第4卷 第2期 计 量 经 济 学 报 Vol. 4, No. 2
2024年3月 China Journal of Econometrics Mar., 2024

doi: 10.12012/CJoE2022-0047

Stacking 框架下 Boosting 算法融合模型量化投资策略设计及应用

陈创练 [1,2], 邹湘妮 [1]

(1. 暨南大学经济学院, 广州 510632; 2. 暨南大学南方高等金融研究院, 广州 510632)

摘 要 金融市场瞬息万变, 对于量化投资策略, 需要及时地调整和优化. 融合模型可以根据市场变化动态调整模型的权重和组合方式, 实现自适应的调整和优化. 基于此, 本文尝试从融合模型的角度来设计量化投资策略. 本文基于 LightGBM、Adaboost、XGBoost 三种不同的 Boosting 类算法构造了三个不同的融合两层 Stacking 模型, 通过沪深 300 成分股上进行选股回测实证分析来对比择优来得到最合适的基学习模型和次级模型最佳的融合效果. 实证结果表明, 三种不同的融合模型在股票市场的预测表现均优于单一算法模型, 其中表现最为优异的是基学习器为 XGBoost 和 LightGBM 算法, AdaBoost 算法作为次级学习器的融合模型. 在持仓数量为 20 只时, 平均年化收益为 13.57%, 夏普比率为 1.23, 最大回撤为 0.48. 此外该回测结果表明, 融合模型在市场波动性较大时有更好的适应性和有效性. 本研究能够为投资者提供一种新的投资思路, 也对如何推动融合模型在金融实践中运用具备一定的启示意义.

关键词 Boosting 算法; 多因子选股模型; Stacking 融合方法; 量化投资策略

Design and Application of Quantitative Investment Strategies for Boosting Algorithm Fusion Models under Stacking Framework

CHEN Chuanglian[1,2], ZOU Xiangni[1]

(1. School of Economics, Jinan University, Guangzhou 510632, China; 2. Southern China Institute of Finance, Jinan University, Guangzhou 510632, China)

Abstract The financial market is changing rapidly, and quantitative investment strategies need to be adjusted and optimized in a timely manner. The fusion model

收稿日期: 2022-05-17
基金项目: 国家自然科学基金面上项目 (72071094)
Supported by National Natural Science Foundation of China (72071094)
作者简介: 陈创练, 博士, 教授, 研究方向: 量化投资与金融风险管理, E-mail: chenchuanglian@aliyun.com; 邹湘妮, 硕士, 研究方向: 量化投资, E-mail: zouxiangni00@163.com.

can dynamically adjust the weights of the model and the way of combination according to the market changes to achieve adaptive adjustment and optimization. In this paper, three different fusion two-layer Stacking models are constructed based on three different Boosting class algorithms, namely LightGBM, Adaboost, and XGBoost, and empirical analyses of stock picking backtesting are carried out on CSI 300 constituent stocks to compare and select the most suitable base learning model and the best fusion effect of the secondary model. The empirical results show that the three different fusion models outperform the single-algorithm models in stock market prediction, with the best performance being the fusion model where the base learners are the XGBoost and LightGBM algorithms, and the AdaBoost algorithm is used as the secondary learner. When the number of holdings is 20, the average annualized income is 13.57%, the Shape ratio is 1.23, and the maximum retracement is 0.48. In addition, the results of this backtest show that the fusion model has better adaptability and effectiveness in times of high market volatility. This study can provide investors with a new way of thinking about investment, and also provides some insights into how to promote the use of fusion models in financial practice.

Keywords Boosting algorithm; multi-factor stock selection model; Stacking; quantitative investment strategy

1 引言

近些年, 机器学习算法和量化投资领域的多因子选股模型常被用来结合来解决投资决策问题 (石荣等, 2023). 尽管机器学习算法可以发现微妙的、上下文相关的和非线性的关系, 但当人们试图从嘈杂的历史数据中提取信号时, 过度拟合会带来重大挑战 (Rahman et al., 2023). Boosting 算法 (Freund et al., 1995) 作为机器学习中一种流行的统计学习技术, 能够更好地应对数据的噪声和复杂性, 提高预测结果的稳定性和准确度.

Boosting 算法主要包括 AdaBoost (Friedman et al., 2000)、XGBoost (Chen et al., 2016) 和 LightGBM (Ke et al., 2017) 三种不同类型, 近年来很多研究者将这三种算法用于超额回报预测、因子模型建模和多因子投资组合管理等领域, 取得了良好的效果. 王燕等 (2019) 对 XGBoost 模型进行参数优化, 并将该模型运用于股票短期预测中, 并获得较好的结果. Ma (2020) 首次基于 LightGBM 模型设计投资策略, 为指数投资提供框架提供了新思路. 研究发现在获得相同收益的情况下, GBDT 模型的风险低于 LightGBM 模型, 但是 LightGBM 模型的准确率高于 GBDT. Chuan et al. (2021) 通过对中国市场的实证研究将 AdaBoost 应用于投资组合管理, 证明了 AdaBoost 是一个成功的分类器, 能对股票实现上涨下跌分类. Li et al. (2022) 使用随机森林、XGBoost 和 LightGBM 对多因素进行滚动测试, 发现 LightGBM 模型效果最好. Dezhkam et al. (2023) 将极端梯度提升 (XGBoost) 作为收盘价趋势分类器, 回测过程表明, 即使市场表现不佳, 该策略也优于基准策略. 但单一机器学习模型存在精确度不高、过拟合、泛化能力差等问题. 多模型融合是通过取各个基模型所长, 结合多个基学习器的预测结果来提高机器学习模型的准确性, 能够有效解决单一模型的缺陷. 但在量化投资领域, 多种机器学习算法融合选股方面的研究论文相对较少. 李斌等 (2019) 指出将机器学习算法引入量化投资领域, 有助于促进人工智能与经济管理学科的交叉融合, 为推进国家人工

智能发展战略提供重要参考. 因此本文通过运用 Boosting 算法中的 AdaBoost、XGBoost 和 LightGBM 三种机器学习算法作为基学习器来构建融合模型, 来提升机器学习在量化选股中的应用效果. 若本文提出的基于融合模型提出的多因子策略, 能在回测验证中获得稳定的收益, 也将为针对 A 股市场量化投资的理论研究和实践提供参考和借鉴.

2 理论基础

2.1 Boosting 算法理论介绍

本文研究对象为 Boosting 算法 (提升算法), 主要思想是通过串行地训练一系列弱分类器, 来提高整体分类器的准确度, 具体算法原理见图 1. 相比于单一的分类器模型, Boosting 算法能够更好地应对数据的噪声和复杂性, 提高预测结果的稳定性和准确度.

Boosting 是一类算法的统称, 有多个不同的变种, 目前较为流行的三类分别为 AdaBoost、XGBoost 和 LightGBM, 每个变种都有其特殊的优点和缺点, 可以根据不同问题的需求进行选择, 下文将详细介绍这三类算法.

