Hans Journal of Data Mining 数据挖掘, 2024, 14(1), 43-63
Published Online January 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.141005

Hans汉斯

基于机器学习的选股模型及投资组合研究

卢相刚，王泽仁

广东工业大学数学与统计学院，广东 广州

收稿日期：2023年12月17日；录用日期：2024年1月19日；发布日期：2024年1月26日

摘　要

股票是金融市场重要的组成部分，其变化存在着一定的内在规律，但是也受到多种因素的制约与影响。因此，如何能够选取好的股票进行操作也成为了很多从业者的研究方向。传统的选股策略有两种，一种为多元回归法，其缺点是对于极端值较为敏感，极端值的存在会影响回归结果，另一种是多因子打分法，其缺点是需要人为给定各个因子的权重，主观性对选股结果有很大影响。本文使用基于决策树的Adaboost模型进行选股，并且构建了投资组合的优化模型，有效提升了投资的收益率。本文的主要工作包括：(1) 建立股票特征指标库，选取更能解释模型的特征指标并对其进行有效性分析和相关性分析；(2) 构建基于决策树的Adaboost选股模型，对模型参数进行优化并且对模型的泛化能力进行评估；(3) 对马科维兹的投资组合模型进行改进，提出一种新的投资组合模型，使得能在降低总体风险的同时将投资收益维持在一个相对高的水平。

关键词

选股模型，机器学习，投资组合

Stock Prediction Method Based on Machine Learning and Portfolio Research

Xianggang Lu, Zeren Wang

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

Received: Dec. 17th, 2023; accepted: Jan. 19th, 2024; published: Jan. 26th, 2024

Abstract

Stock is an important part of the financial market. It has certain intrinsic laws of change, but it is also subject to and influenced by many factors. Therefore, how to select good stocks for trading has become a research direction for many practitioners. There are two traditional stock selection

文章引用: 卢相刚, 王泽仁. 基于机器学习的选股模型及投资组合研究[J]. 数据挖掘, 2024, 14(1): 43-63.
DOI: 10.12677/hjdm.2024.141005

strategies: one is the multiple regression method, which is sensitive to extreme values, the presence of which will affect the regression results; the other is the multi-factor scoring method, which requires artificially assigning weights to each factor and has a great impact on the stock selection results. This paper uses Adaboost model based on decision tree for stock selection and constructs an optimization model for investment portfolio, which effectively improves the investment return. The main work of this paper includes: (1) Establishing a database of stock characteristic indices, selecting characteristic indices with strong explanatory power, and conducting validity analysis and correlation analysis; (2) Constructing an Adaboost stock selection model based on decision trees, optimizing model parameters, and then evaluating the effectiveness and generalization ability of the model; (3) Improving Markowitz's portfolio model and proposing a new investment portfolio model, which can reduce the overall risk and keep the investment return at a relatively high level.

Keywords

Quantitative Stock Selection, Machine Learning, Portfolio

1. 绪论

1.1. 研究背景及意义

1.1.1. 研究背景

自上世纪 90 年代，中国先后在上海和深圳成立证券交易所，才真正揭开了国内证券市场发展的帷幕。至 2021 年 9 月，我国首家公司制证券交易所北交所注册成立，新三板改革得到深化，资本市场为助力中小企业发展提供有力支持，进一步促进后疫情时代经济发展[1]。随着计算机技术与金融数学模型的发展，一项新名词出现在了大众的视野中——量化投资。量化投资，指通过计算机技术和数学建模等手段，用机器语言表述投资策略，利用计算机帮助人们对股票的风险和收益进行量化来参与股票交易的方式。量化投资策略很多，几乎涵盖了投资的全过程，主要有资产配置、量化选股、量化择时、算法交易和风险控制等[2]。

随着大数据时代的到来，各类算法与模型研究更进一步，各种数据处理技术与机器学习方法被应用在各行各业，机器学习因其突出的学习能力与泛化能力成为各项研究的有力工具[3]。机器学习模型相较于传统的线性计量经济学模型，在处理非线性变量方面具有显著优势，在面对金融领域一些不确定的问题时往往能给出更合理的解释，机器学习应用于量化选股将更具优势。另外，投资组合也是金融领域研究中的一个关键问题，一个好的投资组合模型可以得到更好的有效边界，再同等风险下可以给投资者带来更多的收益。但是，有较少投资者注意到根据预测结果选择优质资产作为投入也是形成优质投资组合的可靠保证[4]。因此将选股模型与投资组合理论相结合，在量化投资中具有广阔的应用前景。

基于此，本文对上证 50 股票数据进行分析研究，得出解释能力强的特征指标，采用基于决策树的 Adaboost 模型进行多维特征指标选股和投资组合研究。

1.1.2. 研究意义

(1) 本文利用数据挖掘流程对数据进行预处理，包括数据清洗，数据变换与数据标准化；建立特征指

标库，从多角度探究股票涨跌的原因，准确提取了有效的特征指标。

(2) 使用基于决策树的 Adaboost 模型，对筛选好特征指标的历史数据进行训练，训练后得到模型的参数，然后用最新一期的特征数据来预测未来的收益率情况，弥补了使用时间序列预测方法适用范围受限的不足。为后续将机器学习和深度学习应用于金融领域的研究者提供一定的参考与思路。

(3) 向投资者提供了更加广泛的投资视角。随着市场风格的不断改变，传统因子选股的有效性开始逐渐降低。在目前的经济环境下，建立一个可以跑赢指数的选股模型，并且可以无缝地运用于资产分配，形成机器选股分配资金为主、投资者决策为辅的投资策略，对于降低选取股票的难度以及提高股票收益具有现实意义。

1.2. 国内外研究现状

国内对于这类研究工作起步较晚，许多研究者在总结国外量化模型的基础上，进行了适合本国市场的量化研究。

刘毅[5]选取了成长、估值、质量、动量这四方面的特征指标进行最优因子组合，利用历史数据进行回测检验，结果表明在多种环境下，最优因子组合的股票都跑赢了市场基准。江方敏[6]使用打分法进行选股，根据市场风格构造多种投资组合，同样取得了较好表现。徐景昭[7]进行了行业对股市影响的研究，采用多元回归的方法优化了多因子选股模型。曹正凤等人[8]使用机器学习中的随机森林方法对股票的涨跌进行预测分类，预测准确率达到了较高水平。

国外量化投资的研究工作起步较早，且研究更为全面，对于多因子选股模型的构建，不仅考虑了各类资产的定价，还考虑了特征工程以及机器学习方法的应用。

Markowitz [9]建立了均值－方差模型，将数学工具与金融领域相结合，成为现代量化投资研究与资产配置的基础。Ross [10]提出了套利定价理论，该理论认为证券的价格不能完全由系统性风险解释，而是会受到多种因素的影响。套利定价理论的提出使得研究者开始对多因子进行研究，进而建立了多因子模型，多因子模型已经成为当下应用最为广泛的模型，几乎所有的选股方法都是基于多因子模型而建立的。Fama 和 French [11]提出了三因子模型，他们通过构造市场风险、市值风险和账面市值比三个因子来刻画股票收益的变化。Patalay 等人[12]通过人工智能技术和机器学习方法设计了财务决策支持系统，该系统有助于更高效对股票做出财务决策。Li 等人[13] [14] [15]介绍了人工智能背景下不同的量化投资方法并且使用评分和筛选模型以及卷积神经网络模型进行实证并被证明是有效的。Gao 等人[16]使用了三种机器学习方法以及线性回归来预测收益，结果表明支持向量回归的预测成功率最稳定。

