(pct_chg)、MACD 这些指标与收盘价的相关性相对较弱,相关系数大多在 0.2 及以下。

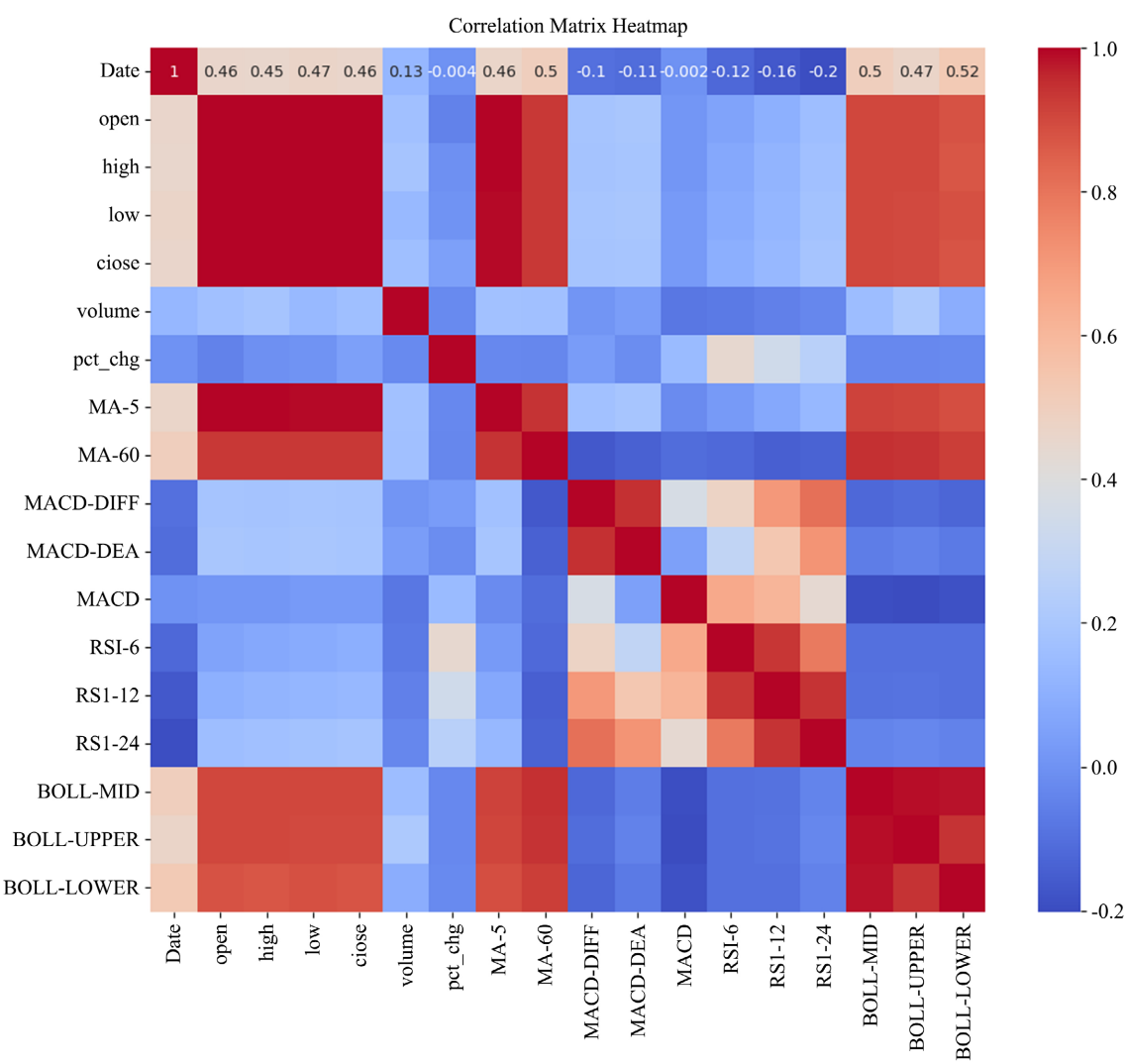


Figure 1. Correlation test result heatmap
图 1. 相关性检验结果热力图

Table 3. Results of correlation test
表 3. 相关性检验结果

	Date	open	high	low	close	volume	pct_chg	MA-5	MA-60	MACD-DIFF	MACD-DEA	MACD	RSI-6	RSI-12	RSI-24	BOLL-MID	BOLL-UPPER	BOLL-LOWER
Date	1	0.46	0.45	0.47	0.46	0.13	0	0.46	0.5	-0.1	-0.1	0	-0.1	-0.2	-0.2	0.5	0.47	0.52
open	0.46	1	1	1	1	0.16	-0.1	1	0.93	0.18	0.19	0.01	0.06	0.1	0.16	0.9	0.9	0.88
high	0.45	1	1	1	1	0.19	0	1	0.93	0.18	0.19	0.01	0.07	0.11	0.17	0.9	0.9	0.87
low	0.47	1	1	1	1	0.14	0	1	0.93	0.19	0.19	0.02	0.08	0.12	0.18	0.9	0.9	0.88
close	0.46	1	1	1	1	0.16	0.04	0.99	0.93	0.19	0.19	0.03	0.1	0.14	0.18	0.9	0.9	0.88
volume	0.13	0.16	0.19	0.14	0.16	1	-0	0.17	0.17	0.01	0.04	-0.1	-0.1	-0.1	-0	0.15	0.21	0.09
pct_chg	0	-0.1	0	0	0.04	-0	1	-0	-0	0.03	-0	0.14	0.44	0.34	0.25	-0	-0	-0
MA-5	0.46	1	1	1	0.99	0.17	-0	1	0.94	0.17	0.19	-0	0.02	0.07	0.14	0.91	0.91	0.89
MA-60	0.5	0.93	0.93	0.93	0.93	0.17	-0	0.94	1	-0.2	-0.1	-0.1	-0.1	-0.2	-0.1	0.95	0.94	0.92