2.1.1 AdaBoost (adaptive boosting)

AdaBoost 是一种经典的梯度提升树算法, 在训练过程中逐步提升弱分类器的效果. AdaBoost 通过给错误分类样本增加权重, 然后重新训练弱分类器, 得到新的权重, 循环迭代, 最终将多个弱分类器组合成强分类器. 所有弱分类器的权值之和并不为 1, 是通过最后结果的符号来决定实例的类别, 该结果的绝对值表示分类的确信度.

在分类任务中, 输入为样本集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, 输出为 -1, $+1$, 弱分类器算法, 弱分类器迭代次数 T. 输出为最终的强分类器 $H(x)$. 步骤如下:

首先初始化所有样本的权重为:

$$D(1)=(\omega_{11},\omega_{12},\cdots,\omega_{1m});\quad \omega_{1i}=\frac{1}{m};\quad i=1,2,\cdots,m. \tag{1}$$

对于每一轮迭代:

第一步, 使用具有权重 D_t 的样本集作为训练数据, 得到弱分类器 h_t.

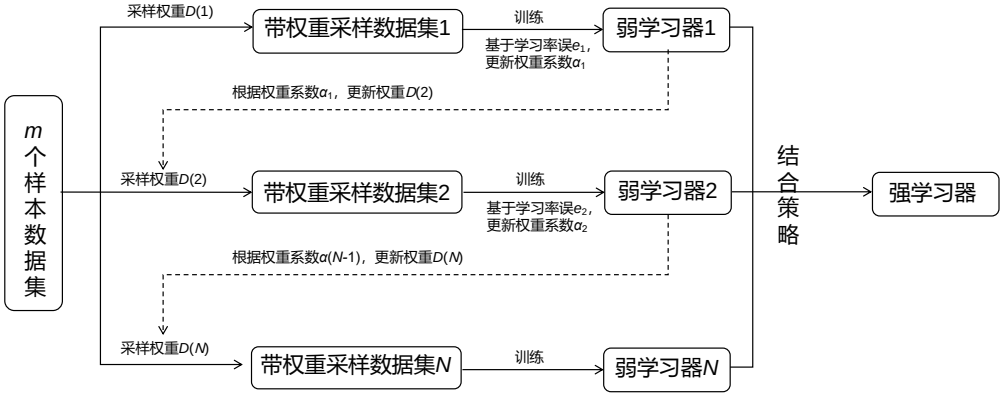

图 1　Boosting 算法原理图

第二步, 计算弱分类器 h_t 的分类误差率 e_t:

$$e_t = P(h_t(x_i) \neq y_i). \tag{2}$$

第三步, 计算弱分类器的权重 α_t:

$$\alpha_t = \frac{1}{2}\log\frac{1-e_t}{e_t}. \tag{3}$$

第四步, 更新样本集的权重分布:

$$\omega_{t+1,i} = \frac{\omega_{ti}}{Z_t}\exp(-\alpha_t y_i h_t(x_i)), \quad i = 1, 2, \cdots, m, \tag{4}$$

其中 Z_t 是规范化因子 $Z_t = \sum_{i=1}^{m}\omega_{ti}\exp(-\alpha_t y_i h_t(x_i))$.

第五步, 构建最终的分类器为:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right). \tag{5}$$

它的主要优点是容易实现, 不容易产生过拟合现象, 能够处理高维数据和离散数据. 然而, 该算法可能对噪声和异常值敏感, 会对迭代过程产生影响.

2.1.2 XGBoost (extreme gradient boosting)

XGBoost 是一种高效的梯度提升树算法, 通过并行计算决策树节点分裂时的增益来加速训练过程, 并采用正则化方法防止过拟合, 也支持对缺失值进行处理. XGBoost 属于一种前向迭代模型, 会训练多棵树, 对于第 t 棵树, 第 i 个样本, 模型的预测值为:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \tag{6}$$

进一步可以得到原始目标函数为:

$$\sum_{i=1}^{m} l(y_i, \hat{y}_i) + \sum_{j=1}^{t}\Omega(f_j). \tag{7}$$

其中 $l(y_i, \hat{y}_i)$ 表示目标值与真实值之间的训练误差, 即模型的损失函数, 后一项是全部 t 棵树的复杂度求和.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2. \tag{8}$$

在 XGBoost 中把它当成是函数中的正则化项, 防止回归树模型复杂度过高, 其中 γ 和 λ 为超参数, 分别控制叶子节点个数和叶子节点权重不会过高.

采用 Boosting 方法训练, 每一次都是在保留原有模型的基础上, 添加一个新函数到模型中, 尽可能让目标函数最大程度的减小, 于是将目标函数 (7) 进一步转化为:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{j=1}^{t}\Omega(f_j) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C. \tag{9}$$

从而可以通过优化 f_t 来优化整个目标函数, 通过泰勒公式对目标函数化简, 将常数项抽离出来, 原有的损失函数进行泰勒二阶展开的结果为:

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) = l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i), \tag{10}$$

其中 g_i 对应损失函数的一阶导数, h_i 对应二阶导数. 进而目标函数可以展开为:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + C. \tag{11}$$

它的优点是目标函数用二阶泰勒展开, 不仅准确率高, 而且能够处理大量数据和复杂特征. 同时利用正则化技术可以避免过拟合, 支持并行计算. 缺点是需要精心调整超参数, 较难解释模型预测结果.

2.1.3 LightGBM (light gradient boosting machine)

LightGBM 算法一种基于梯度提升数算法框架的快速、高效和分布式的 Boosting 算法, 其设计目标是让训练过程更快, 同时保持与其他梯度提升树算法相同的准确性. 核心思想是在每一次迭代中找到当前需要调整的样本, 并将其调整为某个特定的排序位置. 相比其他的 Boosting 算法, 主要的创新点在于使用了基于直方图的决策树算法、带深度限制的 Leaf-wise 的叶子生长策略、基于梯度的单边采样, 这可以极大地降低内存和时间消耗.

2.1.4 算法比较

在 Boosting 算法中基本都使用二叉树作为基础, 使用二叉树的最大问题在于必须考虑切分点. 因此寻找切分点的算法是 Boosting 各种算法之间的主要差异. Adaboost 算法的核心思想是以树桩为基分类器, 每轮迭代后, 增加被错分的数据被抽中的概率, 提高分错样本的权重, 最后将各个基分类器线性加权组合, 只能用于分类问题. XGBoost 和 LightGBM 算法都采用了回归树的方式. 其中 XGBoost 是通过加入了正则项来防止过拟合, LightGBM 则是在 XGBoost 基础上做了进一步的优化, 在精度和速度上都有各自的优点.

2.2 模型融合的理论及其应用

模型融合是指通过将若干多个不同算法组合起来, 以提高整体的准确性和可靠性. 它的基本思想是将多个模型的优点相结合, 从而提高系统的整体性能. 相较于市场组合、等权重组合, 融合模型能够获得更优的投资绩效, 且能在一定程度上抵御风险 (姚海祥等, 2023).