综合以上对于国内外研究现状的分析，目前对于选股与投资组合理论相结合的相关研究较少，基于此，本文计划将选股模型与投资组合优化理论相结合，在选股的同时对资金进行分配，从而构建投资组合进行研究。

2. 相关理论及技术综述

2.1. 机器学习

机器学习的理念是让机器就可以直接从经验或数据中学习如何处理复杂的任务，我们要做的不是告诉机器如何做决策，而是让机器从经验或数据中学习，从而可以自己做出决策。本次案例将会使用到机器学习中的两个经典算法。

2.1.1. 决策树

本文使用的单层决策树(Decision stump)，又称为决策树桩作为弱学习器。如图 1 所示，决策树桩也

就意味着根结点直接与终端结点相连，仅可以对一个属性进行一次划分作为最终的分类结果。若要使得弱学习器在集成时有较好的效果，需要尽可能的找到具有最低错误率的决策树桩。

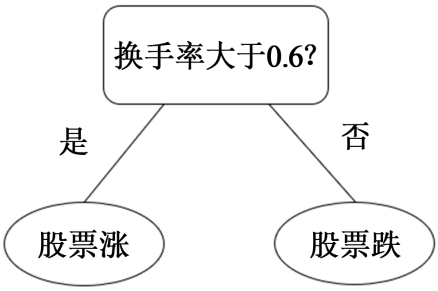

Figure 1. Decision stump
图 1. 单层决策树

2.1.2. Adaboost 集成学习

Adaptive Boosting (自适应增强)，简称 Adaboost。Adaboost 是一种迭代算法，其核心思想是将多个弱分类器以某种方式进行结合，从而形成一个更有助于做决策的强分类器。其工作机制为：先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本的分布进行调整，使得先前基学习器做错的训练样本在后续得到更多的关注，然后基于调整后的样本分布来训练下一个基学习器，直至基学习器的数目达到事先给定的值 T，最终将这 T 个基学习器进行加权结合。Adaboost 方法流程如图 2 所示：

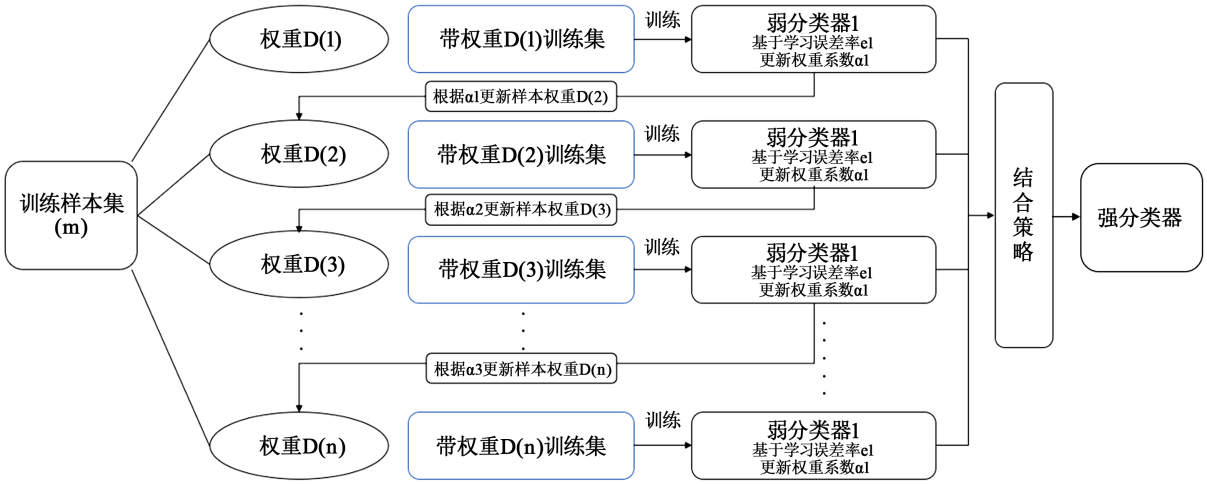

Figure 2. The process of Adaboost
图 2. Adaboost 算法流程

2.2. K 折交叉验证

交叉验证基本思想就是将原训练数据分为两个互补的子集，一部分作为训练数据来训练模型，另一部分作为测试数据来评价模型。交叉验证可以解决数据量不够大的问题，而简单的交叉验证并不能使数据得到充分利用，因此提出 K 折交叉验证。K 折的意思就是将数据集分为 K 份，其中 K-1 份作训练集，剩下 1 份作为测试集。K 折交叉验证的基本思路如图 3 所示。

由于本次案例数据集较小，将使用 10 折交叉验证来训练模型并给出最终模型的评价结果。

Figure 3. The process of K-fold cross validation
图 3. K 折交叉验证流程

3. 特征工程

3.1. 数据预处理

3.1.1. 数据清洗

本文使用 Python 读取数据并对数据进行清洗。

第一步，处理重复值。下载的数据很可能存在重复数据，要对重复值进行删除并重新排列。

第二步，处理缺失值。由于一些股票上市时间较晚，所以存在某些特征指标数据缺失的情况，由于缺失值数量不是很多，我们直接将缺失值所在的行进行删除。

第三步，处理异常值。通常来讲，股票的涨跌的原因往往在这些异常值中更容易挖掘，而过多的异常值又会导致模型性能的下降。这里采取的策略是，只删除高度异常值，保留普通异常值来保证模型的部分泛化性能。

数据清洗流程如图 4 所示：

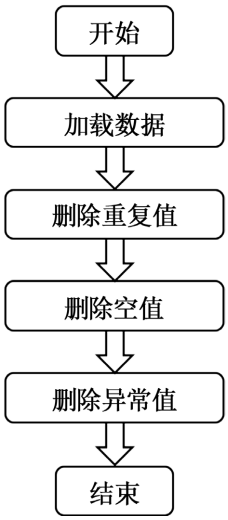

Figure 4. Data cleaning process
图 4. 数据清洗流程

3.1.2. 数据变换

由于股票的某些特征指标是按季度进行统计的，所以要将低频数据转换为高频数据，这里使用了Eviews对季度数据用二次插值的方法将其转换为月度数据。以浦发银行2018年到2020年的市盈率为例，插值前的数据和插值后的数据分别如表1和表2所示：

Table 1. Quarterly data before conversion
表1. 变换之前的季度数据

股票代码	股票名称	日期	市盈率
600000	浦发银行	2018-03-31	6.3206
600000	浦发银行	2018-06-30	5.1335
600000	浦发银行	2018-09-30	5.6092
600000	浦发银行	2018-12-31	5.1445
600000	浦发银行	2019-03-31	5.7018
600000	浦发银行	2019-06-30	5.7666
600000	浦发银行	2019-09-30	5.6919
600000	浦发银行	2019-12-31	6.1633
600000	浦发银行	2020-03-31	4.9809
600000	浦发银行	2020-06-30	5.5693
600000	浦发银行	2020-09-30	4.9838
600000	浦发银行	2020-12-31	4.8715

