【电子与信息科学／Electronics and Information Science】 ·人工智能与数字经济专题·

基于盈余公告漂移的LGBM多因子量化策略

陈怡君[1]，李欣雨[2]，王潇逸[3]，惠永昌[2]

1）西安航空学院图书馆，陕西西安 710077；2）西安交通大学数学与统计学院，陕西西安 710049；
3）中国平安财产保险股份有限公司海南分公司，海南海口 570100

摘　要：在资本市场波动加剧的时代，挖掘有效因子与市场信息，构建合适的投资组合策略，可以实现对风险的控制和获取稳定且持续的超额收益率．选取2018—2022年第1季度中国沪深两市A股上市股票的业绩报告作为研究对象，以公司盈余公告后的1～12周作为时间窗口，通过研究盈余公告后的股价漂移(post-earnings-announcement drift，PEAD)选取市场异象的代理变量预期外盈余因子与其他5个相关市场异象因子，并使用信息系数(information coefficient，IC)、信息比率(information ratio，IR)和双重排序法进行有效因子的筛选和检验．考虑到本次量化选股是低数据量、低频次、特征值高有效性的分类任务，采用基于轻量梯度提升树的多因子量化策略构建投资组合预测股票的收益率，并与传统量化策略(简单打分法、基于预期外盈余的单因子模型、IC值加权的多因子选股模型)、基于其他机器学习模型(支持向量回归(support vector regression，SVR)、人工神经网络(artificial neural network，ANN)与分布式梯度增强(extreme gradient boosting，XGBoost))的量化策略进行对比．实证结果表明，在基于A股市场第1季度PEAD效应的股票超额收益率预测中，轻量级梯度提升机(light gradient boosting machine，LGBM)机器学习多因子量化策略构建的投资组合在多空组合中实现的年均收益率达到21.633%，超过基准年化收益率20.184%．LGBM多因子量化策略构建的投资组合在A股市场表现优异，较其他量化策略有显著提升且更为稳定，可更好地控制组合风险并获取更高的超额收益．

关键词：数字经济；量化投资；多因子选股；轻量梯度提升树；盈余公告后漂移；异象因子

中图分类号：TP18；TP391　　**文献标志码：**A　　**DOI：**10.3724/SP.J.1249.2024.03313

Multi-factor quantification strategy of LGBM based on post-earnings announcement drift

CHEN Yijun[1], LI Xinyu[2], WANG Xiaoyi[3], and HUI Yongchang[2]

1) Xi'an Aeronautical University Library, Xi'an 710077, Shaanxi Province, P. R. China
2) School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, Shaanxi Province, P. R. China
3) Hainan Branch, Pingan Property Insurance Co. Ltd., Haikou 570100, Hainan Province, P. R. China

Abstract: In the era of intensified volatility of domestic and international capital markets, exploring effective factors and market information to construct appropriate investment portfolio strategies is of great significance to control risks and obtain stable and sustainable excess returns. In this study we select the performance reports of A-share listed stocks from the first quarter of 2018 to 2022 as research objects. Taking 1 to 12 weeks after the company's earnings announcement as the time window, we examine the post-earnings-announcement drift (PEAD) of stock prices, selecting the expected excess earnings factor, a proxy variable of market anomaly, and other 5 related market

Received: 2023-07-28; **Accepted:** 2023-09-04; **Online** (CNKI)**:** 2024-04-10
Foundation: National Social Science Foundation of China (23BTJ057)
Corresponding author: Associate professor HUI Yongchang (huiyc180@xjtu.edu.cn)
Citation: CHEN Yijun, LI Xinyu, WANG Xiaoyi, et al. Multi-factor quantification strategy of LGBM based on post-earnings announcement drift [J]. Journal of Shenzhen University Science and Engineering, 2024, 41(3): 313-322. (in Chinese)

anomaly factors. The effective factors are screened and tested by information coefficient (IC), information ratio (IR) and double sorting methods. Considering that the quantitative stock selection is a classification task with low data volume, low frequency and high validity of eigenvalues, a multi-factor quantitative strategy based on a lightweight gradient boosting tree is used to construct a portfolio to predict stock returns. This approach is compared with traditional quantitative strategies (simple scoring method, single factor model based on expected surplus, IC value weighted multi-factor stock selection model) and quantitative strategies based on other machine learning models such as support vector regression (SVR), artificial neural network (ANN) and extreme gradient boosting (XGBoost). Empirical results show that in forecasting stock excess returns based on the PEAD effect in the A-share market, the portfolio constructed by the light gradient boosting machine (LGBM) multi-factor quantitative strategy achieves an average annual return of 21.633% in the long-short combination portfolio, exceeding the benchmark annualized return of 20.184%. The comprehensive analysis of various indicators reflect that the investment portfolio constructed by LGBM multi-factor quantitative strategy has excellent performance in the A-share market and more show significant improvement compared to other quantitative strategies while maintaining stability, which can better control portfolio risks and achieve higher excess returns.

Key words: digibal economy; quantitative investment; multi-factor stock selection; lightweight gradient lifting tree; post-earnings-announcement drift; anomaly factor

量化投资是一种以历史数据为出发点、以模型为核心，利用程序化交易手段进行投资，从而客观地获取稳定收益的投资策略[1]，具有客观和反应快等优势，现已成为资本市场构建投资策略最主要的方式之一．量化选股是量化投资的关键之一．20世纪60年代，SHARPE[2]首次提出资本资产定价模型，该理论认为资产的期望收益与系统风险显现出正向的线性关系．ROSS[3]建立了套利定价模型，认为资产价格变动是受多因素综合影响的结果．基于此，学术界针对多因子的选择进行了深入的理论研究和长足的探索．