MACD-DIFF	-0.1	0.18	0.18	0.19	0.19	0.01	0.03	0.17	-0.2	1	0.94	0.37	0.48	0.7	0.81	-0.1	-0.1	-0.1
MACD-DEA	-0.1	0.19	0.19	0.19	0.19	0.04	-0	0.19	-0.1	0.94	1	0.04	0.28	0.54	0.71	-0.1	-0.1	-0.1
MACD	0	0.01	0.01	0.02	0.03	-0.1	0.14	-0	-0.1	0.37	0.04	1	0.65	0.61	0.44	-0.2	-0.2	-0.2
RSI-6	-0.1	0.06	0.07	0.08	0.1	-0.1	0.44	0.02	-0.1	0.48	0.28	0.65	1	0.93	0.79	-0.1	-0.1	-0.1
RSI-12	-0.2	0.1	0.11	0.12	0.14	-0.1	0.34	0.07	-0.2	0.7	0.54	0.61	0.93	1	0.94	-0.1	-0.1	-0.1
RSI-24	-0.2	0.16	0.17	0.18	0.18	-0	0.25	0.14	-0.1	0.81	0.71	0.44	0.79	0.94	1	-0	-0	-0.1
BOLL-MID	0.5	0.9	0.9	0.9	0.9	0.15	-0	0.91	0.95	-0.1	0.06	-0.2	-0.1	-0.1	-0	1	0.99	0.98
BOLL-UPPER	0.47	0.9	0.9	0.9	0.9	0.21	-0	0.91	0.94	-0.1	-0.1	-0.2	-0.1	-0.1	-0	0.99	1	0.94
BOLL-LOWER	0.52	0.88	0.87	0.88	0.88	0.09	-0	0.89	0.92	0.13	-0.1	-0.2	-0.1	-0.1	-0.1	0.98	0.94	1

4.2. 基于 LSTM 模型的富时中国 A50 股指期货价格预测

在对数据进行 Min-Max 方法将数据归一化后，本文选择 open、high、low、ma5 对 close 进行 LSTM 建模构成四变量模型，是因其与 close 相关性高，能提供价格动态信息，构建价格波动框架与趋势线索。并且在交易逻辑中受关注，具潜在预测价值且相互补充，提供多维度输入，同时数据频率和时效性与 close 匹配，保障时间序列同步连贯，利于模型捕捉动态关联以进行有效预测。