Table 2. Monthly data after conversion
表2. 变换之后的月度数据

日期	市盈率	日期	市盈率	日期	市盈率
2018M01	7.024226	2019M01	5.588996	2020M01	5.112693
2018M02	6.259015	2019M02	5.720041	2020M02	4.915315
2018M03	5.678559	2019M03	5.796363	2020M03	4.914693
2018M04	5.282859	2019M04	5.765667	2020M04	5.547078
2018M05	5.071915	2019M05	5.771767	2020M05	5.612778
2018M06	5.045726	2019M06	5.762367	2020M06	5.548044
2018M07	5.589952	2019M07	5.635896	2020M07	5.108863
2018M08	5.644030	2019M08	5.671674	2020M08	4.966274
2018M09	5.593619	2019M09	5.768130	2020M09	4.876263
2018M010	5.147993	2019M010	6.251174	2020M010	4.838830
2018M011	5.106648	2019M011	6.224552	2020M011	4.853974
2018M012	5.178859	2019M012	6.014174	2020M012	4.921696

3.1.3. 数据标准化

由于不同的因子所描述的对象单位不同，因此不同因子之间的取值差异可能很大。为了避免这些差异对模型的训练产生坏的影响，因此要对数据进行标准化，这里使用 Z-score 标准化方法。Z-Score 是数据标准化处理的一种常用方法，这种方法根据原始数据的均值和标准差进行标准化，经过处理后的数据符合标准正态分布，即均值为 0，标准差为 1。通过 Z-Score 可以将不同量级的数据转化为统一量度的 Z-Score 分值，并进行比较。转化函数为(1)：

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

其中 μ 为所有样本数据的均值，σ 为所有样本数据的标准差。

这里使用 sklearn 库的方法直接对数据进行 Z-score 标准化。

3.2. 特征指标选取

3.2.1. 选取思路

股票的超额收益是由不同特征指标的共同驱动下才产生的，因此特征指标的选择对于模型的构建与未来预测的准确性是非常重要的。若同时选择更多更有效的特征指标，可以更好地增强多因子模型的解释能力，更好地刻画因子与收益之间的关系，从而带来更多的收益。

3.2.2. 特征指标库的构建

本次案例考虑到对于模型有效的因特征指标应该具有可持续性、鲁棒性、可解释性等特征。因此从股票量化的因子库中，在市值类、估值类、成长能力类、情绪类、利润表类、质量类、动量类、每股指标类中初步挑选了如表 3 所示的候选特征指标。

Table 3. Characteristic indicators library
表 3. 特征指标库

特征指标所属类目	特征指标名称	特征指标描述
市值类	Mc	总市值
	Tmv	流通市值
估值类	PE	市盈率
	PB	市净率
	PS	市销率
	PCF	市现率
	ROE	净资产收益率
成长能力类	DivYield	股息率
	OperevYOY	营业收入同比增长率
	NetprfYOY	净利润同比增长率
	OpeprTOR	营业利润/营业总收入
	TopecostTOR	营业总成本/营业总收入
情绪类	FulTurnR	换手率

续表

利润表类	TotOpRev	营业总收入
	OpePrf	营业利润
	NPParentComp	归属母公司净利润
	NetprfCut	扣除非经常性损失后净利润
质量类	ROA	资产回报率
	Gincmrt	销售毛利率
	Netprfrt	销售净利率
	OPItprf	经营活动净收益 / 利润总额
	Curtotast	流动资产/总资产
动量类	REVS60	过去三个月的价格动量
每股指标类	OpeprfPS	每股营业利润
	MincmPS	每股营业收入
	BasicEPS	基本每股收益

3.2.3. 特征指标的有效性分析

信息系数(Information Coefficient，简称 IC)，表示所选特征指标与股票下期收益的相关系数，通过 IC 的值可以判断该特征指标对于下期收益率的预测能力，IC 越高，表明该特征指标对股票收益的预测能力越强，通常使用 Rank IC 来刻画 IC 值。Rank IC，即某时点某因子在全部股票暴露值排名与其下期回报排名的截面相关系数。Rank IC 的公式为(2)：

$$\text{Rank IC} = corr\left(order_{t-1}^{f} - order_{t}^{r}\right) \tag{2}$$

其中 $order_{t-1}^{f}$ 为 $t-1$ 期个股票的因子值排名， $order_{t}^{r}$ 为 t 期股票收益率排名。

一般而言，一个特征指标的 IC 值的绝对值高于 0.02，便可认为该特征指标的有效性较好，本次案例中，将使用 Rank IC，即斯皮尔曼相关系数来求候选池中各个特征指标的 IC 值。如表 4 所示：

Table 4. The IC value of the characteristic indicators
表 4. 特征指标的 IC 值

特征指标名称	IC 均值	IC 标准差	|IC| > 0.02 比率
总市值	0.02	0.03	48.57%
流通市值	0.03	0.06	74.29%
市盈率	−0.02	0.06	73.00%
市净率	−0.04	0.11	74.29%
市销率	−0.01	0.09	77.14%
市现率	0.03	0.07	83.37%
净资产收益率	0.02	0.08	80.00%

续表

股息率	−0.02	0.05	71.34%
营业收入同比增长率	0.02	0.07	77.14%
净利润同比增长率	0.02	0.08	80.00%
营业利润/营业总收入	0.03	0.08	85.71%
营业总成本/营业总收入	0.03	0.08	85.71%
换手率	−0.09	0.10	94.29%
营业总收入	0.05	0.06	82.86%
营业利润	0.05	0.06	82.86%
归属母公司净利润	0.05	0.06	82.86%
扣除非经常性损失后净利润	0.05	0.06	91.43%
资产回报率	0.02	0.08	94.29%
销售毛利率	0.01	0.07	68.57%
销售净利率	0.02	0.07	85.71%
经营活动净收益/利润总额	0.05	0.06	71.43%
流动资产/总资产	0.00	0.06	74.29%
过去三个月的价格动量	0.01	0.05	65.31%
每股营业利润	0.03	0.09	88.57%
每股营业收入	0.02	0.05	80.00%
基本每股收益	0.03	0.09	94.29%

根据表 4 我们可以初步对因子进行筛选，例如，在表中可以看到换手率的 IC 均值为−0.09，且大于 0.02 的比率达到 90%多，说明该特征指标对股票收益的预测能力很强，属于优质因子。在表中又可以看到销售毛利率的 IC 均值较小且大于 0.02 比率比较低，其对于股票收益的预测能力较弱，因此我们可以将其剔除。

由于各支股票的业务场景不同，其财报的内容也有所差异。例如大多数银行的财报对于市场销售方面的指标比较少，为了避免存在太多缺失值，我们又剔除了例如市销率、过去三个月的价格动量等指标。

最终经过以上有效性分析，我们初步筛选得到的特征指标如表 5 所示。

Table 5. Effective characteristic indicators
表 5. 有效因子

有效因子汇总				
市值类	总市值	流通市值		
估值类	市盈率	市净率	市现率	净资产收益率
成长能力类	股息率	营业收入同比增长率	净利润同比增长率	
情绪类	换手率			