近些年来，被发现并发表出来的股票因子多达数百种[4]，但仍有许多因子难以被市场捕捉．纵观针对中国沪深两市A股市场的研究发现，目前的大多数研究都只是将常见的影响因子作为数据特征纳入量化模型中[5]，然而，随着量化交易的盛行和市场有效性的提高，实证检验发现部分已被发掘的因子对股价的影响逐渐减弱，难以持续地提供超额收益[6]．因此，深入挖掘更稳定的影响因子至关重要．

本研究将资本市场上一种长期稳定存在的市场异象——股价的盈余公告后漂移(post-earnings-announcement drift，PEAD)作为一个重要的异象因子纳入量化投资模型框架中．PEAD效应指当前盈余意外与随后的股票收益之间存在显著的正相关关系，因此，股价分别呈现向上或向下持续漂移的趋势，且漂移存在滞后性[7]．研究证明，PEAD效应在各资本市场持续存在，且标准化预期外盈利(standardized unexpected earnings，SUE)、公司规模、行业板块和机构持股等因素都会对盈余惯性的方向和显著性产生影响[8-10]．

此外，为解决因子选择缺乏理论依据和经济意义等问题，本研究将PEAD效应的代理因子(SUE因子)与其他市场异象因子融合，构建多因子选股模型．多因子选股模型在早期主要使用传统的计量方法来构建，如简单打分法和多元线性回归方法等[11-12]．但是，随着金融市场日趋复杂，传统的量化投资计量方法逐渐遇到了技术瓶颈：一是因子数量日益庞大且功能相似，传统方法难以区分因子间的相互作用；二是传统研究方法在高维因子的处理中存在一定的困难[13]．随着科技的进步与机器学习的日益发展，量化投资与算法的结合逐渐紧密，学者们开始采用支持向量机[14]、遗传算法[15]、反向传播(back propagation，BP)神经网络[16]、决策树[17-18]、自注意机制[19]和集成模型[20-21]等各种非线性方法建立不同的模型，并对模型进行改进和优化．本研究尝试使用轻量级梯度提升机（light gradient boosting machine，LGBM）与多因子选股模型相结合的方法构建投资组合[22]，并进行实证研究．LGBM算法在A股市场投资量化中表现效果较为优异，相较于传统方法可以更好的识别非线性和多因子特征数据[23]．

综上，本研究基于2018—2022年第1季度A股上市股票的业绩报告数据，以盈余公告后的1～12

周作为时间窗口，通过研究PEAD效应与相关历史文献构建市场异象的代理变量SUE因子与其他市场异象因子，并采用双重排序法进一步筛选对股价漂移有影响的市场异象因子，最后利用简单打分法、信息系数(information coefficient，IC)值加权的多因子选股模型、基于SUE的单因子模型、支持向量回归(support vactor regression，SVR)、人工神经网络(artificial neural network，ANN)、极端梯度提升(extreme gradient boosting，XGBoost)和LGBM多因子量化策略分别构建投资组合预测股票超额收益．LGBM多因子量化策略构建投资组合的整体流程如图1．本研究的贡献主要包括以下两方面．

1）因子选择更具有理论依据和合理性；在理论分析和文献综述的基础上选择有效且稳定的市场异象因子，并将PEAD效应的代理变量SUE因子与其他市场异象因子相融合构建多因子量化策略．

2）结合在A股市场具有良好的适应性的LGBM算法构建量化策略，为研究A股市场基于PEAD效应的量化投资方法提供了一个新思路．与传统量化策略及基于其他机器学习模型的量化策略相比，所构建的LGBM多因子量化策略在样本期内可获得更高且更稳定的超额收益率．

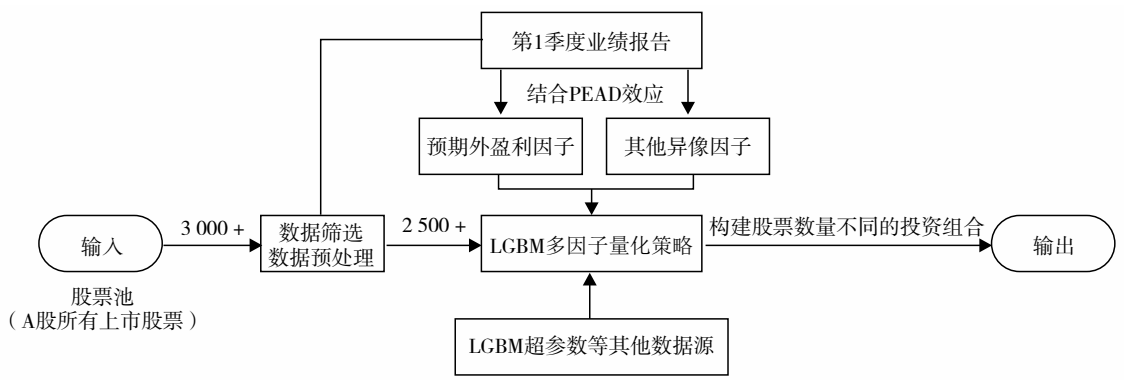

图1　LGBM多因子量化投资策略构建投资组合的整体流程(3 000 +和2 000 +分别表示当期股票样本数量超过3 000和2 000)

Fig. 1　Process of building a portfolio by LGBM multi-factor quantitative investment strategy. (3 000 + and 2 000 + represent more than 3 000 and 2 000 stock samples in the current period, respectively.)