融合方式上可以采用投票、平均值、加权平均值等方式进行, 也可以使用更复杂的多层组合模型, 如 Stacking (田利辉等, 2014) 模型、Blending (田利辉等, 2014) 模型等. Stacking 方法的思想是用初始模型训练出若干个基学习器, 然后将基学习器的结果作为新的特征进行建模再训练一个新的学习器. 它在本质上是一种分层结构, 通常结合 K 折交叉验证进行训练或者在次级学习器的选择上使用简单的线性模型. 考虑到在实际操作中, Stacking 具有更好的性能和更精确的结果. 从机器学习在量化交易领域的应用来看, 单一模型在量化领域获取超额收益的能力总体有限, 而 Stacking 模型作为一种集成算法, 能够有效提升量化交易过程中的总体收益.

关于 Stacking 融合方法对 Boosting 算法研究上, 一些学者应用 Stacking 方法进行了量化投资策略研究. 李佩琛 (2018) 利用 Stacking 集成学习思想将随机森林、GBDT、SVM、Adaboost、K 邻近、决策树这六个算法进行组合, 构建了一个复杂有效的机器学习模型. 用训练好的模型对沪深 300 的成分股实证分析, 发现集成后的模型无论在收益指标还是风险指标上都有了超越子模型的良好表现. Ribeirod et al. (2020) 使用 Stacking 算法将随机森林、XGBoost 和 LightGBM 等多种算法的结果进行融合, 获得了更加准确的股价预测结果. 岳腾强 (2022) 使用 XGBoost 和 LightGBM 两种机器学习树模型构建基础学习器, 借助 Stacking 方式进行融合以获得更强的模型融合强学习器, 结果表明模型融合策略可以更大发挥机器学习算法的预测优势. 丁卿纯 (2023) 通过融合 XGBoost 算法、Light GBM 等多种不同机器学习模型优缺点, 尝试探索不同模型能否克服对方的缺点, 融合成的新模型相比单一模型更具有超额收益, 能够有效战胜市场.

另外, 还有一些研究集中于提出新的 Stacking 变体, 以更好地适应量化投资的需求. 比如, 一些学者提出了基于改进的 Stacking 结构的投资策略, 其主要思路是通过对不同层之间的权重进行优化或者参数优化方式改进, 从而进一步提高 Stacking 的融合效果. 周申豪 (2022) 本文从 Bagging、Boosting 和 Stacking 三种集成策略中分别选 1 至 2 种算法, 共 4 种集成算法. 结果表明 4 种集成算法在长期中均有较好的选股能力, RXL-Stacking 算法的选股能力最佳, 即第一层为随机森林算法, 第二层使用 XGBoost 和 Light GBM 算法. 同时在选取 10 只股票时, 为最佳投资组合. 钟正豪 (2022) 采用了 Stacking 两层融合的方式, 采用随机森林、XGBoost 算法作为基学习器进行预测, 再将结果输入到 LightGBM 算法构成的次级学习器, 通过对不同层的权重优化, 得到最终模型, 并在对沪深 300 成分股实证分析中验证了该融合模型能够获得更高的超额收益. 陈雨芳 (2022) 借助 LR、DT、XGBoost 等机器学习算法, 采用 Voting 和 Stacking 的融合方法构建融合模型, 发现融合后的模型精确率更高.

总之, Stacking 融合模型能够应用于各种进行预测和分类的场景, 通过综合多种算法的结果, 提高了预测结果的准确性和鲁棒性. 但是 Boosting 算法在量化投资领域而言, 大部分的研究集中在选取 Boosting 算法中某一类代表性算法和其他的机器学习算法进行 Stacking 融合. 采用具有相似特性的 AdaBoost、XGBoost 和 LightGBM 三类算法共同融合 Stacking 模型在量化投资领域中运用还没有学者尝试, 这也是本文的主要创新思路.

3 基于 Stacking 框架的融合模型设计

3.1 融合策略设计

本文使用两层 Stacking 方案, 即进行两个阶段的训练: 第一阶段是为每个基模型单独训练; 第二阶段是使用基模型的预测结果进行次级模型训练. 此外, 还需要对模型进行交叉验证和调参等操作. 模型融合过程如图 2 所示.

本文从 AdaBoost、XGBoos、LightGBM 三种模型中任意选取两种为第一层基学习模型, 剩下的一种为第二层的次级学习模型. 具体来说, 将第一层使用 LightGBM 模型和 XGBoost 模型作为基学习模型进行预测, 并将预测结果输入到第二层的 AdaBoost 次级模型中进行最终预测, 即为 Stacking-Ada 模型; 将第一层使用 LightGBM 模型和 AdaBoost 模型进行预

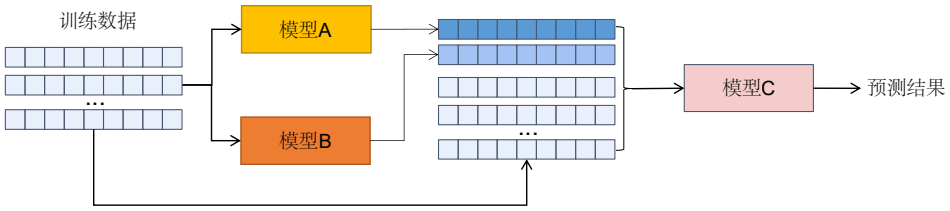

图 2　融合过程

测, 并将结果输入到第二层的 XGBoost 模型中进行最终预测, 即为 Stacking-XGB 模型; 第一层使用 AdaBoost 模型和 XGBoost 模型进行预测, 并将结果输入到第二层的 LightGBM 模型中进行最终预测, 即为 Stacking-LGB 模型.

3.2 因子体系构建

本文选择沪深 300 指数的成分股作为研究对象, 数据的时间跨度为 2011 年至 2021 年共 11 年 132 个月的月频数据. 本文数据来源为万得金融数据库, 选取了沪深 300 指数的所有成分股及其 74 个因子数据, 具体见附录. 对预处理后的数据采用 IC 值检验来验证因子的有效性. IC (Information Coefficient) 是指信息系数, 代表因子值与股票下期收益率的截面相关性. IC 值的绝对值大于 0.05 时, 认为因子有效. 因此本文以 0.05 作为临界值筛选得到 43 个因子. 因子筛选结果见表 1 所示, 并在后续实证分析环节使用这些因子的数据.