选取 70%数据为训练集，15%数据为验证集，15%数据为测试集。LSTM 模型的多参特性导致手动调节参数不易找到较优的匹配参数。为了得到较好的模型效果，本文结合网格搜索方法对 Lstm 模型进行调参。针对 LSTM 模型超参数的调整，主要调整 epochs (训练轮数)、batch_size (批量大小)和 loss (损失函数)。epochs 取值为 40, 50, 60，训练轮数过少可能导致模型未充分学习数据特征，过多则可能引发过拟合；batch_size 取值 8, 16, 32, 64，合适的批量大小有助于模型更快收敛且能有效利用计算资源，过小会使训练过程不稳定，过大则可能导致内存占用过高和训练效率降低；loss 函数选择“mean_squared_error”和“mae”，均为常用的回归损失函数，通过比较不同组合在验证集上的表现来确定最优参数。主要参数的调整范围如表 4 所示：

Table 4. The main parameter adjustment ranges of the LSTM model
表 4. LSTM 模型主要参数调整范围

epochs	batch_size	loss
40, 50, 60	8, 16, 32, 64	‘mean_squared_error’, ‘mae’

此外，为避免特征变量选取遗漏导致模型性能下降，再选取与 `close` 相关系数大于等于 0.5 的变量作为特征变量，并采取网格搜索的方式调参后建模构成多变量模型。