续表

利润表类	营业总收入	营业利润	归属母公司净利润	扣除非经常性损失后净利润
质量类	销售净利率			
每股指标类	每股营业收入	基本每股收益		

3.2.4. 特征指标的相关性分析

相关性是不同的变量之间可能存在非严格的不确定关系，对于变量间的相关关系的分析，称为相关性分析。为了避免特征指标的冗余，需要对特征指标计算其相关系数，剔除相关系数很高的两个特征指标，从而保持特征指标之间存在较高的独立性。相关性分析得到结果如图 5 所示。

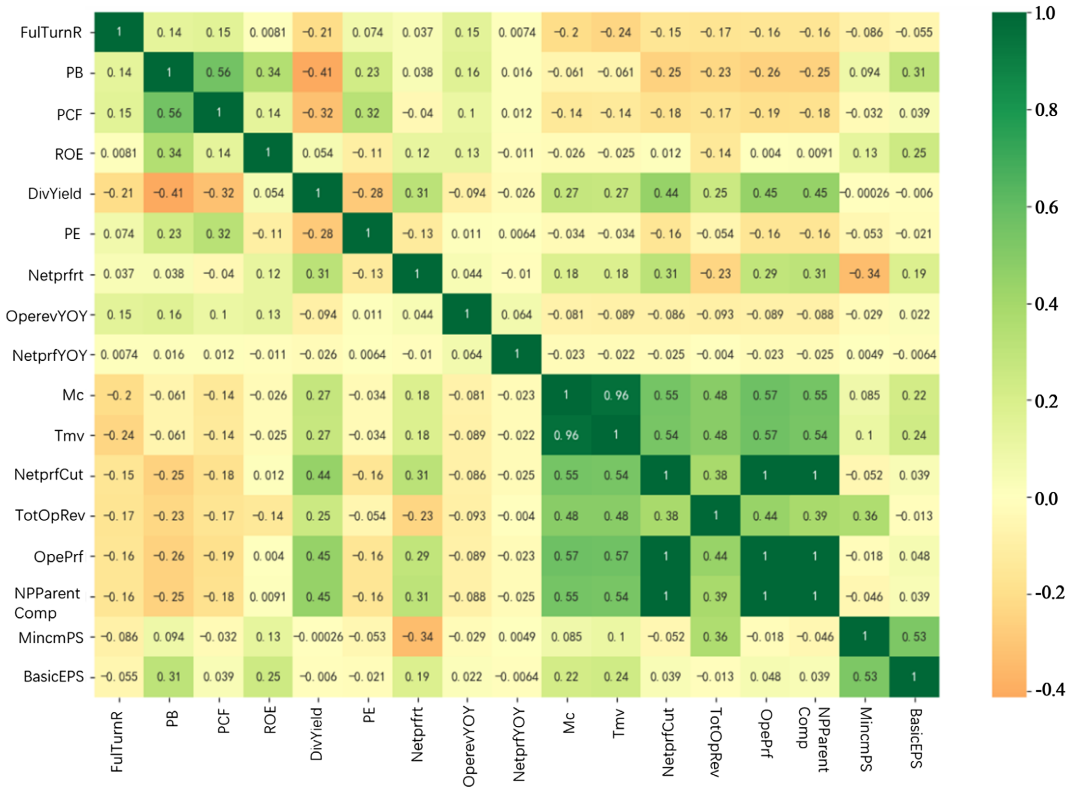

Figure 5. Correlation analysis
图 5. 相关性分析

从图中可以直观的看到月总市值和月流通市值存在很强的正相关关系，达到了 0.96；并且营业利润与扣除非经常性损益后净利润与归属于母公司所有者的净利润的相关关系近似于 1。综合以上相关性分析，最终决定剔除月总市值、营业利润、归属于母公司所有者的净利润这三个特征指标，保留月流通市值和扣除非经常性损益后净利润。

4. 基于决策树的 Adaboost 选股模型

4.1. Adaboost 模型思想

Adaboost 模型的主要思想主要有两个方面：1) 如何修正每一轮训练数据的权重。2) 如何将一组弱学习器组合成强学习器。

对于第一个方面，Adaboost 模型的做法是将第一轮弱学习器分类错误的样本在之后的学习过程中受到更多的关注，即提高分类错误的样本的权重，这样就可以保证在下一轮训练时的弱学习器会更多的处理这些之前出错的样本。理论上来讲，训练过程的分类错误率可以趋于 0。

对于第二个方面，Adaboost 模型不是将弱学习器的分类结果进行简单平均后组合，而是利用弱学习器的分类错误率对权重进行改变，加大分类误差率小的弱分类器权值，使其在分类的表决中起较大的作用，减小分类误差率大的弱分类器的权值，使其在表决中起较小的作用。

4.2. 基于决策树的 Adaboost 算法步骤

首先假设有 m 个样本的训练集，标签为−1 和 1，若有 n 个弱分类器的预测结果分别是 $\left(h_1(x),h_2(x),\cdots,h_n(x)\right)$，算法的基本原理如下：

(1) 计算样本权重

赋予训练集中每个样本一个权重，构成权重向量 D，权重向量 D 一般初始化为 $1/m$。

(2) 计算错误率

在训练集上训练一个弱分类器，并计算分类器的错误率：$\varepsilon=$ 分错的数量/样本总量

(3) 计算弱分类器的权重

为当前弱分类器赋予权重 alpha，如(3)所示：

$$\alpha=\frac{1}{2}\ln\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{3}$$

(4) 调整权重值

根据上一次训练结果，调整权重值(上一次分对的样本权重降低，分错的样本权重增加)

如果第 i 个样本被正确分类，则该样本权重更改为(4)：

$$D_i^{(t+1)}=\frac{Di_i^{(t)}\mathrm{e}^{-\alpha}}{sum(D)} \tag{4}$$

如果第 i 个样本被分错，则该样本权重更改为(5)：

$$D_i^{(t+1)}=\frac{Di_i^{(t)}\mathrm{e}^{\alpha}}{sum(D)} \tag{5}$$

(5) 最终强分类器的结果

循环结束，得到强分类器的预测结果(6)：

$$H(x)=Sign\left\{\sum_i^n\left(\alpha_t h_t(x)\right)\right\} \tag{6}$$

这个公式相当于弱分类器分类结果的线性组合得出强分类器的结果，若得出的值为正值，则强分类器的预测结果为 1，若为负值，则强分类器的预测结果为−1。

可以看到，强分类器是由弱分类器的给出的结果加权相加之后，根据该值的正负给出最终的判别结果。显然我们对于一组股票仅仅是判断其未来的涨跌，仍然很难选择出其中的几支股票进行入手。若能够给出一组股票未来上涨或下跌的“机率”，这对我们选择股票将会是很重要的依据。而由弱分类器的给出的结果加权相加后的值，恰恰可以给出我们选股的依据，这里我们称这个值为类别估计值。即(7)：

$$\sum_i^n\left(\alpha_t h_t(x)\right) \tag{7}$$

4.3. 模型构建

本文将上证 50 股票的特征指标数据作为样本，将月度收益率用符号函数处理后作为数据集的标签。为了避免有太多相似数据影响模型的性能，本文将样本依照月度收益率作降序排序，只取位于前百分之三十及后百分之三十的数据，剔除位于中间的数据。由于剔除数据后数据集变小，所以将 2010 年 1 月至 2018 年 12 月的月度数据利用十折交叉验证的方法给出模型的精确度，将 2019 年 1 月至 2019 年 12 月的数据作为测试集进行验证。