表 1　筛选后得到的因子池

编号	因子名称	因子描述
1	PE	市盈率
2	PS	市销率
3	VAL_PBINDU_SW	(个股市净率 – 行业 PB 均值)/PB 行业标准差
4	DIVIDEND YIELD2	股息率 (12 个月)
5	RISK_VARIANCE20	20 日收益方差
6	RISK_VARIANCE60	60 日收益方差
7	RISK_LOSSVARIANCE20	20 日损失方差
8	RISK_LOSSVARIANCE60	60 日损失方差
9	RISK_CUMRETURN12M	12 月累计收益
10	FA_ROE_AVG	净资产收益率
11	FA_GROSSPROFITMARGIN_TTM	销售毛利率
12	FA_ROA_TTM	资产回报率
13	FA_OCFPS_TTM	每股经营活动产生的现金流量净额 TTM
14	FA_OPPS_TTM	每股营业利润
15	FA_PBTTOOR_TTM	利润总额/营业收入
16	FA_CURASSETSRATIO	流动资产比率
17	FA_TATURN_TTM	总资产周转率
18	FA_ARTURNDAYS_TTM	应收账款周转天数
19	FA_ORGR_TTM	增长率 – 营业收入
20	TECH_TURNOVERRATE5	5 日平均换手率

表 1 (续)

编号	因子名称	因子描述
21	TECH_TURNOVERRATE10	10 日平均换手率
22	TECH_TVMA6	6 日成交金额的移动平均值
23	TECH_TVMA20	20 日成交金额的移动平均值
24	TECH_TVSTD6	6 日成交金额的标准差
25	TECH_TVSTD20	20 日成交金额的标准差
26	TECH_OBV	能量潮指标
27	TECH_OBV6	6 日能量潮指标
28	TECH_ATR6	6 日均幅指标
29	TECH_ATR14	14 日均幅指标
30	TECH_EMA5	5 日指数移动均线
31	TECH_EMA10	10 日指数移动均线
32	TECH_BOLLUP	上轨线
33	FA_GP_TTM	毛利
34	FA_PROFIT_TTM	净利润
35	FA_GR_TTM	营业总收入
36	FA_OR_TTM	营业收入
37	FA_EBITTOGR_TTM	息税前利润/营业总收入
38	FA_BPS	每股净资产
39	FA_EPS_BASIC	基本每股收益
40	TECH_RSI	相对强弱指标
41	VAL_LNFLOATMV	对数流通市值
42	FA_OPTOGR_TTM	营业利润/营业总收入
43	FA_OCTOGR_TTM	营业总成本/营业总收入

3.3 选股模型设计

金融数据是时序数据的一种, 因子的有效性随着时间的推移会受到许多因素的影响, 因此需要选择合适的建模方案. 在训练过程中需要考虑模型对于数据的处理能力与数据总量的关系, 且考虑到这段时间内股票的起伏较大, 因此本文使用滚动建模的方法对时间序列的股票数据进行建模处理, 选择前六个月的数据作为训练数据, 后一个月的数据作为预测数据, 这样便完成了一个周期的建模. 然后将窗口向后滑动一个月, 进行下一个周期的建模处理. 具体方法见图 3 所示.

本策略通过滚动训练来训练模型并进行测试, 即根据当前截面的因子数据, 预测下一个交易日的股票价格, 最后将优化好的模型用来做股票回测, 验证方案策略的有效性. 具体方案如下:

1) 初始资金: 选择进行股票回测的初始资金为 100 万元;

2) 选择股票: 以前六个月作为训练集, 第七个月作为测试集滚动训练和预测, 获得当日上涨前 30% 的为当月候选清单, 按照收益率预测值从高到低排序;

3) 调仓: 每月调仓一次, 更新投资组合时对于不再需要持有的股票全仓卖出, 对于需要持仓的股票进行等额买入;

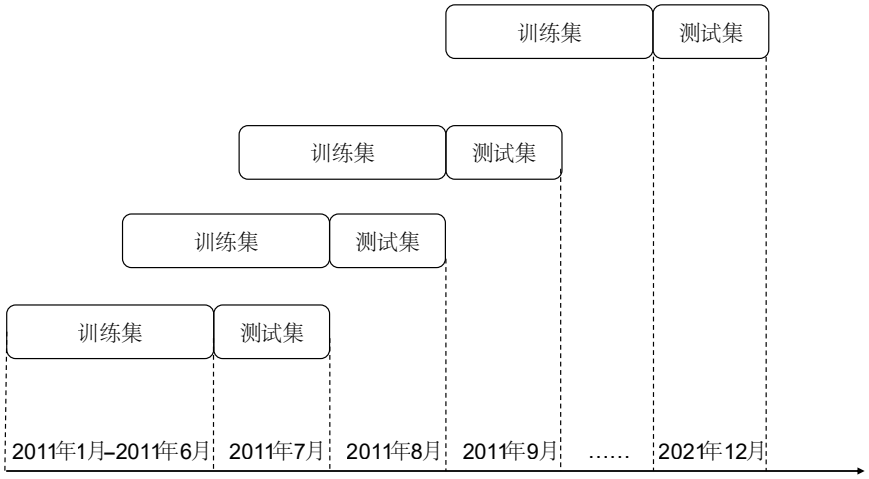

图 3　滚动建模过程

4) 建仓: 选择四种建仓方式, 在每一个运行周期内, 选择候选清单中前 k 只股票在本月持仓, $k \in \{5,\ 10,\ 20,\ 30\}$. 调仓时间为每个周期内第一个交易日, 按照上个月末收盘价进行交易, 对候选股票池的股票是等权分配资金;

5) 股票交易过程中可以购买任意数量的股票, 不设置最小交易量和交易单位;

6) 买入股票前先对停牌、ST 类股票和退市的股票进行剔除, 并过滤掉涨停跌停的股票;

7) 每笔交易时的手续费为佣金的万分之三, 在卖出股票时加收千分之一印花税;

8) 投资组合收益是通过每个交易月持仓股票的月末收益率等权重相加得到.

3.4 评价体系构建

本文将从模型评价、策略评价两方面构建评价体系. 本文建立了在分类问题的模型, 对实际中排名前 30% 的模型定义为正例. 模型评价指标有精确率、召回率、准确率、F1 分数. 投资策略评价指标中, 采用总收益率、平均年化收益率和超额收益率衡量模型策略的收益, 采用夏普比率、最大回撤两个指标度量模型的风险性.

4 实证检验

4.1 参数调优

本文对 Boosting 算法中的 AdaBoost、XGBoost 和 LightGBM 模型使用 Optuna 框架进行参数优化, Optuna 将每一个优化过程作为研究对象, 通过一定数量的独立采样之后识别不同超参数之间潜在的共现关系, 然后进行超参数选取, 同时会进行高效的并行计算以及检测计算过程不符合预定条件的试验, 大大提高了参数选择的时间效率.

4.2 基模型实证分析

本文对参数优化后的三个基础模型 (AdaBoost、XGBoost 和 LightGBM 算法模型) 来进行选股回测实证分析, 滚动训练长度为六个月, 持仓数量分别为 5 只、10 只、20 只、30 只, 得到结果如表 2 所示.