四变量模型预测结果见图 2，多变量模型预测结果见图 3：

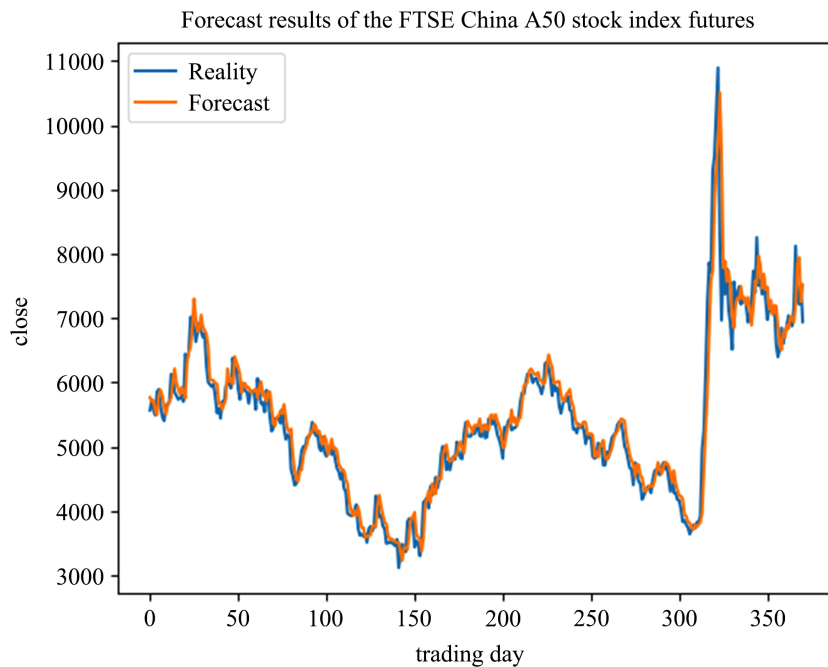


Figure 2. Four-variable model

图 2. 四变量模型

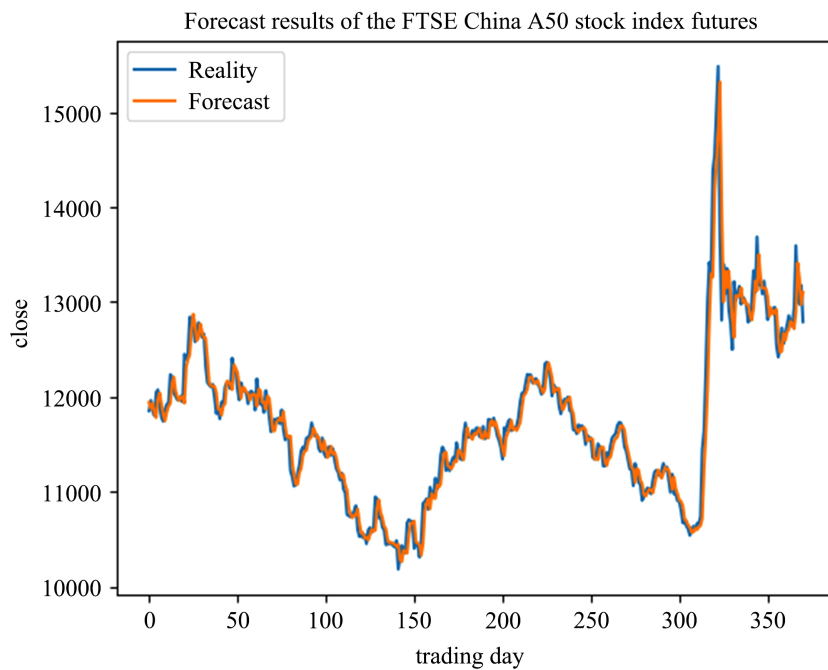


Figure 3. Multivariable model

图 3. 多变量模型

Table 5. Results of model evaluation
表 5. 模型评估结果

	四变量 LSTM	多变量 LSTM
MSE	44253.61	112809.87
MAE	126.28	210.28
R ²	0.93	0.92

两模型的性能如表 5 所示，四变量的 LSTM 模型拥有更优异的性能。

4.3. 基于随机森林的富时中国 A50 股指期货价格预测

随机森林是一种集成学习算法，由多个决策树组成。每个决策树在构建过程中会随机选择特征子集进行分裂，这种随机性使得模型能够充分利用所有特征之间的复杂关系。对于包含多个不同类型指标的完整特征集，随机森林可以挖掘出其中隐藏的交互信息。并且随机森林对特征的重要性评估是基于多个决策树的综合结果。一些在相关性分析中表现为弱相关的特征，在随机森林的集成环境下可能会发挥重要作用。

为了获得更好的模型效果，本文将变量组合分为三组，以获得更好的模型性能。A 组是与 LSTM 相同的 open、high、low、MA-5 四个变量；B 组采用与 close 相关性系数大于 0.5 的所有变量，包括 open、high、low、MA-5、MA-60、MACD-DIFF、MACD-DEA、BOLL-MID、BOLL-UPPER、BOLL-LOWER；C 组取全变量。同时，考虑随机森林对参数调节要求，三组变量均使用网格搜索。每组变量按照网格搜索最优参数建模。考虑到模型的复杂度和计算效率，选择 `n_estimators` (决策树的数量)、`max_depth` (决策树的最大深度)、`min_samples_split` (节点划分所需的最小样本数)和 `min_samples_leaf` (叶子节点所需的最小样本数)作为主要调整参数。通过网格搜索方法，设定 `n_estimators` 的取值范围为 50, 100, 150, 200，这是因为过少的决策树可能导致模型欠拟合，过多则会增加计算负担且可能过拟合；`max_depth` 取值为 None, 5, 10, 15, None 表示不限制树的深度，但可能导致过拟合，较小的深度值有助于控制模型复杂度；`min_samples_split` 取值 2, 5, 10, 较小的值会使模型对数据更加敏感，容易过拟合，较大的值则可能导致欠拟合；`min_samples_leaf` 取值 1, 2, 4，同样是在模型的拟合能力和泛化能力之间寻求平衡。使用 80%的数据用作训练集，20%数据作为测试集。网格搜索超参数范围如表 6 所示：

Table 6. The adjustment range of the main parameters of the random forest model
表 6. 随机森林模型主要参数调整范围

<code>n_estimators</code>	<code>max_depth</code>	<code>min_samples_split</code>	<code>min_samples_leaf</code>
50, 100, 150, 200	None, 5, 10, 15	2, 5, 10	1, 2, 4

由表 7 所示，性能测试结果可以看出，特征变量组合 C 整体表现最优，其 MSE 和 MAE 最小，意味着它在预测收盘价时误差相对最小，准确性更高，同时 R² 与组合 A 并列最高，说明对收盘价变化的解释能力也很强；组合 A 表现稍逊于 C，不过其 R² 也较高，预测准确性也比较可观；组合 B 的 MSE 和 MAE 最大，R² 最低，相对来说在这三个特征变量组合中预测效果最差。综合而言，全变量的组合是用于随机森林预测收盘价较优的特征变量选择。

Table 7. Performance test results of the random forest model
表 7. 随机森林模型性能测试结果

	MSE	MAE	R ²
特征组合 A	12348.28	73.8	0.9973
特征组合 B	13494.46	77.5	0.997
特征组合 C	11858.17	70.98	0.9973