1　PEAD效应

1.1　数据来源

本研究的初始数据集涵盖了中国沪深两市所有上市股票，盈余公告的时间范围为2018—2022年，有效避免了幸存者偏差问题．存在可能会使PEAD现象失真的股票有：① 上市未满2 a的新股相较于同行业股票会存在普遍的溢价情况，而较频繁的融资扩张在短期内也会使盈余公告数据好于预期；② 停牌的股票在复牌时会产生巨大的股价波动，影响其盈余公告数据；③ 当期特别处理(special treatment，ST)股票存在诸多限制，存在盈利缺陷．因此，本研究剔除了初始数据集中上市未满2 a的新股、盈余公告前1 d处于停牌状态的股票，以及当期被标注为“ST”的股票，使选定股票的盈余公告尽可能地展示真实且有效的数据．处理后的数据集内有效股票数量如图2．其中，“半年度”指上半年(第1季度与第2季度的和)；“前3季度”指第1、2和3季度的和；截止数据采集时2022年度数据尚

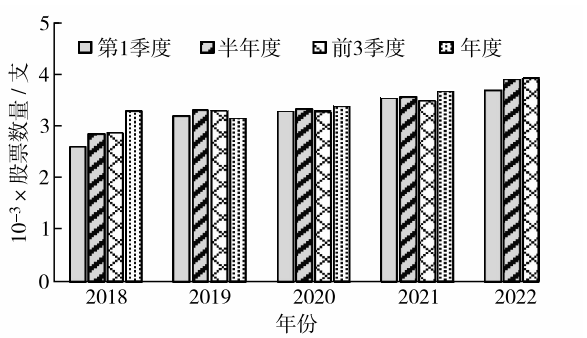

图2　2018—2022年各季度盈余公告后有效股票数量

Fig. 2　Number of shares in force after earnings announcement for each quarter from 2018 to 2022.

未全部公布，因此2022年度无年度数据．所有股票数据源自万得(Wind Information，WIND)数据库(https://www.wind.com.cn/portal/zh/EDB/index.html)．

1.2　SUE因子选择

SUE因子与盈余公告中的实际盈余与市场预期盈余相关，并与股价存在正相关关系．为找出最适合A股市场的SUE因子，本研究选取4种常用的SUE计算方法，包括：

1）以企业每股收益与净资产收益率计算.

$$E_{\mathrm{SU},t} = E_{\mathrm{S},t} - E_{\mathrm{S},t-1}\left(1 + E_{\mathrm{RO},t-1}\right) \tag{1}$$

其中，$E_{\mathrm{SU},t}$为第t年企业每股收益与净资产收益率；下标t为年份索引编号；$E_{\mathrm{S},t}$为企业单季度每股收益(earnings per share，EPS)；$E_{\mathrm{RO},t}$为单季度净资产收益率(rate of return on common stockholders′ equity，ROE).

2）以企业实际盈利与机构预期盈利之间的差距计算.

$$E_{\mathrm{SU},t} = P_{\mathrm{Y},t} - \hat{P}_{\mathrm{Y},t} \tag{2}$$

其中，$P_{\mathrm{Y},t}$为年同比净收益(year-on-year net proceeds)；$\hat{P}_{\mathrm{Y},t}$机构预测的年同比净收益.

3）以企业实际盈利与特定数值的差距计算.

$$E_{\mathrm{SU},t} = P_{\mathrm{Y},t} - C \tag{3}$$

其中，C为市场的预期盈利.

4）不受机构关注股票企业盈利和股价漂移.

$$E_{\mathrm{SU},t} = P_{\mathrm{Y},t} \tag{4}$$

1.3 盈余公告种类、SUE与PEAD效应

公司盈余公告可分为经营情况简报、业绩预报和业绩报告，经对比分析，本研究选取信息详实且经过审计的业绩报告作为研究对象，以求展示客观准确的盈余公告数据. 为选取出能产生明显的PEAD效应的业绩报告种类，基于式(1)至式(4)计算各季度业绩报告的SUE，将数据集中的股票按照SUE值升序排列并平均分为5组，记为A—E，计算每组股票30 d的股价漂移. SUE对各业绩报告的区分程度见图3. 其中，SUE1至SUE4分别对应以式(1)至式(4)计算出的各季度业绩报告的SUE.

分析图3可知，第1季度报告的SUE与超额收益率表征的股价漂移之间存在正相关关系；半年度报告的部分SUE与股价漂移之间存在正相关关系，但相关性较弱；前3季度报告的SUE完全无法体现与股价漂移之间的关系；年度报告的SUE与股价漂移之间关系出现了分化，存在负相关关系. 可见，并非所有的业绩报告都会导致股价漂移. 相比其他3个SUE，SUE1与股价漂移具有最强的正相关性.

为更好地解释我国A股市场中的PEAD效应，选取第1季度业绩报告与SUE1用于后续研究.

1.4 时间窗口与PEAD效应

PEAD效应是一个持续性过程，发布盈余公告后本应即时发生的价格调整往往需要数周甚至数月才能完全实现[24]. 为确保发布盈余公告后股价实现充分调整，本研究选择第1季度业绩报告发布后的