表 2 基学习模型的选股回测指标结果

模型	持仓数量	总收益率	年化收益率	超额收益率	夏普比率	最大回撤
AdaBoost	5	233.2%	11.56%	181.34%	0.4983	0.7872
	10	222.93%	11.25%	171.07%	0.7239	0.7135
	20	175.52%	9.65%	123.66%	0.8827	0.6118
	30	172.28%	9.53%	120.42%	0.9882	0.5235
XGBoost	5	119.65%	7.42%	67.79%	0.4485	0.7522
	10	127.83%	7.77%	75.97%	0.532	0.6597
	20	151.5%	8.75%	99.64%	0.7893	0.5706
	30	196.57%	10.39%	144.71%	1.0849	0.5419
LightGBM	5	269.86%	12.63%	218%	0.6312	0.7655
	10	173.3%	10.13%	121.44%	0.7632	0.7356
	20	157.73%	8.99%	105.73%	0.8321	0.6228
	30	155.75%	8.91%	103.89%	0.9482	0.5537
Base	-	51.86%	3.9%	0	0.5436	0.4285

表 2 中 Base 表示沪深 300 指数基准. 在不同持仓数量下, 训练数据的质量和数量可能会有所不同, 会影响算法的表现. 从表中可以看出, 无论在何种持仓情况下, 三个基学习模型均取得超额收益, 收益率表现效果较好. 由于不同算法的模型结构和优化方法上存在一定差异, 导致在不同持仓情况下表现不同. LightGBM 模型在持股 5 只时达到了 269.86% 的总收益, 218% 的超额收益. 在最大回撤方面, 相较于基准指数有较大的波动, 随着持仓数增多, 最大回撤也逐渐降低.

4.3 Stacking-Boosting 融合模型实证分析

然后本文尝试了三种不同组合的模型融合方案, 即 Stacking-Ada [1]、Stacking-XGB [2]、Stacking-LGB [3]. 简单的说, 就是在第一层任选两种模型进行预测, 再将结果输入到第二层, 再采用剩下的一种模型来进一步预测, 得到的结果如表 3 所示.

根据表 3, 从收益指标来看, 在持仓数量为 5 只、10 只、20 只、30 只的情况下, Stacking-Ada 模型的年化收益均达到了 10% 以上, 且都要高于其他两种融合模型, 远超出基准收益. 从风险指标上看, 最大回撤要低于基准模型, 并且随着持仓数量的增加, 最大回撤会逐渐降低. 在持仓数量为 20 只的时候, Stacking-Ada 模型的收益和风险指标均最优, 累计收益率达到了 306.93%, 超额收益率为 255.07%, 最大回撤为 0.4816.

[1]Stacking-Ada 模型是指在第一层使用 LightGBM 模型和 XGBoost 模型进行预测, 并将结果输入到第二层的 AdaBoost 模型中进行最终预测.

[2]Stacking-XGB 模型是指第一层使用 LightGBM 模型和 AdaBoost 模型进行预测, 并将结果输入到第二层的 XGBoost 模型中进行最终预测.

[3]Stacking-LGB 模型是指第一层使用 AdaBoost 模型和 XGBoost 模型进行预测, 并将结果输入到第二层的 LightGBM 模型中进行最终预测.

表 3　Boosting 融合模型的选股回测指标结果

模型	持仓数量	总收益率	年化收益率	超额收益率	夏普比率	最大回撤
Stacking-Ada	5	306.93%	13.68%	255.07%	0.5996	0.7936
	10	209.38%	10.81%	157.52%	0.6941	0.7107
	20	305.61%	13.57%	253.75%	1.2362	0.4816
	30	220.32%	11.16%	168.46%	1.1596	0.502
Stacking-XGB	5	119.65%	7.42%	67.79%	0.4485	0.7522
	10	127.83%	7.77%	75.97%	0.532	0.6597
	20	151.5%	8.75%	99.64%	0.7893	0.5706
	30	196.57%	10.39%	144.71%	1.0849	0.5419
Stacking-LGB	5	121.76%	7.51%	69.9%	0.4487	0.7561
	10	144.55%	8.47%	92.69%	0.5686	0.7119
	20	119.41%	7.4%	67.55%	0.688	0.6261
	30	138.43%	8.22%	86.57%	0.8599	0.5427
Base	-	51.86%	3.9%	0	0.5436	0.4285

4.4 结果对比分析

4.4.1 基学习器与融合模型预测性能对比分析

在算法运行时间上, 模型训练时间具有一定的差距, 几种算法模型的平均运行时间如下所示. 在基础模型中, 由于 AdaBoost 没有并行操作, 只能串行进行优化, 因此训练时间最长.XGBoost 和 LightGBM 模型都加入了并行操作, 因此在时长上有所优化, 如表 4 所示. 同时在分类效果上, 本文对排名前 30% 的股票进行预测, 对预测结果与真实结果计算准确率和 F1 分数, 分类的指标对比如表 5 所示.

从表中可以看出 AdaBoost 模型在准确率上是最高的, 这是因为该模型在迭代训练过程中以准确率为目标进行模型优化. 而 Stacking-Ada 融合模型在结合了 AdaBoost 准确率高的优点外, 加入了 XGBoost 和 LightGBM 的预测结果作为特征, 所以其稳定性更好, F1 分数较高, 且准确率上也比 XGBoost 和 LightGBM 单模型要好.

表 4　模型训练时间对比值

模型	时长 (秒)
AdaBoost	887
XGBoost	258
LightGBM	67
Stacking-Ada	894

表 5　模型评价指标平均值

模型	准确率	F1 分数
AdaBoost	0.834	0.74
XGBoost	0.773	0.729
LightGBM	0.776	0.733
Stacking-Ada	0.814	0.764

4.4.2 融合模型回测分析

根据上述分析, 选取以持仓五只股票, 以第一层为 LightGBM 和 XGBoost 模型, 第二层选择 AdaBoost 模型的融合模型来进行回测. 后续的 Stacking-Ada 融合模型验证也选择该模型, 得到收益结果如图 4. 同时为保证数据的有效性, 本文选取 2011 年至 2021 年沪深 300 指数收盘价作为原始数据, 结合统计学知识, 将发生小概率事件时刻定义为非平稳时期, 即以

“均值 ±2 倍标准差” 为警戒线来判断股市是否处于稳定. 经测算后收益率均值为 0.56%, 标准差为 0.0642, “均值 +2 倍标准差” 为 13.40%, “均值 −2 倍标准差” 为 − 12.27%. 具体如图 5.