4.4. 基于梯度提升树的富时中国 A50 股指期货价格预测

梯度提升树是基于树模型的迭代式提升算法，通过不断拟合前一轮模型的残差来逐步优化预测结果，它能够有效处理特征之间的复杂关系，并且对于数据中的异常值和噪声有一定的鲁棒性。在股指期货价格预测方面，其优势在于能处理变量间复杂非线性关系；可自动筛选关键变量，有效处理高维数据；采用正则化等技术抗过拟合，集成学习增强稳定性；基于决策树结构对异常值和噪声有鲁棒性，且通过逐步迭代减少噪声影响。

同样按照(三)中的变量分组方式，并采用网格搜索进行超参优化，选择 `n_estimators` (决策树的数量)、`learning_rate` (学习率)、`max_depth` (决策树的最大深度)、`min_samples_split` (节点划分所需的最小样本数)和 `min_samples_leaf` (叶子节点所需的最小样本数)作为主要调整参数。`n_estimators` 取值 50, 100, 150, 200, 学习率取值 0.05, 0.1, 0.2, 较大的学习率会使模型收敛速度加快，但可能导致错过最优解，较小的学习率则训练时间较长；`max_depth` 取值 3, 4, 5, 控制树的深度以平衡模型的拟合能力和复杂度；`min_samples_split` 取值 2, 5, 10, `min_samples_leaf` 取值 1, 2, 4, 目的是在模型的拟合能力和泛化能力之间找到最佳平衡点。主要参数调整范围如表 8:

Table 8. The adjustment range of the main parameters of the gradient boosting tree
表 8. 梯度提升树主要参数调整范围

<code>n_estimators</code>	<code>learning_rate</code>	<code>max_depth</code>	<code>min_samples_split</code>	<code>min_samples_leaf</code>
50, 100, 150, 200	0.05, 0.1, 0.2	3, 4, 5	2, 5, 10	1, 2, 4

如表 9 所示，从 MSE、MAE、R² 三个指标看，使用全变量的组合进行梯度提升树模型训练预测的结果都优于其他两组。在拟合程度都比较高的前提下，全变量的特征组合模型准确性提升较大。

Table 9. Performance test results of the gradient boosting tree model
表 9. 梯度提升树模型性能测试结果

	MSE	MAE	R ²
特征组合 A	12831.38	75.01	0.9971
特征组合 B	13218.7	78.12	0.997
特征组合 C	9266.46	65.38	0.998

4.5. 富时中国 A50 期指结果比较

不同模型有其独特结构和算法原理，对数据适应性各异。为充分发挥各模型优势，挖掘数据中的有

效信息以实现最佳预测效果，针对各模型特性进行差异化的数据处理是必要的。三种模型对于富时中国 A50 期指的预测都取其最优精度模式，结果如表 10 所示。梯度提升树在对 A50 期指价格预测中表现最佳，其均方误差(MSE)为 9266.46，平均绝对误差(MAE)为 65.38，决定系数(R^2)为 0.998；随机森林次之，MSE 为 11858.17，MAE 为 70.98， R^2 为 0.9973；LSTM 表现最差，MSE 为 44253.61，MAE 为 126.28， R^2 为 0.93。综合来看，在此次预测中，梯度提升树模型对富时中国 A50 期指价格的预测精度最高。

Table 10. The prediction accuracies of the three models
表 10. 三种模型预测精度

模型	MSE	MAE	R^2
LSTM	44253.61	126.28	0.93
随机森林	11858.17	70.98	0.9973
梯度提升树	9266.46	65.38	0.998

梯度提升树和随机森林在处理富时中国 A50 期指数据时，能有效挖掘特征间复杂关系，对数据特征提取和利用更优，而 LSTM 可能在处理该数据的长期依赖关系或特征选择上存在不足，导致预测精度不如前两者。同时，梯度提升树在拟合数据和优化预测结果上可能比随机森林更具优势，从而表现出更高的预测精度。