4.4. 模型参数优化

本文采用单层决策树作为弱分类器，单层决策树仅仅基于单个特征指标做出决策，由于这棵树只有一次分裂过程，所以在这里我们需要对每个特征指标的分裂阈值参数进行优化。另外，对弱分类器的个数也需要进行优化。若只有一个弱分类器，它只能根据一个特征指标进行分类，显然其性能很弱；但若增加弱分类器的数目，其训练的误差就会逐渐减小。但并不是弱分类器的数目越多越好，随着弱分类器数目的增加很可能会发生过拟合进而影响测试的准确率。图 6 即为分类器个数与训练集测试集准确率的关系。

分类器数目	训练集准确率	测试集准确率
10	57.63	54.28
50	60.07	57.75
100	60.90	58.56
300	62.68	58.29
500	63.75	58.82
1000	62.83	58.29
1500	62.83	58.29

Figure 6. Relationship between the number of classifiers and accuracy
图 6. 分类器数量与准确率之间的关系

接下来验证模型的有效性，我们将数据集进行十折交叉验证，将数据随机打乱后分为十份，九份作为训练集，剩下的一份作为测试集，每一份数据集都会作为测试集，迭代十次。之后分别计算训练集与测试集的准确率，最后求其均值作为模型有效性的最终结果。如图 7 所示。

从交叉验证的均值图中可以看到，随着分类器的个数的增加，训练集和测试集的准确率都在增长，分类器个数从 10 到 50 增加时，其准确率增长较快，准确率在分类器个数为 1000~1500 趋于平缓，且训练集准确率始终大于测试集的准确率，符合模型的预期表现。

接下来对模型的泛化能力进行评估。我们的模型是根据一系列特征指标产生一个实值，然后将这个预测值与分类阈值进行比较，若大于该阈值则将其划分为正类，若小于该阈值则将其划分为反类，这个实值的好坏决定了模型的泛化性能，而 ROC 曲线就是从这个角度来评估模型泛化能力的有力工具。ROC 曲线根据模型的预测结果对样本进行排序，按此顺序把样本作为正例进行预测，计算出“真正例率”与“假正例率”，以它们为横纵坐标轴。图 8 即为该模型的 ROC 曲线图。

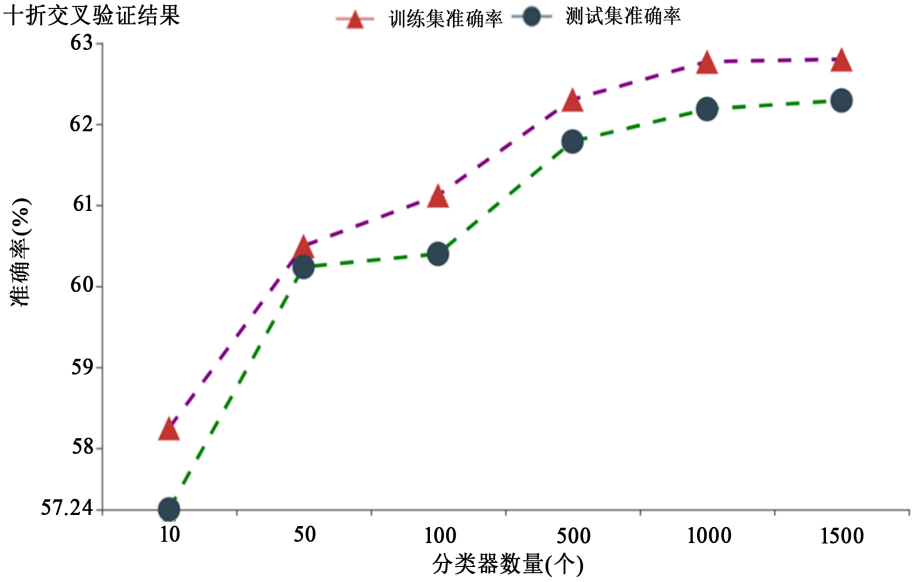

Figure 7. Average results of the ten fold cross validation
图 7. 十折交叉验证的平均结果

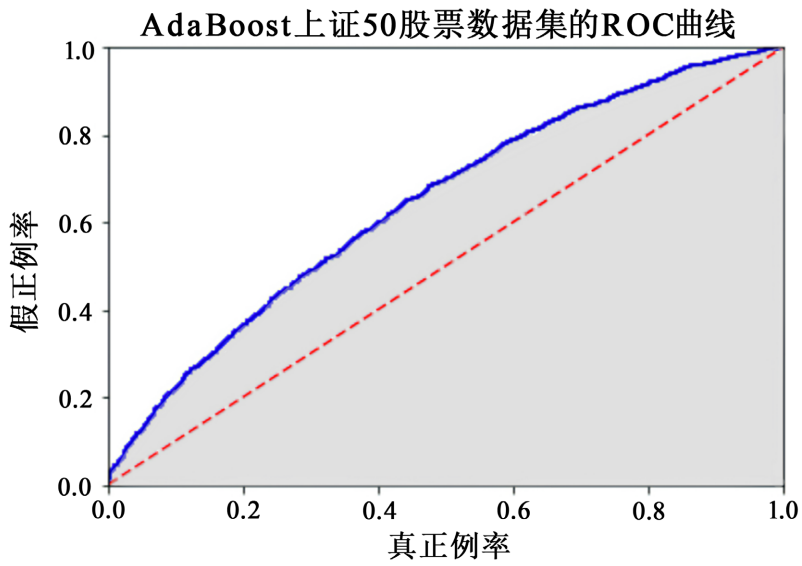

Figure 8. ROC curve of the model
图 8. Adaboost 模型的 ROC 曲线

一般来说，我们用 ROC 曲线下的面积，即 AUC (Area Under Curve)来作为模型的评价标准。该方法可以直观的通过图示观察从而分析学习器的准确性，并可用肉眼作出判断。从图中可以看到该模型的 AUC 明显大于 0.5，说明该模型有一定的预测价值。

5. 基于决策树的 Adaboost 选股模型

5.1. 马科维兹理论

1952 年，马科维兹提出了一种投资组合优化模型，又称均值方差模型。该模型的建立基于以下假设：证券市场是有效的，即投资者充分了解各种风险资产的期望收益率及标准差；各种证券的收益率有一定

的相关关系；投资者都是风险规避者且对是收益不知足者；证券的收益率服从正态分布；每一个资产都是无限可分的；税收和交易成本忽略不计。因为风险较难度量，所以这里我们给出在一定收益下风险最小的均值－方差模型，该模型如(8)所示：

$$\begin{cases} \min \delta_Q = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i \cdot w_j \cdot \rho_{ij} \cdot \delta_i \cdot \delta_j} \\ \text{s.t. } E\left(r_Q\right) = \sum_{i=1}^{n} w_i \cdot r_i = r_0 \\ \sum_{i=1}^{n} w_i = 1 \quad w_i \geq 0, i = 1,2,\cdots,n \end{cases} \tag{8}$$

其中 Q 代表投资组合，δ_Q 为投资组合的期望标准差，$E\left(r_Q\right)$ 为组合 Q 的期望收益，r_i 为第 i 支股票的期望收益，w_i 为第 i 支股票的购买权重，δ_i 为第 i 支股票预期收益率的标准差，ρ_{ij} 为第 i 支和第 j 支股票的相关系数。