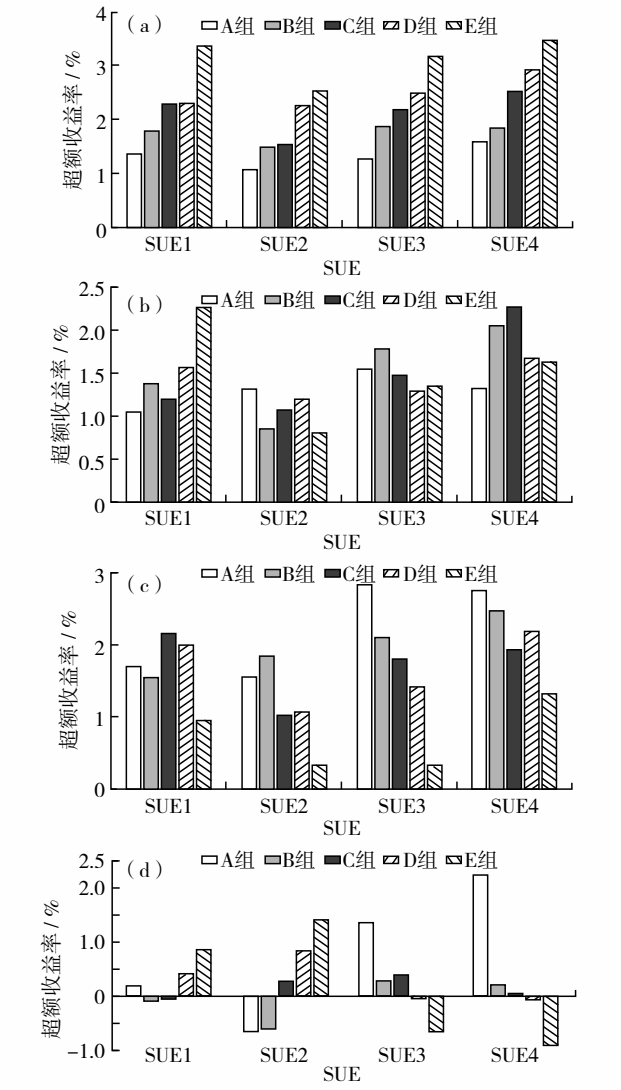

图3 SUE对(a)第1季度、(b)半年度、(c)前3季度和(d)年度业绩报告的区分程度

Fig. 3 The degree to which SUE distinguishes among (a) the 1st quarter, (b) semi-annual, (c) 3 quarters, and (d) annual results reports.

1～12周作为时间窗口. 需注意的是，虽然12周后股价漂移可能仍未结束，但此时已经是半年度业绩报告的公告期了，为避免新发布的业绩报告对过去的业绩报告产生的股价漂移造成影响，本研究未将12周以后的股价漂移纳入时间窗口.

2 其他异象因子的选取与检验

基于对PEAD效应的研究，本研究还结合对市场异象的分析寻找其他有效异象因子，包括以双重排序法检验其他异象因子与SUE之间是否存在信息增益；以皮尔森相关系数验证各个因子之间的相关

性，避免共线性问题.

2.1　其他异象因子的选取

2.1.1　跳空程度

跳空指当股票受到重大事件影响或流动性冲击时会产生股价跳空波动的现象，可用于测量市场对事件反应的强度与方向. 当发生盈余公告事件时，若企业实际财务数据与市场预期的数据存在偏差就会导致跳空现象的出现，此时跳空的强度和方向所隐含的信息可反应预期盈利动量的强弱[25].

2.1.2　交易成本

交易成本是造成PEAD效应的重要因素，即由于交易成本的存在导致投机者无法正确确定股价与及时调整投资策略，进而出现股价漂移现象，交易成本越高的股票漂移效应越强[26].

2.1.3　异常交易量与意见分歧

异常交易量指市场在盈余公告之前经常会出现异常高换手率的现象. 由于中国A股市场中投资经验不足的散户居多，很容易受非理性因素的影响导致异常交易行为的发生[27]. 意见分歧指由于投资者注意力有限、经验差异和投资方法不同等，易在盈余公告后产生意见分歧而采取不同的交易行为. 盈余公告前的异常交易量行为代表着投资者的投机行为，投机行为的剧烈程度需通过意见分歧程度来衡量. 意见分歧程度的扩大会刺激盈余公告前的投机行为，从而发生异常高交易量现象. 因此，在投资者意见分歧程度越高的前提下，异常高交易量意味着更高的投机泡沫，预示着盈余公告后预期偏差会导致更低的股票超额收益[28].

2.1.4　机构集中度

机构投资者在市场上扮演着重要的角色，被认为是市场上的聪明投资者. 他们积极收集信息从而发现市场上潜在的盈利机会并通过套利行为消除市场的错误定价[29]. 但A股市场具有特殊性，如散户投资者占主导地位、法治薄弱和所有权高度集中. 同时，由于机构羊群行为的存在，机构持股在A股有较高的PEAD值和较慢的价格发现过程. 即机构投资者持股占比越高，股价漂移效应越高，这种正向关系取决于制度羊群效应的强度和方向[30].

2.1.5　异质波动率

异质波动率异象是典型的市场异象，异质波动率通过影响未来预期收益，进而影响PEAD效应. 本研究选取了除SUE以外其他6个异象因子(跳空程度、交易成本、异常交易量、意见分歧、机构集中度与异质波动率)，用于构建LGBM多因子量化策略. 代理因子SUE及6个异象因子的信息如表1.

表1　代理因子SUE及异象因子信息表

Table 1　Proxy factor SUE and anomaly factors information table

因子名称	简称	解　释
预期外盈利	SUE	当期业绩实际增长与预期增长之间的差值
跳空程度	Jump	盈余公告发布日后首日集合竞价成交价对前1 d收盘价的涨跌幅度
交易成本	Cost	公司盈余公告发布当日股票的交易成本
异常交易量	AbnVol	盈余公告发布日前12～2 d的平均成交量与60～12 d的平均成交量的比值
意见分歧	Doo	盈余公告发布日前30～12 d股票成交量的方差
机构集中度	ICD	企业在第1季度报告公告前最新的机构持仓比例
异质波动率	IVOL	盈余公告发布日前30～1 d内股票价格的异质波动标准差

2.2　IC与IR检验法

本研究以IC检验法对7个异象因子作有效性检验(IC绝对值大于0.05则说明因子具有较强的预测能力)；以信息比率(information ratio，IR)检验法作稳定性检验(IR值大于0.5即可认为因子稳定性较好). 将股票的当期因子和下一期的预期收益进行分别排名，并计算两组排名的IC绝对值均值与IR值，结果见表2. 由表2可见，各个因子的IC绝对值均值均大于0.05且IR值均大于0.5，证明所选择因子兼具有效性与稳定性.

表2　7个异象因子的IC和IR检验表

Table 2　IC and IR test tables of 7 anomaly factors

因子	IC绝对值均值	IC标准差	IR值
SUE	0.138	0.047	2.916
Jump	0.100	0.029	3.425
Cost	0.113	0.124	0.902
AbnVol	0.055	0.065	0.846
Doo	0.109	0.121	0.898
ICD	0.091	0.106	0.865
IVOL	0.092	0.117	0.781

2.3 双重排序法

以2018年万德全A指数为基线，将SUE按升序平均划分为5组得到SUE-1—SUE-5，其他6个异象因子进行同样方法划分，再以双重排序法交叉构建投资组合并计算组合的超额收益，对比验证其他6个异象因子与SUE之间是否存在信息增益，结果如表3.