5. 富时中国 A50 股指期货对 A 股定价的影响

5.1. 数据来源及预处理

A 股价格依照现在主流的研究文献，使用沪深 300 作为 A 股的替代变量。选择近 3 年的沪深 300 成分股每日收盘价数据作为原始数据。沪深 300 成分股每半年更新一次，每次调整的比例通常不超过 10%，同时会遵守一些特定的规则，保证指数的稳定性和连续性，避免成分股的大幅变动对指数造成过大冲击。为了保证数据的时效性和稳定性，本文选取的股票数据为 2024 年 12 月被选入沪深 300 指数中，稳定性高且数据完整的 288 只股票，剔除掉有数据缺失的交易日，共计 726 个交易日，数据量为 209088，数据来源于 baostock。描述性统计结果见表 11：

Table 11. Descriptive statistics of stock data
表 11. 股票数据描述性统计

	open	high	low	close	preclose	pctChg
count	215,031	215,031	215,031	215,031	215,031	215,031
unique	21,640	22,340	21,809	23,431	23,428	73,341
top	6.0000	4.6500	5.5500	6.0000	6.0000	0.000000
freq	141	131	139	143	143	5610

为了对不同股票的收益率形成统一衡量标准，准确衡量股票价格的相对变化，本文对沪深 300 成分股股票每日收盘价和富实 A50 期指每日收盘价进行 Max-Min 归一化处理。此外，由于境内股票市场与新加坡股票市场存在不同的节假日，本文以 A 股市场的交易日为准，有助于去除局部噪音数据。

5.2. Fama-Macbeth 检验

5.2.1. Fama-Macbeth 回归

本文在对富时中国 A50 指数期货的价格进行预测之后，进一步验证 A50 期指是否对 A 股市场的定价有影响，旨在为境内投资者说明前者对后者的参考作用。Fama-Macbeth 检验可以单纯的观察这两个变量之间的关联，直接得到它们之间的定价关系，因此不再加入其他的控制变量。Fama-Macbeth 两步检验原理已在本文第三部分说明，这里不再赘述。

Table 12. The regression results of the second stage of Fama-Macbeth
表 12. Fama-Macbeth 第二阶段回归结果

	coef	stderr	t	P > t
γ_0	0.0006	0.002	0.246	0.806
γ_1	0.0784	0.025	3.143	0.002
η	4.5534	0.454	10.026	0

Fama-MacBeth 第二阶段回归结果如表 12 所示， γ_1 为 0.0784，t 值 3.143，系数为正说明 A50 期指相关变量每增加 1 单位，A 股市场定价相关的因变量预期增加 0.0784 单位，显示出富时中国 A50 期指对 A 股市场定价有正向的影响，并且由于 P 值较小，这种影响在统计上是显著的。

5.2.2. 稳健性检验

滚动回归是一种时间序列分析方法。它不是对整个数据集进行一次回归，而是在数据上按照一定的窗口大小逐步移动来进行多次回归。本文采用滚动回归的方式，对 Fama-Macbeth 回归结果进行稳健性检验。窗口设置为 22 个交易日，约为 1 个月，可以得到共 33 期的 γ_1 ，若通过 t 检验其显著则稳健。

Table 13. Results of rolling regression
表 13. 滚动回归结果

	coef	stderr	t	P > t
const	0.0025	0.001	2.117	0.035
beta	0.2693	0.011	25.145	0
res_std	23.248	0.571	40.707	0

结果如表 13 所示，表明富时中国 A50 期指的价格对于 A 股市场的价格是具有显著影响的，可以参考为 A 股的定价因子。

6. 结论与启示

本研究通过一系列严谨的分析与检验，得出以下关键结论：在富时 A50 股指期货价格预测上，梯度提升树凭借其出色的处理复杂关系、抗噪声和优化预测能力脱颖而出，其预测精度显著高于随机森林和 LSTM，为投资者把握 A50 期指价格走势提供了有力工具。在定价因子探究中，Fama-Macbeth 检验及滚动回归稳健性检验有力证实富时 A50 期指是 A 股市场重要定价因子，其价格变动与 A 股定价紧密相连且呈显著正向关联。

对于投资者而言,这一研究成果具有重要实践意义。在投资决策过程中,应高度重视富时 A50 期指价格波动,将其纳入投资分析框架,依据其走势提前布局或调整 A 股投资组合,有效降低风险、提升收益。在市场分析方面,要充分认识到 A50 期指对 A 股市场的引导作用,密切关注其在不同时段的变化,洞察市场预期与投资情绪转变,增强投资决策的科学性与前瞻性。