接下来使用该理论对 2018 年 12 月的五种风险资产进行组合：五只股票分别是万华化学、中国平安、中国中免、洛阳铝业与贵州茅台。将前三个月的平均月收益率作为期望收益，接下来根据前十一个月的月收益率计算其相关系数，代入该模型。可以得出各支股票权重如表 6 所示。

Table 6. Quarterly data before conversion
表 6. 变换之前的季度数据

股票名称	投资权重
万华化学	0.0005
中国平安	0.0000
中国中免	0.0014
洛阳铝业	0.2871
贵州茅台	0.7110

五支股票的权重分别为 0.0005，0.0000，0.0014，0.2871，0.7110，即使用大约 30%的资金购买第四支股票，70%的资金购买第五支股票。

5.2. 马科维兹理论的不足

从上文可以看到马科维兹的均值－方差模型得到的最优解出现了对某支股票的投资占据了非常大的权值，即投资者仅对某一只股票重仓而不进行分散组合，这无疑违背了风险分散化的投资原理，使得非系统性风险得不到有效的分散与减少。同时，马科维兹的均值－方差模型对于期望收益和方差指标的给定，都是根据自身经验所做出的判断，参数的估计等会有很大的偏误。另外，马科维兹理论没有充分考虑投资组合收益率的偏态特征，即对于真实收益高于期望收益与真实收益低于期望收益的情况作同等对待，但是人们普遍将风险视为期望收益低于实际收益的情况，而对于实际收益大于期望收益的情况，则认为是投资成功。

5.3. 投资组合模型的优化

为了降低风险，可以对股票的权值进行限定，即对于所选股票都给予一定的初始权值，该初始权值

由其股票的类别估计值给出，从而达到对风险进行分散的目的。

同时，本文将要对期望收益进行预测，得出的预测收益率将代替均值方差模型中的期望收益率。这里采用 LSTM 长短时记忆网络对收益率进行预测，通过 LSTM 对于收益率的长期依赖设定能够得到精度更高的预测值，从而达到减小误差的目的。

另外，本文将以半方差的思想重新刻画风险指标。假设真实收益率为r，预测收益率为$\hat{r}$，风险为ε，则将ε表示为(9)：

$$\varepsilon = \begin{cases} r - \hat{r} & r < \hat{r} \\ 0 & r > \hat{r} \end{cases} \tag{9}$$

由上式可知，若预测收益率大于真实收益率，即$\varepsilon < 0$，这无疑是投资者不能接受的情形，因为其对于预测收益率大的股票所给的权重也大，这样会影响最终收益；当ε为 0 时，预测收益率小于真实收益率，这种情况是投资者所能接受的，虽然存在一定的误差，使得实际的权重减小，但可以降低不必要的风险。基于以上分析，构建的风险指标如下：$\gamma_{ij} = \varepsilon_i * \varepsilon_j$，此指标代表了第 i 支股票与第 j 支股票的相关风险，可以看到只有两种情况下γ_{ij}为 0，一种为ε_i和ε_j都大于 0，即两支股票都是真实收益率大于预测收益率，这种情况可以为我们带来超额收益；另一种为ε_i和ε_j其中一个大于 0，一个小于 0，这种情况两只股票为负相关，可以使得投资组合的风险降低。γ_{ij}大于 0 时，说明两支股票为正相关且会使得最终收益减少，所以我们要使得该指标的值尽可能小。针对优化的目标得到新的投资组合模型如(10)所示：

$$\begin{cases} \min \sum_{i=1}^{n} \sum_{j=1}^{n} w_i \cdot w_j \cdot \gamma_{ij} - \sum_{i=1}^{n} w_i \cdot \hat{r}_i \\ \text{s.t.} \ \sum_{i=1}^{n} w_i = 1 \\ \quad w_i \geq w_0, i = 1, 2, \cdots, n \end{cases} \tag{10}$$

其中w_i为第 i 支股票的购买权重，$\hat{r}_i$为第 i 支股票的预测收益率，γ_{ij}为构建的风险指标，w_0为各支股票的初始限定权值，对股票的初始权值进行限定意味着不会出现某些股票买入权值为 0 的情况，降低了系统风险。

5.4. 投资组合模型求解

本文对 2020 年 1~6 月份的股票进行投资组合，以 2020 年 1 月为例，模型的基本求解步骤如下：

第一步：使用 Adaboost 选股模型从股票池中选取股票。

这里将 2019 年 12 月上证 50 股票的特征指标进行输入，得出 1 月份的累计类别估计值，按照该值的大小选取前十位进行投资组合。选取的股票如表 7 所示：

Table 7. The selected stocks
表 7. 选取的股票

股票名称	累计类别估计值
贵州茅台	0.56
中国平安	0.56
兴业银行	0.54

续表

三安光电	0.52
中信建投	0.44
华泰证券	0.40
海尔智家	0.40
建设银行	0.36
中国中免	0.34
工商银行	0.34

第二步：确定股票的初始限定权值

根据类别估计值对股票的初始购买权值进行限定。我们将初始权值设置为占总金额的 30%，将各支股票类别估计值所占该值总和的比例与初始权值相乘，即可得到各支股票的初始限制权值。如表 8 所示：

Table 8. The selected stocks
表 8. 选取的股票

股票名称	初始限定权值
贵州茅台	0.04
中国平安	0.04
兴业银行	0.04
三安光电	0.03
中信建投	0.03
华泰证券	0.03
海尔智家	0.03
建设银行	0.03
中国中免	0.02
工商银行	0.02

第三步：使用 LSTM 长短时记忆网络对收益率进行预测。

利用历史数据训练 LSTM 模型，除了使用特征工程中选取的特征指标外，还增加了开盘价、收盘价、最低价、最高价、成交量等基本盘面指标，以及 ASI、ATR、MTM、RSI、SAR 等与时间序列相关的指标。这里使用前三个月的历史数据来预测下个月的收益率。图 9 为对中国平安的预测效果图，在图中红色线条为真实收益率，蓝色线条为预测收益率，可以看到 LSTM 模型预测的收益率只能模拟出收益率的大概走势，对于幅度过大振荡的预测都较为保守。

第四步：计算风险指标 γ_{ij}

根据上文风险指标的计算方法，以 1 月为例，得出风险指标矩阵为(11)：

$$\gamma_{ij} = \begin{bmatrix} 0.0273 & 0.0150 & 0.0000 & 0.0000 & 0.0000 & 0.0146 & 0.0132 & 0.0023 & 0.0144 & 0.0060 \\ 0.0150 & 0.0083 & 0.0000 & 0.0000 & 0.0000 & 0.0081 & 0.0073 & 0.0012 & 0.0079 & 0.0033 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0146 & 0.0081 & 0.0000 & 0.0000 & 0.0000 & 0.0078 & 0.0071 & 0.0012 & 0.0077 & 0.0032 \\ 0.0132 & 0.0073 & 0.0000 & 0.0000 & 0.0000 & 0.0071 & 0.0064 & 0.0011 & 0.0069 & 0.0029 \\ 0.0023 & 0.0012 & 0.0000 & 0.0000 & 0.0000 & 0.0012 & 0.0011 & 0.0002 & 0.0012 & 0.0005 \\ 0.0144 & 0.0079 & 0.0000 & 0.0000 & 0.0000 & 0.0077 & 0.0069 & 0.0012 & 0.0076 & 0.0032 \\ 0.0060 & 0.0033 & 0.0000 & 0.0000 & 0.0000 & 0.0032 & 0.0029 & 0.0005 & 0.0032 & 0.0013 \end{bmatrix} \quad (11)$$