4.5 有效性验证及改进

由图 4 可以看出, 无论是单算法模型还是 Stacking-Ada 融合模型都在 2014 年下半年至 2015 年年底经历了暴涨暴跌现象, 且在 2015 年 6 月 (即牛市顶峰) 达到收益巅峰值. 主要是由于 2014 年 7 月起, 我国股市在经历了 5 年的调整之后, 逐步恢复上升, 继而快速上涨. 到 2015 年 6 月开始, 开始出现了三轮大幅下跌, 直到 2016 年 1 月起, A 股市场引入熔断机制. 从整体收益上来看, Stacking-Ada 融合模型和单一算法模型在不同阶段的收益率都远高于沪深 300 成分股, 说明模型能适应市场环境变化. 同时, 峰值的出现导致后续预测结果就会出现一定偏差, 但相较于单一算法模型, 融合模型在后续预测对偏差的调整也更加及时, 能够在市场中得到更加稳定的效果.

从图 5 可以看出, 2012 年下半年到 2015 年下半年的沪深 300 指数收益率多次突破 “两倍标准差上限” 与跌破 “两倍标准差下限”, 处于动荡时期. 而 2016–2021 年的沪深 300 指数波动处于上限和下限的区间之内, 市场风格相对稳定. 因为量化投资本质上就是根据历史判断未来, 假设的是市场风格会继续演绎, 若出现股灾这类突发事件, 市场风格可能会在某个区间的训练样本中发生改变, 模型在训练时就会产生混乱, 预测的准确性自然也就下降了, 所以该段时间的训练得到的 Stacking-Ada 融合模型有效性可能无法准确评估. 所以本文重新选取 2016–2021 年的数据来对 Stacking-Ada 模型训练并进行有效性验证. 具体如下:

从图 5 可以看出, 2012 年下半年到 2015 年下半年的沪深 300 指数收益率多次突破 “两倍标准差上限” 与跌破 “两倍标准差下限”, 处于动荡时期. 而 2016–2021 年的沪深 300 指数波动处于上限和下限的区间之内, 市场风格相对稳定. 因为量化投资本质上就是根据历史判断未来, 假设的是市场风格会继续演绎, 若出现股灾这类突发事件, 市场风格可能会在某个区间的训练样本中发生改变, 模型在训练时就会产生混乱, 预测的准确性自然也就下降了, 所以该段时间的训练得到的 Stacking-Ada 融合模型有效性可能无法准确评估. 所以本文重新选取

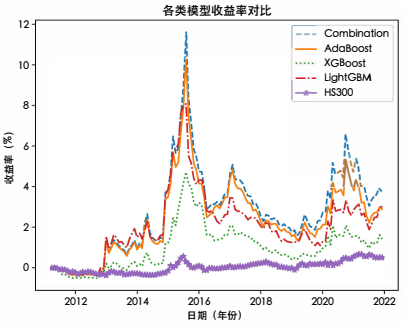

图 4 模型收益率比较

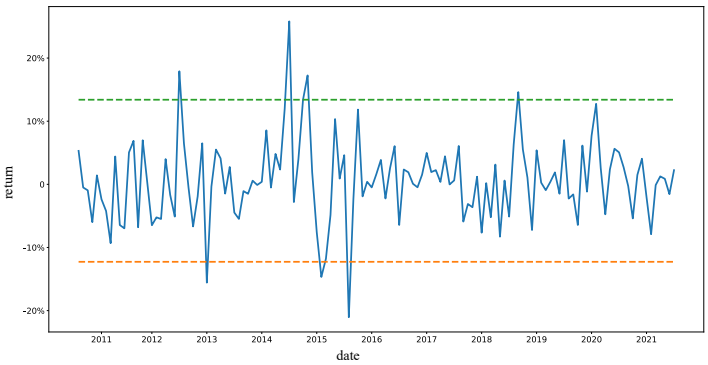

图 5 2011–2021 年沪深 300 走势

2016–2021 年的数据来对 Stacking-Ada 模型训练并进行有效性验证. 具体如下:

本文首先重新对模型进行参数优化, 再用滚动训练方式来得到训练模型, 并在相同的区间上进行选股回测. 得到的结果如表 6 所示. 同上持仓数量分别为 5 只、10 只、20 只、30 只.

上述表格中可以看出, 经过重新训练得到的模型, 相比于原来训练的模型在收益率上有更好的表现. 在持仓数量为 20 只的时候, 达到了最优的风险-收益表现, 即验证了 Stacking-Ada 融合模型有效性. 基础算法模型与集成模型的收益结果如图 6 所示.

然后将训练前后的收益率曲线综合对比, 模型收益如图 7 所示. 图例中的 Combination_N 和 Combination_O 分别表示重新训练后的模型 (数据选取区间为 2016–2021 年) 和原来训练的 Stacking-Ada 模型 (数据选取区间为 2011–2021 年).

从图 7 中可以看出, 无论是在哪段时间内, 重新训练后的模型的表现效果都要比原训练融合模型效果好. 虽然在中间某个阶段重新训练的 AdaBoost 模型效果最好, 但表现并不稳定, 后期收益逐渐下降. 不同的是, Stacking-Ada 融合模型的表现效果较为稳定, 尤其是在最后阶段收益效果较为明显. 由于从 2021 年开始, 受疫情影响股市波动较大, Boosting 算法在预测过程中并没有表现出较好的效果, 准确性有所下降了. 但从在收益率表现来看, 2021 年开始, 该融合模型明显高于三种单一算法模型, 说明该模型在一定程度上能经受住市场的考验.

5 研究结论与启示

本文提出了一种基于 Boosting 算法的多模型融合的量化因子投资策略, 该策略通过结合 AdaBoost、XGBoost 和 LightGBM 这三种 Boosting 代表性算法组合成三个不同的新的 Stacking 融合模型来进行交易策略的优化, 并将该策略应用到沪深 300 成分股中, 从风险、

表 6　重新训练的模型在新区间的回测指标

模型	总收益率	年化收益率	超额收益率	夏普比率	最大回撤
AdaBoost	−1.01%	−0.16%	−40.64%	−0.0394	0.5316
	113.44%	13.47%	73.81%	1.348	0.3188
	129.99%	14.89%	90.36%	1.9241	0.2709
	120.96%	14.13%	81.33%	2.0586	0.2688
XGBoost	7.35%	1.19%	−32.28%	0.3876	0.4215
	90.9%	11.38%	51.27%	1.2464	0.301
	108.56%	13.03%	68.93%	1.786	0.3018
	125.72%	14.53%	86.09%	2.0947	0.3168
LightGBM	5.11%	0.83%	−34.52%	0.3258	0.4485
	38.49%	5.58%	−1.14%	0.5800	0.3046
	143.95%	16.02%	104.32%	2.0291	0.2588
	114.71%	13.58%	75.08%	1.905	0.3022
Stacking-Ada	8.21%	1.32%	−31.42%	0.1176	0.4914
	186.02%	19.14%	146.39%	1.9954	0.2367
	172.8%	18.21%	133.17%	2.4054	0.2324
	140.99%	15.79%	101.36%	2.2857	0.2926
Base	39.63%	5.8%	0	1.0127	0.3006