分析表3可知，SUE分别与Jump、Doo和ICD因子存在正相关关系，在SUE最高的情况下，盈余公告后首日跳空最高、意见分歧度最大，以及机构集中度最高的一组股票分别实现了8.206%、8.596%和8.338%的超额收益. SUE与Cost和IVOL因子都存在负相关关系，在SUE最高的情况下，交易成本最小和异质波动率最小的股票分别实现了7.622%与8.327%的超额收益；AbnVol与SUE无明显正反相关关系，但AbnVol趋于中位数的股票组合可实现6.673%的超额收益，AbnVol过高或过低都会使股票组合的超额收益降低. 综上，其他6个异象因子与SUE存在信息增益，共同构建的投资组合获得的超额收益均高于仅使用SUE构建的投资组合获得的6.103%的超额收益.

2.4 皮尔森相关系数检验

通过计算异象因子的皮尔森相关系数，检验异象因子之间的相关性，所得相关性热力图如图4. 由图4可见，异象因子之间的相关系数均位于合理区间，无共线性问题.

3 投资组合构建

本研究基于异象因子SUE、Jump、Cost、AbnVol、Doo、ICD与IVOL利用模型构建投资组合. 模型的选择基于数据量和任务特点相关经验确定. 基于预期外盈利效应的量化投资研究是一个低频次任务，数据规模较小且特征值非稀疏矩阵，利用机器学习模型即可完成处理. 本研究中的量化选股实际为面板数据中的分类任务，而深度学习模型通常更适用于高频的时间序列模型选股而非本研究中的数据；相关文章的研究分析中主流的做法依然是采用机器学习模型用于量化选股，但本研究在此基础上基于金融市场特性加入择时手段，能提升因子的有效性，实现高于主流做法的市场超额收益.

通过传统量化策略(简单打分法、多元线性回归和基于SUE的单因子模型)、基于其他机器学习模型(SVR、ANN和XGBoost)的量化策略与LGBM多因子量化策略构建投资组合.

表3 双重排序法超额收益率

Table 3 Double ranking excess return

异象因子	组别	SUE-1	SUE-2	SUE-3	SUE-4	SUE-5
SUE	—	−1.085	−0.552	0.740	1.909	6.103
Jump	Jump1	−1.364	−2.579	0.872	−0.814	3.903
	Jump2	−1.060	−1.359	0.056	2.629	5.932
	Jump3	−1.562	−0.412	0.438	1.612	3.397
	Jump4	−0.515	0.762	0.716	1.351	6.082
	Jump5	1.575	2.199	2.732	3.565	8.206
Cost	Cost1	1.402	1.870	3.501	4.052	7.622
	Cost2	−0.857	−0.715	0.407	1.804	6.134
	Cost3	−1.638	−1.551	−0.668	2.115	6.402
	Cost4	−2.013	−1.106	1.227	2.732	6.732
	Cost5	−2.488	−2.436	−1.085	−1.346	5.894
AbnVol	AbnVol1	−3.241	−1.617	−0.375	0.030	5.430
	AbnVol2	−0.153	−0.694	0.152	0.091	4.010
	AbnVol3	−1.515	−0.633	2.629	3.811	6.673
	AbnVol4	−0.101	0.541	2.371	−0.352	4.999
	AbnVol5	−3.008	−3.125	−1.701	−2.517	1.832
Doo	Doo1	−2.451	−0.510	−1.957	−0.198	2.279
	Doo2	−0.115	0.492	1.015	2.129	4.111
	Doo3	0.735	0.334	0.721	3.104	6.743
	Doo4	−0.035	0.571	1.884	2.426	7.154
	Doo5	0.779	1.743	1.777	3.158	8.596
ICD	ICD1	−2.460	−1.787	−2.325	0.409	4.016
	ICD2	−1.414	−0.309	0.235	1.435	5.725
	ICD3	−0.379	−0.961	2.035	1.596	6.023
	ICD4	−0.094	1.218	2.132	4.025	8.238
	ICD5	1.061	1.520	3.178	2.475	8.338
IVOL	IVOL1	3.714	4.532	4.521	4.813	8.327
	IVOL2	3.121	5.298	3.513	5.374	7.192
	IVOL3	1.984	4.242	2.216	3.090	5.294
	IVOL4	−1.517	3.214	2.174	3.317	4.275
	IVOL5	−2.199	2.357	−0.381	−0.129	3.261

注：灰底和下划线数据分别为该项目下的最优值和次优值.

3.1 传统量化策略

传统量化策略是量化投资领域的重要组成部分，本研究选定3种传统量化策略构建投资组合作

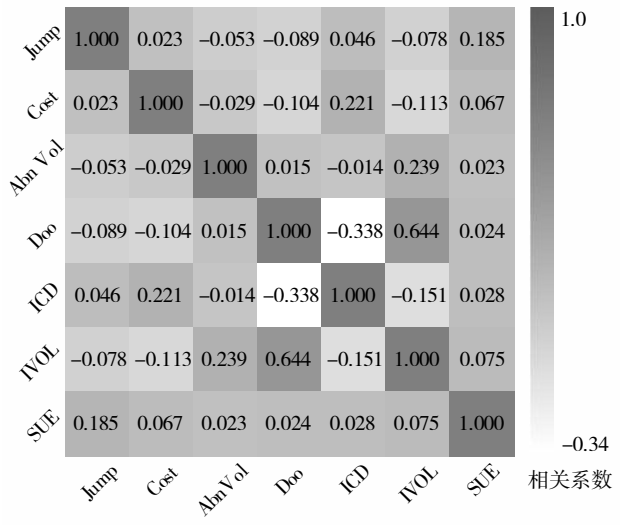

图4　皮尔森相关系数热力图

Fig. 4　Pearson correlation coefficient heat map.