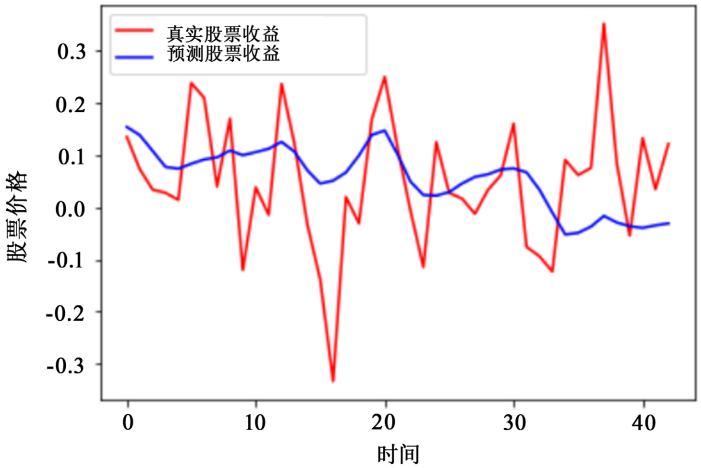

Figure 9. The predictive effect of China Ping An
图 9. 中国平安的预测效果

第五步：将参数带入模型，得出股票权重。

将参数带入优化模型中，最终得到股票的购买权重。图 10 为结果图：

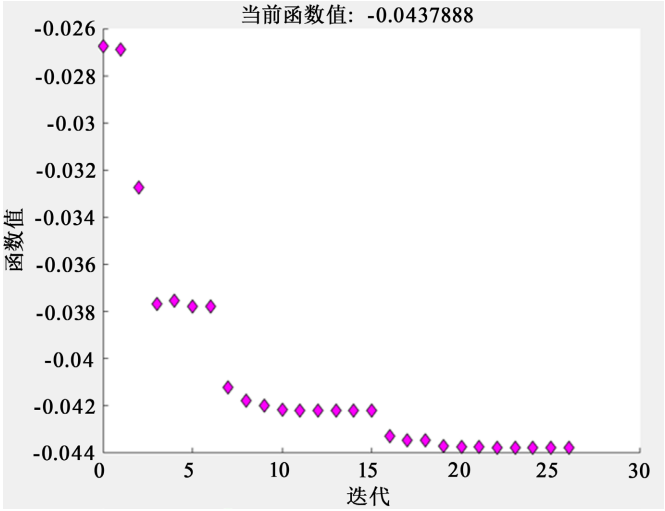

Figure 10. The process of iterative optimization
图 10. 迭代优化的过程

每种股票的权重如表 9 所示。

Table 9. The selected stocks
表 9. 选取的股票

股票名称	投资权重
贵州茅台	0.2497
中国平安	0.2500
兴业银行	0.0400
三安光电	0.2500
中信建投	0.0301
华泰证券	0.0300
海尔智家	0.0300
建设银行	0.0802
中国中免	0.0200
工商银行	0.0200

5.5. 回测与对比

本文将回测的初始金额设置为 100 万人民币。由于上证 50 中贵州茅台为高价股，买入时至少买入 100 股才能进行交易，所以设置贵州茅台的初始资金为 20 万。即在所选股票有贵州茅台的交易中回测金额为 120 万人民币。

为了消除择时买入卖出等策略因素的影响，本文在回测时不在盘中进行操作，仅在每月的第一个交易日，把所选股票买入，持有一个月，在月末卖出，然后将本月股票数据输入 Adaboost 选股模型对下月股票进行排名，再进行调仓，持续六个月。同时，本文还进行了平均权值的回测，即给每支股票相同权值的金额进行买入卖出。将最终得到的结果进行对比来体现投资组合的有效性。

接下来展示 1~6 月份投资组合收益率与平均权值收益率的回测对比，图 11 为投资组合的收益率曲线，图 12 为平均权值为的收益率曲线。其中，蓝色线条为回测曲线，红色线条为上证 50 指数。

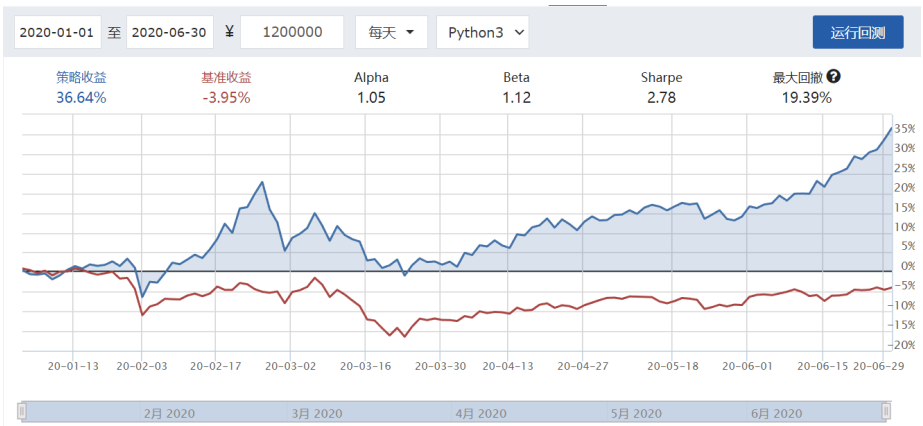

Figure 11. Portfolio strategy returns trend
图 11. 投资组合策略收益趋势

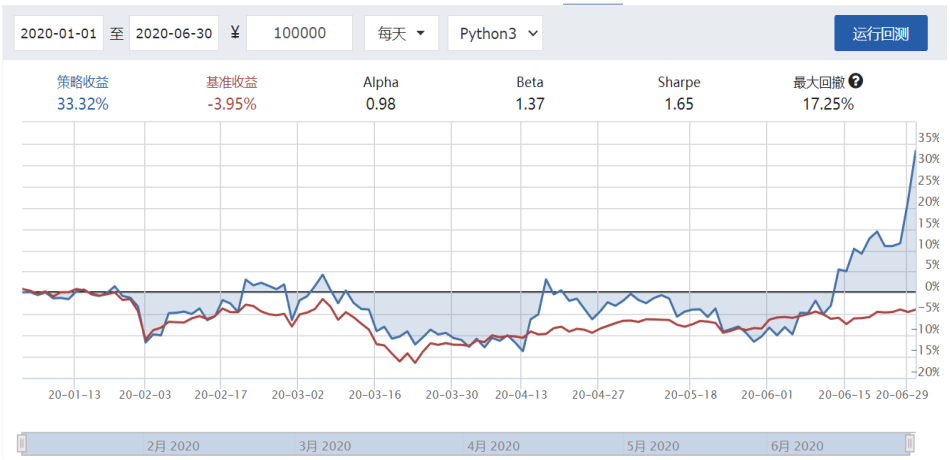

Figure 12. Average weight strategy returns trend
图 12. 平均权重策略收益趋势

从图中可以看出，通过选股模型筛选出的股票，无论是以平均权重买入还是通过组合模型分配权重后买入，其表现都优于上证 50 指数。此外，在组合模型对股票进行加权后，收益率曲线明显更好。接下来，对两种策略的结果进行分析，表 10 对两种策略的数据指标进行了比较：