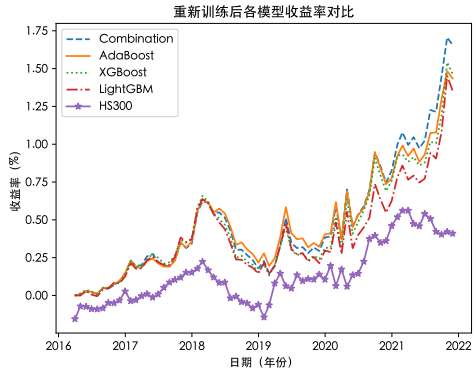

图 6 重新训练的模型在新区间的月收益图

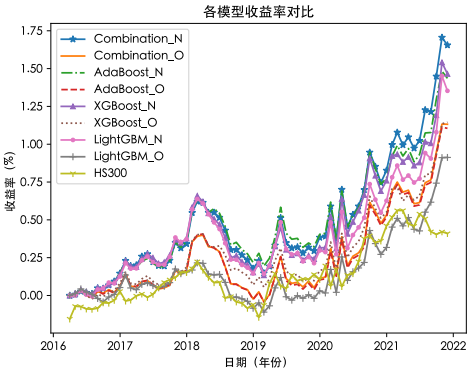

图 7 训练前后模型月收益曲线对比

收益、模型性能这三个维度对比分析, 得到了最优的融合模型为 Stacking-Ada 模型. 并且验证了该模型在不同的回测区间上对股票市场的适用性和有效性. 针对该结论, 作出了如下理论分析:

1) Stacking 的基本原理是将不同的基础模型的预测结果作为新的特征, 再利用另一个模型来对这些新特征进行分类或回归, 从而得到最终的预测结果. 2) LightGBM 和 XGBboost 作为基学习器, 是因为它们都是梯度提升树算法, 在高维稀疏数据上表现较优, 具有高效性和准确性的特点. 而选择 AdaBoost 作为次级学习器, 是因为 AdaBoost 是一种迭代算法, 是通过迭代过程纠偏来减少模型误差. 该算法只能用于分类问题, 在处理分类问题时优于其他两种算法. 在融合模型中, AdaBoost 可以挖掘基学习器之间的相互依赖关系, 并将基学习器的输出结果进行加权平均, 除了保留了模型原来的纠偏能力, 在一定程度上通过加入其他模型的结果降低了模型预测的方差, 因此模型在稳定性和预测能力上都有一定的提升.

同时在研究过程中也得到了如下启示:

1) 本文选取的是沪深 300 指数成分股, 不能代表全部股票市场, 因此该模型不一定能够适用整个交易市场. 而且只选取了 74 个对股票影响较大的因子作为研究对象, 通过筛选后只留下了 43 个因子, 由于企业之间存在差异性, 很可能错过部分对于某些股票影响大的特征因子.

2) 本文提出的融合方法在实际应用中仍然面临很多挑战, 例如如何选择最优的参数、如何解决数据样本不平衡问题、如何处理缺失数据等等. 因此, 在使用这种融合方法时, 需要谨慎地选择合适的算法和参数, 并且要考虑到数据质量和实际投资需求.

3) 融合模型是当前机器学习领域的研究热点, 这一领域仍有很大的发展空间. 如何选择合适的基模型、如何选择次级模型、如何处理过拟合和欠拟合等问题都将会成为研究方向. 也有望出现更多创新性的融合方法. 但需要注意的是, 金融市场交易存在各方面的限制, 应考虑如何去更大限度地将模型贴近真实交易市场, 从而创造实际使用价值.

4) 尽管融合模型有很好的准确性和泛化能力, 但是它相对于单一模型来说, 可能会有更多的波动和不稳定性, 特别是在面对复杂的市场环境时. 股票市场风云诡谲, 投资策略的成功并不取决于单一的技术手段. 在实际应用中需要综合考虑市场环境、投资目标、风险控制等

多个因素来制定完备的投资方案和策略.

总体来说, 在机器学习量化投资中, 选择适合特定问题的算法和模型是一个关键问题, 但如何选择最好的模型并没有一个标准答案, 需要不断优化和探索. 模型融合与量化投资结合的研究仍处于初级阶段, 随着机器学习技术和数据源的不断发展, 基于 Boosting 算法的模型融合未来有望成为量化投资领域的重要研究方向, 并且会得到更加广泛的关注和应用. 还需要更多的研究去优化融合算法, 以提高其预测性能和稳定性.

参　考　文　献

陈雨芳, 赵英杰, 吴昕劼, (2022). 基于融合模型的机器学习算法识别财务造假的研究——以制造业为例 [J]. 统计与管理, 37(7): 116–121.

Chen Y F, Zhao Y J, Wu X J, (2022). A Study on Machine Learning Algorithm Based on Fusion Model to Identify Financial Fraud — A Case Study of Manufacturing Industry[J]. Statistics and Management, 37(7): 116–121.

丁卿纯, (2023). 基于机器学习的多因子量化选股策略设计 [D]. 武汉: 中南财经政法大学.

Ding Q C, (2023). Design of Multi-factor Quantitative Stock Selection Strategy Based on Machine Learning[D]. Wuhan: Zhongnan University of Economics and Law.

李斌, 邵新月, 李玥阳, (2019). 机器学习驱动的基本面量化投资研究 [J]. 中国工业经济, (8): 61–79.

Li B, Shao X Y, Li Y Y, (2019). Machine Learning Driven Quantitative Investment Research on Fundamentals[J]. China Industrial Economy, (8): 61–79.

李佩琛, (2018). 用 stacking 算法堆积随机森林、Gbdt、Svm、Adaboost 等七种算法的多因子选股模型 [D]. 杭州: 浙江工商大学.

Li P C, (2018). Multi-factor Stock Selection Model with Stacking Algorithm Stacking Seven Algorithms Including Random Forest, Gbdt, Svm, and Adaboost[D]. Hangzhou: Zhejiang Gongshang University.

石荣, 张特, 杨国涛, (2024). 计量经济学中的机器学习方法: 回顾与展望 [J]. 统计与决策, (1): 52–56.

Shi R, Zhang T, Yang G T, (2024). Machine Learning Methods in Econometrics: A Review and Outlook[J]. Statistics and Decision Making, (1): 52–56.

田利辉, 王冠英, 张伟, (2014). 三因素模型定价: 中国与美国有何不同?[J]. 国际金融研究, (7): 37–45.

Tian L H, Wang G Y, Zhang W, (2014). Three Factor Model Pricing: How is China Different from the United States?[J] Studies of International Finance, (7): 37–45.

田利辉, 王冠英, (2014). 我国股票定价五因素模型: 交易量如何影响股票收益率?[J]. 南开经济研究, (2): 54–75.

Tian L H, Wang G Y, (2014). A Five-factor Model of Stock Pricing in China: How Does Trading Volume Affect Stock Returns?[J]. Nankai Economic Studies, (2): 54–75.

王燕, 郭元凯, (2019). 改进的 XGBoost 模型在股票预测中的应用 [J]. 计算机工程与应用, 55(20): 202–207.