为传统量化策略对照实验组，具体构建思路如下：简单打分法将异象因子分组排序并按照20%分位分组打分，选择分数最高的一组股票做多，同时选择分数最低的一组股票做空以构建多空投资组合；多元线性回归模型将7个异象因子作为自变量，股票的12周超额收益作为因变量，以第$t-1$年和第$t-2$年共2 a的自变量与因变量数据进行回归分析并计算其回归方程，进而利用该方程预测第t年的超额收益并构建多空投资组合；基于SUE的单因子模型做多SUE值最高的一组股票，并做空SUE值最低的一组股票以构建多空投资组合.

3.2　基于其他机器学习模型的量化策略

机器学习模型已广泛应用于量化交易. 其中，SVR模型有着极强的非线性近似能力；ANN模型广泛应用于分类问题；XGBoost模型则经常在各类量化投资领域脱颖而出. 因此，本研究选用上述3种模型作为机器学习模型对照组，统一设置神经网络参数，以7个异象因子作为特征值输入模型，进而构建多空投资组合.

3.3　LGBM多因子量化策略

LGBM属于决策树模型，不但可区分对股价漂移存在线性影响的异象因子，对AbnVol这类对股价漂移有着非线性影响的因子同样有很好的区分能力.

本研究将7个异象因子作为特征值输入LGBM模型，将股价漂移这一离散数据标签化，从而将投资组合的构建转化为二分类任务. 需要注意的是，标签化需兼顾样本均衡与投资组合超额收益最大化：若正例与反例标签数量差距过大，模型会因为样本数量太小而难以区分两种标签；若正例标签数量过多，会将一部分涨幅未达到预期目标的股票纳入正例，影响整体组合业绩. 因此，本研究在构建做多组合时，将12周内获得的超额收益大于样本整体超额收益均值加上标准差的股票标记为1，剩余股票标记为0. 构建做空组合时，将12周内获得的超额收益小于样本整体超额收益均值减去标准差的股票标记为1，剩余股票则标记为0. 通过调整标签中正例的比例来调整投资组合的股票数量，提高正例的比例会使得组合中股票数量提高，降低投资组合的收益波动和风险，但也会降低投资组合的超额收益.

4　实证研究

为验证LGBM多因子量化策略的有效性与稳定性，将LGBM多因子量化策略应用于A股市场，基于第1季度业绩报告选择异象因子并构建投资组合，并通过计算投资组合的超额收益、收益标准差与夏普比率.

4.1　业绩评价指标

为对量化策略的表现进行客观评价，选用5 a平均超额收益(R_{AE})、标准差(σ)与夏普比率(R_{S})3个评价指标来评价各个策略构建的投资组合. R_{AE}值可直观得展示投资组合的盈利能力，σ值越小代表投资组合的波动性越小，R_{S}值越大代表投资组合的相对性价比越高. 各评价指标计算公式为

$$R_{\mathrm{AE}}=\frac{1}{N}\sum_{t=1}^{N}\left[R_t-E(R_t)\right] \tag{5}$$

$$\sigma=\sqrt{\frac{\sum_{t=1}^{N}(R_t-\mu)^2}{N}} \tag{6}$$

$$R_{\mathrm{S}}=\frac{E\left(R_t\right)-R_{\mathrm{f}}}{\sigma} \tag{7}$$

其中，N为预测的年数；R_t为第t年投资组合的收益；$E\left(R_t\right)$为第t年投资组合的期望收益；μ为R_t的均值；R_{f}为无风险收益.

4.2　业绩评价基线

选取万德全A指数作为业绩评价基线，计算各个投资组合的超额收益. 同时计算样本整体的超额收益，衡量各个投资组合的超额收益是否超过市场回报率. 需要注意的是，由于股票样本中已剔除了当期上市未满2 a的新股、停牌的股票以及ST股

票，且每支股票的超额收益的计算期间是第1季度业绩报告后首日起至12周结束，因此，整体样本的平均超额收益并不为0.

4.3 对照实验

对照试验旨在回答两个问题：基于PEAD效应所构建的投资组合能否在A股市场获得超额收益？LGBM多因子量化策略构建的投资组合获得的超额收益相较于传统量化策略与基于其他机器学习模型的量化策略有多大提升？

本研究将简单打分法、多元线性回归模型、基于SUE的单因子模型、基于其他机器学习模型(SVR、ANN和XGBoost)的量化策略和LGBM多因子量化策略构建的投资组合的业绩进行比较分析. 其中，简单打分法、多元线性回归模型和基于SUE的单因子模型选择当期得分最靠前的20支股票构建多空投资组合计算超额收益；LGBM多因子量化策略与基于其他机器学习模型的量化策略则通过调整正例标签进而调整投资组合的股票数量，最终用于构建多空投资组合的股票也为20支. 各模型构建的股票组合业绩如表5.

表5 各模型以20支股票构建的投资组合业绩表

Table 5 A portfolio performance table of 20 stocks for each model

模型	R_{AE}/%	σ	R_S
单因子模型	7.790	8.358	0.932
简单打分法	15.850	9.735	1.628
多元线性回归	12.069	4.449	2.713
SVR	12.556	5.056	2.483
ANN	13.644	4.373	3.120
XGBoost	16.872	5.427	3.109
LGBM	21.633	6.883	3.143

注：灰底和下划线数据分别为该项目下的最优值和次优值.

由表5可见，基于SUE的单因子模型构建的投资组合所实现的超额收益为7.790%，高于样本整体的超额收益(1.449%). 由于LGBM模型识别因子间线性关系和非线性关系的能力更强，其构建的投资组合实现的年均收益达到21.633%，超过次优值4.761%；R_S = 3.143，能够较好的平衡风险与收益，其中，n为组合中股票数量.