Table 10. The selected stocks
表 10. 选取的股票

指标名称	投资组合购买策略	平均权重购买策略
策略收益	36.64%	33.32%
策略年化收益	94.85%	84.86%
超额收益	42.24%	38.81%
阿尔法	1.045	0.977
贝塔	1.116	1.373
夏普比率	2.776	1.650
盈亏比	2.005	0.670
最大回撤	19.39%	17.25%
超额收益最大回撤	11.54%	16.30%
日胜率	0.598	0.487
信息比率	4.828	2.380
策略波动率	0.327	0.490

从表中可以看出，投资组合策略在大部分指标上的表现都优于平均权值策略，可见投资组合策略更为合理有效。此外，投资组合策略在行业配置方面也更加合理，如图 13 所示，平均权值的策略资金大都分配到了金融行业，投资组合策略的行业配置则较为分散，对于资金的配置涉及到了较多的行业，有利于行业中性化。

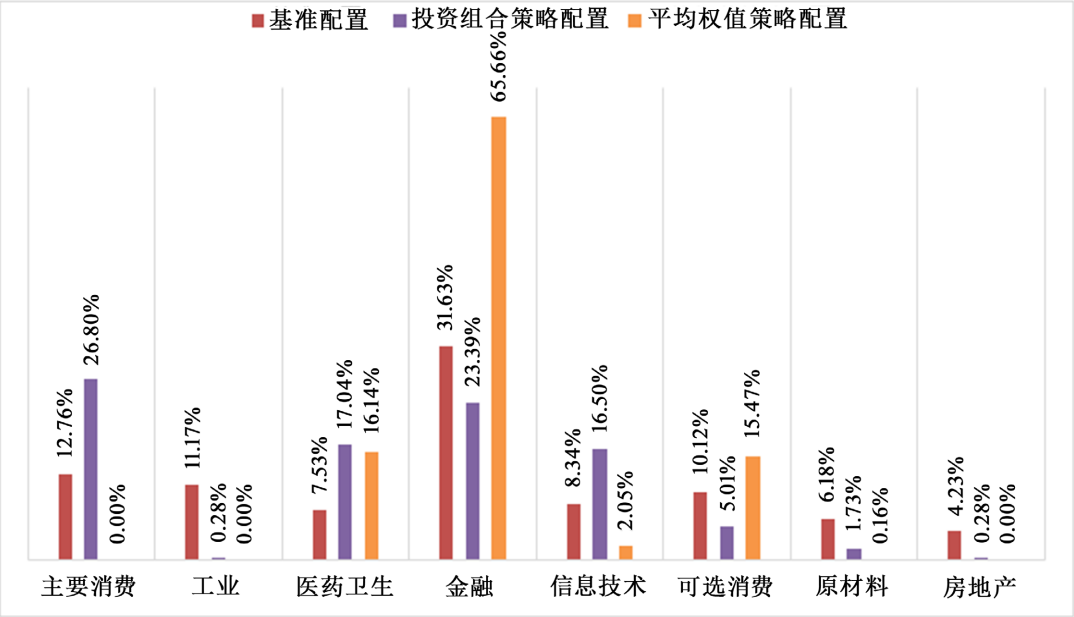

Figure 13. Comparison of industry allocation
图 13. 行业配置比较

6. 总结

本文的主要工作内容及总结如下：

(1) 特征指标库的构建及筛选。本文构建了基于市值类、估值类、成长能力类、情绪类、利润表类、质量类、动量类、每股指标类等 26 个特征指标。经过对特征指标的有效性分析和相关性分析，最终筛选出了 14 个具有解释与预测能力的特征指标。

(2) 基于决策树的 Adaboost 选股模型的构建。本文根据决策树在分类方面的优势，将有效特征指标作为决策树桩的根节点，使用 Adaboost 算法将弱分类器进行组合，建立了一个可以根据当月数据来预测下月股票涨跌的模型。并且使用十折交叉验证及 ROC 曲线验证了模型的准确率与有效性。

(3) 对投资组合模型实证及优化。本文对马科维兹的均值－方差模型进行实证，总结出其理论存在的不足并且对不足进行针对性优化，建立了新的投资组合模型。该模型结合了选股模型的累计类别估计值，对各支股票的初始购买权值进行限定，并且构建了一个更加合理的风险指标。从回测结果看，经过投资组合后的收益不仅高于市场基准，也高于平均权中买入的情形，证明了投资组合的有效性。

参考文献

[1] 王娴, 闫琰. 北交所助力中小企业创新发展[J]. 中国金融, 2021(18): 68-70.
[2] 张晓燕, 张远远. 量化投资在中国的发展及影响分析[J]. 清华金融评论, 2022(1): 44-45.
[3] 姜霁航. 基于大数据时代探究机器学习的发展趋势[J]. 电子元器件与信息技术, 2021, 5(10): 32-33.
[4] 尹兴广. 基于机器学习的投资组合选择研究[J]. 信息系统工程, 2021(12): 128-131.
[5] 刘毅. 因子选股模型在中国市场的实证研究[D]: [硕士学位论文]. 上海: 复旦大学, 2012.
[6] 江方敏. 基于多因子量化模型的 A 股投资组合选股分析[D]: [硕士学位论文]. 成都: 西南交通大学, 2013.
[7] 徐景昭. 基于多因子模型的量化选股分析[J]. 金融理论探索, 2017(3): 30-38.
[8] 曹正凤, 纪宏, 谢邦昌. 使用随机森林算法实现优质股票的选择[J]. 首都经济贸易大学学报, 2014 (2): 21-27.
[9] Markowitz, H. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77-91.

https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

[10] Ross, S.A. (1976) The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic*, **13**, 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

[11] Fama, E.F. and French, K.R. (1993) Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **33**, 3-56. https://doi.org/10.1016/0304-405X(93)90023-5

[12] Patalay, S. and Bandlamudi, M.R. (2021) Decision Support System for Stock Portfolio Selection Using Artificial Intelligence and Machine Learning. *ISI*, **26**, 87-93. https://doi.org/10.18280/isi.260109

[13] Chaweewanchon, A. and Chaysiri, R. (2022) Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning. *International Journal of Financial Studies*, **10**, 64. https://doi.org/10.3390/ijfs10030064

[14] Mba, J.C., Ababio, K.A. and Agyei, S.K. (2022) Markowitz Mean-Variance Portfolio Selection and Optimization under a Behavioral Spectacle: New Empirical Evidence. *International Journal of Financial Studies*, **10**, 28. https://doi.org/10.3390/ijfs10020028

[15] Mynbayeva, E., Lamb, J.D. and Zhao, Y. (2022) Why Estimation Alone Causes Markowitz Portfolio Selection to Fail and What We Might Do about It. *European Journal of Operational Research*, **301**, 694-707. https://doi.org/10.1016/j.ejor.2021.11.036

[16] Gao, J., Guo, H. and Xu, X. (2022) Multifactor Stock Selection Strategy Based on Machine Learning: Evidence from China. *Complexity*, **2022**, Article ID 7447229. https://doi.org/10.1155/2022/7447229