从金融市场研究视角,本研究丰富了股指期货与股票市场关系的理论内涵,为后续研究提供了实证依据与方法借鉴,有助于推动金融市场定价理论的深化与完善,促进金融市场的稳定健康发展。

参考文献

- [1] 卜林,李纪欣.全球金融周期对我国企业尾部系统风险的影响研究[J].国际金融研究,2024(8):39-48.
- [2] 霍林,黄俊杰.新加坡富时中国 A50 股指期货的波动溢出效应研究[J].区域金融研究,2021(3):21-28.
- [3] 武佳薇,孙玉奎,杨阳.富时中国 A50 股指期货对我国期现货市场的影响——理论分析及实证检验[J].证券市场导报,2019(12):51-59,74.
- [4] 叶栋梁,李彩熔.新加坡富时 A50 股指与中国股市影响关系的实证检验[J].中外企业家,2016(13):263-266.
- [5] 闵豫南.离岸与在岸股指期货市场的价格联动关系研究——基于国内股指期货市场受限的实证分析[J].南京审计大学学报,2020,17(6):79-89.
- [6] 高雅,任慧婷,熊熊.中国 A 股市场的风险-收益关系:基于隔夜和日内两时段的研究[J].计量经济学报,2024,4(6):1483-1514.
- [7] 古志辉,郝项超,张永杰.卖空约束、投资者行为和 A 股市场的定价泡沫[J].金融研究,2011(2):129-148.
- [8] 杨学威.基于机器学习算法的股指期货价格预测模型研究[J].软件工程,2022,25(12):1-8.
- [9] 李斌,屠雪永.基于机器学习和资产特征的投资组合选择研究[J].系统工程理论与实践,2024,44(1):338-359.
- [10] Tobek, O. and Hronec, M. (2021) Does It Pay to Follow Anomalies Research? Machine Learning Approach with International Evidence. *Journal of Financial Markets*, 56, Article ID: 100588. <https://doi.org/10.1016/j.finmar.2020.100588>
- [11] Freyberger, J., Neuhierl, A. and Weber, M. (2020) Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, 33, 2326-2377. <https://doi.org/10.1093/rfs/hhz123>
- [12] Ghosh, P., Neufeld, A. and Sahoo, J.K. (2022) Forecasting Directional Movements of Stock Prices for Intraday Trading Using LSTM and Random Forests. *Finance Research Letters*, 46, Article ID: 102280. <https://doi.org/10.1016/j.frl.2021.102280>
- [13] 张兆芹,王秋雨,张少华.A 股市场的盈利效应及其定价能力——基于 Fama-French 五因子模型的实证研究[J].金融理论探索,2022(1):39-50.
- [14] 李志冰,杨光艺,冯永昌,等.Fama-French 五因子模型在中国股票市场的实证检验[J].金融研究,2017(6):191-206.
- [15] 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,2004,39(10):75-83.
- [16] 刘丽文,王镇.投资者情绪对不同类型股票收益影响的实证研究[J].金融理论与实践,2016(2):90-97.
- [17] 田大伟.宏观因子套利定价模型的因子筛选及在中国股票市场的应用[J].西安财经学院学报,2006,19(5):40-45.
- [18] 杨小蛮,刘勇.我国货币政策的资本市场传导方式分析[J].农村经济,2004(1):64-66.
- [19] 史永东,王超.股票流动性影响企业生产效率吗?[J].经济管理,2021,43(11):156-175.
- [20] 朱顺伟,刘海龙,周春阳.量价趋势、信息不对称与股票收益率:基于中国 A 股市场的实证研究[J].系统管理学报,2023,32(4):774-783.
- [21] 邓晶,李路.参数优化随机森林在股票预测中的应用[J].软件,2020,41(1):178-182
- [22] 闫政旭,秦超,宋刚.基于 Pearson 特征选择的随机森林模型股票价格预测[J].计算机工程与应用,2021,57(15):286-296
- [23] Mahla, M., et al. (2019) Comparative Analysis of Machine Learning Models for Stock Closing Price Prediction. *Journal of Financial Analytics*, 25, 123-145.
- [24] 李斌,龙真.中国股票市场可预测性研究:基于机器学习的视角[J].管理科学学报,2023,26(10):138-158.