综上，基于PEAD效应所构建的投资组合能够在A股市场获得超过市场平均水平的超额收益. LGBM多因子量化策略构建的投资组合获得的超额收益比传统量化方法与其他机器学习模型都有显著提升.

4.4 稳定性检验

稳定性检验旨在回答随着选定的股票数量的增加，LGBM多因子量化策略构建的投资组合能否稳定得获取超额收益？

将投资组合的股票数量由20支增加到50支与100支并进行稳定性检验. 以标准差衡量其投资风险，以夏普比率综合比较其收益与稳定性，结果如表6.

表6 各模型以不同股票数量(50与100支)构建的投资组合业绩表

Table 6 A portfolio performance table with different stock numbers (50 versus 100) for each model

n	模型	R_{AE}/%	σ	R_S
50	单因子模型	12.503	6.968	1.794
	简单打分法	12.461	9.431	1.321
	多元线性回归	13.728	11.686	1.104
	SVR	9.281	3.904	2.377
	ANN	11.187	4.554	2.456
	XGBoost	11.436	3.020	3.787
	LGBM	16.750	2.857	5.862
100	单因子模型	10.261	5.109	2.008
	简单打分法	12.900	11.686	1.104
	多元线性回归	11.957	5.845	2.046
	SVR	7.908	1.340	5.902
	ANN	8.944	3.587	2.493
	XGBoost	10.559	3.522	2.997
	LGBM	13.318	2.183	6.101

注：灰底和下划线数据分别为该项目下的最优值和次优值.

由表6可见，基于PEAD效应构建的多空投资组合收益超过样本整体的平均超额收益(1.449%)，且基于多因子构建的投资组合的收益相较于单因子模型构建的投资组合有更好的收益表现. 在多因子投资组合中，基于LGBM多因子量化策略构建的投资组合的收益率最高，为16.750%(50只股票样本)和13.318%(100只股票样本)；标准差均为最优值与次优值，分别为2.857与2.383；夏普比率始终保持最优，分别为5.862与5.589.

综上，随着组合中股票数量的增多，基于LGBM多因子量化策略构建的投资组合的超额收益率符合均值回归预期. LGBM多因子量化策略选定

不同数量的股票构建的投资组合的夏普比率始终保持最优，证明其能较好得控制组合风险，同时稳定得获取超额收益.

5 结　论

针对中国沪深两市A股市场的PEAD效应开展研究，采用IC、IR、双重排序法与皮尔森相关系数检验了SUE、跳空程度、交易成本、异常交易量、意见分歧、机构集中度和异质波动率7个市场异象因子的有效性、稳定性，以及因子之间的信息增益与相关性；分别采用简单打分法、多元线性回归模型、基于SUE的单因子模型、基于其他机器学习模型（SVR、ANN和XGBoost）的量化策略和LGBM多因子量化策略构建投资组合，测算各投资组合的超额收益率与夏普比率并进行对比分析. 得出以下主要结论：

1）第1季度业绩报告能产生明显的PEAD效应，因此，基于PEAD效应所构建的投资组合可在A股市场获得超额收益.

2）经因子检验与实证研究，本研究选取的7个市场异象因子有效且稳定. 将上述异象因子作为特征值加入量化策略，可有效提高投资组合的收益率.

3）基于LGBM的多因子量化策略在A股市场表现优异，与传统量化策略与其他基于主流机器学习模型的量化策略相比，可更好地控制组合风险并获取更高的超额收益.

基金项目： 国家社会科学基金资助项目（23BTJ057）
作者简介： 陈怡君(201907034@xaau.edu.cn)，西安航空学院副研究馆员. 研究方向：数据分析、知识图谱，以及推荐系统.
引　　文： 陈怡君，李欣雨，王潇逸，等. 基于盈余公告漂移的LGBM多因子量化策略[J]. 深圳大学学报理工版，2024，41(3)：313-322.

参考文献 / References：

[1] 李占军. 量化投资在我国投资市场中的应用[J]. 投资与合作，2022(11)：22-24.
LI Zhanjun. The application of quantitative investment in China′s investment market [J]. Investment and Cooperation, 2022 (11): 22-24. (in Chinese)

[2] SHARPE W F. Capital asset prices: a theory of market equilibrium under conditions of risk [J]. The Journal of Finance, 1964, 19(3): 425-442.

[3] ROSS S A. The arbitrage theory of capital asset pricing [J]. Journal of Economic Theory, 1976, 13(3): 341-360.

[4] NOVY-MARX R. The other side of value: the gross profitability premium [J]. Journal of Financial Economics, 2013, 108(1): 1-28.

[5] YAO Haixiang, XIA Shenghao, LIU Hao. Six-factor asset pricing and portfolio investment via deep learning: evidence from Chinese stock market [J]. Pacific-Basin Finance Journal, 2022, 76: 101886.

[6] 刘宇轩，金伟泽，袁亮. 多因子量化选股模型优化与实证研究——引入金融周期指标的分析[J]. 价格理论与实践，2022(4)：141-145.
LIU Yuxuan, JIN Weize, YUAN Liang. Optimization and empirical research on multi-factor quantitative stock selection model: introduce the analysis of financial cycle indicators [J]. Price: Theory & Practice, 2022 (4): 141-145. (in Chinese)

[7] HOU Kewei, XUE Chen, ZHANG Lu. Replicating anomalies [J]. The Review of Financial Studies, 2020, 33(5): 2019-2133.

[8] 冯海涵，马哲坤，吴一夫. 中国A股主板市场PEAD实证研究[J]. 当代经济，2015(7)：10-16.
FENG Haihan, MA Zhekun, WU Yifu. Empirical study on PEAD in China′s A-share main board market [J]. Contemporary Economics, 2015(7): 10-16. (in Chinese)

[9] BALL R, BROWN P. An empirical evaluation of accounting income numbers [J]. Journal of Accounting Research, 1968, 6 (2):159-178.