Wang Y, Guo Y K, (2019). Application of Improved XGBoost Model in Stock Forecasting[J]. Computer Engineering and Applications, 55(20): 202–207.

姚海祥, 李晓鑫, 房勇, (2023). 基于 AdaBoost 集成算法和 Black-Litterman 模型的资产配置 [J]. 系统工程理论与实践, 43(11): 3182–3196.

Yao H X, Li X X, Fang Y, (2023). Asset Allocation Based on ADABOOST Integrated Algorithm and Black-Litterman Model[J]. Systems Engineering — Theory & Practicee, 43(11): 3182–3196.

岳腾强, (2022). 基于机器学习模型融合的中证 800 指数增强量化策略设计 [D]. 武汉: 中南财经政法大学.

Yue T Q, (2022). Design of Enhanced Quantitative Strategy for CSI 800 Index Based on Machine Learning Model Fusion[D]. Wuhan: Zhongnan University of Economics and Law.

钟正豪, (2022). 基于机器学习的 A 股量化交易策略研究 [D]. 成都: 西南财经大学.
Zhong Z H, (2022). Research on A-share Quantitative Trading Strategy Based on Machine Learning[D]. Chengdu: Southwest University of Finance and Economics.

周申豪, (2022). 基于集成算法的多因子量化选股模型研究 [D]. 成都: 西南财经大学.
Zhou S H, (2022). Research on Multi-factor Quantitative Stock Selection Model Based on Integration Algorithm[D]. Chengdu: Southwest University of Finance and Economics.

Chen T, Guestrin C, (2016). Xgboost: A Scalable Tree Boosting System[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785–794.

Chuan Y, Zhao C, He Z, Wu L, (2021). The Success of Adaboost and Its Application in Portfolio Management[J]. International Journal of Financial Engineering, 8(2): 2142001.

Dezhkam A, Manzuri M T, (2023). Forecasting Stock Market for an Efficient Portfolio by Combining XGBoost and Hilbert-Huang Transform[J]. Engineering Applications of Artificial Intelligence, 118: 105626.

Freund Y, Schapire R E, (1995). A Desicion-theoretic Generalization of On-line Learning and an Application to Boosting[C]// European Conference on Computational Learning Theory. Springer Berlin Heidelberg: 23–37.

Friedman J, Hastie T, Tibshirani R, (2000). Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors)[J]. The Annals of Statistics, 28(2): 337–407.

Ke G L, Meng Q, Finley T, Wang T F, Chen W, et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C]// Proceedings of the 31st International Conference on Neural Information Processing SystemsDecember: 3149–3157

Ma S, (2020). Predicting the SP500 Index Trend Based on GBDT and LightGBM Methods[J]. E3S Web of Conferences, 214(5): 02019.

Rahman M J, Zhu H, (2023).Predicting Accounting Fraud Using Imbalanced Ensemble Learning Classifiers — Evidence from China[J]. Accounting & Finance, 63(3): 3455–3486.

Ribeiro M H D M, Dos Santos Coelho L, (2020). Ensemble Approach Based on Bagging, Boosting and Stacking for Short-term Prediction in Agribusiness Time Series[J]. Applied Soft Computing, 86: 105837.

Li Z, Xu W J, Li A, (2022). Research on Multi Factor Stock Selection Model Based on LightGBM and Bayesian Optimization[J]. Procedia Computer Science, 214: 1234–1240.

附录

附录 A 本文研究的候选因子

附表　初始选取的全部因子

因子名称	因子描述	因子名称	因子描述
PE	市盈率	PS	市销率
PCF_NCF	市现率 (现金净流量)	Dividend	股息率 (12 个月)
risk_variance20	20 日收益方差	risk_variance60	60 日收益方差
risk_lossvariance20	20 日损失方差	risk_lossvariance60	60 日损失方差
risk_beta20	20 日 beta 值	risk_beta60	60 日 beta 值
risk_kurtosis20	收益 20 日峰度值	risk_treynorratio20	20 日特雷诺比率
risk_cumreturn12m	12 月累计收益	fa_roe_avg	净资产收益率
fa_grossprofit-margin_ttm	销售毛利率	fa_roa_ttm	资产回报率
fa_pbttoor_ttm	利润总额/营业收入	fa_opps_ttm	每股营业利润
fa_curassetsratio	流动资产比率	fa_equityassetsratio	股东权益比率
fa_current	流动比率	fa_taturn_ttm	总资产周转率
fa_arturndays_ttm	应收账款周转天数	fa_currtasset-strate_ttm	流动资产周转率
fa_orgr_ttm	增长率-营业收入	tech_vr	成交量比率
tech_turnoverrate5	5 日平均换手率	tech_turnoverrate10	10 日平均换手率
tech_turnoverrate20	20 日平均换手率	tech_obv	能量潮指标
tech_turnoverrate-volatility20	换手率相对波动率	tech_obv6	6 日能量潮指标
tech_ATR6	6 日均幅指标	tech_ATR14	14 日均幅指标
tech_EMA5	5 日指数移动均线	tech_EMA10	10 日指数移动均线
tech_revs5	过去 5 日的价格动量	tech_cci10	10 日顺势指标
tech_bias5	5 日乖离率	tech_bias10	10 日乖离率
tech_bollup	上轨线	tech_macd	平滑异同移动平均线
tech_roc6	6 日变动速率	tech_roc20	20 日变动速率
tech_rvi	相对离散指数	tech_mass	梅斯线
fa_gp_ttm	毛利	fa_profit_ttm	净利润
fa_gr_ttm	营业总收入	fa_gc_ttm	营业总成本
fa_or_ttm	营业收入	fa_ebittogr_ttm	息税前利润/营业总收入
fa_debttoequity	产权比率	fa_bps	每股净资产
fa_operactcash-flow_ttm	经营活动现金净流量	fa_eps_basic	基本每股收益
tech_rsi	相对强弱指标	fa_cfps_ttm	每股现金流量净额
fa_cfotocurliabs_ttm	现金流动负债比	val_lnfloatmv	对数流通市值
fa_optogr_ttm	营业利润/营业总收入	fa_octogr_ttm	营业总成本/营业总收入
tech_vema5	成交量的 5 日指数移动平均	tech_vema10	成交量的 10 日指数移动平均
tech_tvma20	20 日成交金额的移动平均值	tech_tvma6	6 日成交金额的移动平均值
fa_ocftodebt	经营活动产生的现金流量净额/负债总额	tech_tvstd20	20 日成交金额的标准差
fa_ocftoor	经营活动产生的现金流量净额/营业收入	fa_ocfps_ttm	每股经营活动产生的现金流量净额 TTM
tech_tvstd6	6 日成交金额的标准差	tech_revs10	过去 10 日的价格动量
fa_debttoasset	资产负债率	val_pbindu_sw	(个股市净率 − 行业 PB 均值)/PB 行业标准差