[10] GRIFFIN J M, KELLY P J, NARDARI F. Do market efficiency measures yield correct inferences? A comparison of developed and emerging markets [J]. The Review of Financial Studies, 2010,23(8):3225-3277.

[11] HUNG M, LI Xi, WANG Shiheng. Post-earnings-announcement drift in global markets: evidence from an information shock [J]. The Review of Financial Studies, 2015, 28(4):1242-1283.

[12] PIOTROSKI J D. Value investing: the use of historical financial statement information to separate winners from losers [J]. Journal of Accounting Research, 2000, 38: 1-41.

[13] 王春丽，刘光，王齐. 多因子量化选股模型与择时策略[J]. 东北财经大学学报，2018，19(5)：81-87.
WANG Chunli, LIU Guang, WANG Qi. Research on multi-factor quantitative stock selection model and timing strategy [J]. Journal of Dongbei University of Finance and Economics, 2018, 19(5): 81-87. (in Chinese)

[14] 侯晓辉，王博. 基于基本面分析的量化投资：研究述评与展望[J]. 东北师大学报哲学社会科学版，2021(1)：124-131，141.
HOU Xiaohui, WANG Bo. Review and prospect of quantitative investment based on analysis of fundamental plane [J]. Journal of Northeast Normal University Philosophy and Social Sciences, 2021(1): 124-131, 141. (in Chinese)

[15] 邓晶，李路. 基于TWSVM的核函数评估及其在量化投资中的应用[J]. 计算机应用与软件，2022，39(10)：87-93.
DENG Jing, LI Lu. Kernel function evaluation based on TWSVM and its application in quantitative investment [J]. Computer Applications and Software, 2022, 39(10): 87-93. (in Chinese)

[16] 任君，王建华，王传美，等. 基于正则化LSTM模型的股票指数预测[J]. 计算机应用与软件，2018，35(4)：44-48，108.
REN Jun, WANG Jianhua, WANG Chuanmei, et al. Stock index forecast based on regularized LSTM model [J]. Computer Applications and Software, 2018, 35(4): 44-48, 108. (in Chinese)

[17] 王淑燕，曹正凤，陈铭芷. 随机森林在量化选股中的应用研究[J]. 运筹与管理，2016，25(3)：163-168，177.
WANG Shuyan, CAO Zhengfeng, CHEN Mingzhi. Research on application of random forests in the quantitative stock selection model [J]. Operations Research and Management Science, 2016, 25(3): 163-168, 177. (in Chinese)

[18] 王燕，郭元凯. 改进的XGBoost模型在股票预测中的应用[J]. 计算机工程与应用，2019，55(20)：202-207.
WANG Yan, GUO Yuankai. Application of improved XGBoost model in stock forecasting [J]. Computer Engineering and Applications, 2019, 55(20): 202-207. (in Chinese)

[19] 张虎，沈寒蕾，刘晔诚. 基于自注意力神经网络的多因子量化选股问题研究[J]. 数理统计与管理，2020，39(3)：556-570.
ZHANG Hu, SHEN Hanlei, LIU Yecheng. The study on multi-factor quantitative stock selection based on self-attention neural network [J]. Journal of Applied Statistics and Management, 2020, 39(3): 556-570. (in Chinese)

[20] PAIVA F D, CARDOSO R T N, HANAOKA G P, et al. Decision-making for financial trading: a fusion approach of machine learning and portfolio selection [J]. Expert Systems with Applications, 2019, 115: 635-655.

[21] 张宁，石鸿伟，郑朗，等. 基于PCANet的价值成长多因子选股模型[J]. 计算机科学，2020，47(增刊2)：64-67.
ZHANG Ning, SHI Hongwei, ZHENG Lang, et al. PCANet-based multi-factor stock selection model for value growth [J]. Computer Science, 2020, 47(Suppl. 2): 64-67.

[22] KE Guolin, MENG Qi, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree [C]// The 31st International Conference on Neural Information Processing Systems. New York, USA: ACM, 2017: hal-03953007.

[23] LI Zimo, XU Weijia, LI Aihua. Research on multi factor stock selection model based on LightGBM and Bayesian optimization [J]. Procedia Computer Science, 2022, 214: 1234-1240.

[24] FINK J. A review of the post-earnings-announcement drift [J]. Journal of Behavioral and Experimental Finance, 2021, 29: 100446.

[25] ZHOU Haigang, ZHU Qi. Jump on the post-earnings announcement drift (corrected) [J]. Financial Analysts Journal, 2012, 68(3): 63-80.

[26] BHUSHAN R. An informational efficiency perspective on the post-earnings announcement drift [J]. Journal of Accounting and Economics, 1994, 18(1): 45-65.

[27] LIU Dehong, GU Hongmei, LUNG P. The equity mispricing: evidence from China′s stock market [J]. Pacific-Basin Finance Journal, 2016, 39: 211-223.

[28] 赵宣凯，何宇. 黎明前的疯狂：盈余公告前的投机泡沫——基于异常交易量的视角[J]. 会计研究，2021(10)：28-42.
ZHAO Xuankai, HE Yu. Pre-dawn madness: speculative bubbles before earnings announcements: based on the perspective of abnormal trading volume [J]. Accounting Research, 2021 (10): 28-42. (in Chinese)

[29] BARTOV E, RADHAKRISHNAN S, KRINSKY I. Investor sophistication and patterns in stock returns after earnings announcements [J]. The Accounting Review, 2000, 75(1): 43-63.

[30] CAI Guilong, LIN Bingxuan, WEI Minghai, et al. The role of institutional investors in post-earnings announcement drift: evidence from China [J]. Accounting and Business Research, 2021, 51(2): 206-236.

【中文责编：英子；英文责编：木柯】