Statistics and Application 统计学与应用, 2025, 14(3), 1-14
Published Online March 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.143052

Hans汉斯

股票选购的策略

蔡天睿，吕　平*

杭州师范大学数学学院，浙江　杭州

收稿日期：2025年2月7日；录用日期：2025年2月27日；发布日期：2025年3月12日

摘　要

本次研究构建了股票选购的全套理论和方法，首先通过简单移动平均线和相对强弱指数选出优质股票，再结合随机森林模型和ARIMA模型预测出这些优质股票将来的收盘价，最后通过二次规划计算出最优的投资组合方案。这套股票选购的方法操作流程简洁高效且通过大量的实践发现其实用性较好，为那些追求短期利润的新手投资者以及相关研究人员提供了参考依据。

关键词

优质股票，随机森林模型，ARIMA模型，二次规划

Strategies for Stock Shopping

Tianrui Cai, Ping Lyu*

School of Mathematics, Hangzhou Normal University, Hangzhou Zhejiang

Received: Feb. 7th, 2025; accepted: Feb. 27th, 2025; published: Mar. 12th, 2025

Abstract

This research constructs a full set of theories and methods for stock shopping, firstly, selecting high-quality stocks through simple moving averages and relative strength indices, then predicting the future closing prices of these high-quality stocks by combining the Random Forest model and the ARIMA model, and finally calculating the optimal investment portfolio plan through quadratic programming. This method is simple, efficient and practical, providing a reference for novice investors and researchers seeking short-term profits.

Keywords

Quality Stocks, Random Forest Model, ARIMA Modeling, Quadratic Programming

*通讯作者。

文章引用: 蔡天睿, 吕平. 股票选购的策略[J]. 统计学与应用, 2025, 14(3): 1-14. DOI: 10.12677/sa.2025.143052

1. 引言

股票投资由于其高风险高回报的特点，一直是人们投资理财的热门选择，然而对于大多数新手投资者而言，面对股票市场的风云变幻往往会选择知难而退[1]，为此，许多研究者基于种种复杂的理论为股票的选购提出了数不胜数的方法，但对于不从事相关研究的普通人而言，过于复杂的分析方法往往由于难以操作实现而显得实用性欠佳。

本次研究希望通过尽可能简单高效的方法，为新手投资者们提供一套股票选购的策略并搭建一套股票投资的理论体系，使其向着实现收益的方向迈出第一步。股票选购所需的三个步骤依次为：选择优质股票，预测这些优质股票之后的收盘价，根据预测出的收盘价选择最优购买方案。下面将一一介绍这三个步骤并提供相应的研究方法，希望对股票投资感兴趣的读者可以在此基础上进行更深入的研究。

2. 挑选优质股票

2.1. 指标的选择

一只优质股票往往具有下面两个特点，一是其收盘价在之前较长一段时间内都是平稳增长的，二是近期其处于价格过低的超卖状态，将来有很强的增长潜力。至于如何判别某只股票是否具有这两个特点，不同的研究人员提出过许多指标，如国信证券工程师焦健等人就提出过以市净率、市盈率、ROA、EPS 一致预期变化率等指标来衡量[2]。但其中大部分指标的计算要用到的数据需要通过收集整理公司官网、分析机构和财经媒体发布的报告来获得，显得十分繁琐。

本次研究将通过简单移动平均线(SMA)来判别股价在之前较长一段时间内是否是平稳增长的，通过相对强弱指数(RSI)来判断股票是否处于超卖状态，计算这些指标要用到的数据可以通过数据挖掘的方法直接从证券交易所网站的历史行情中获得。下面以股票福莱新材 2024 年 1 月 1 日至 2024 年 12 月 30 日的数据为例，研究其收盘价是否是平稳增长的以及在 2024 年 12 月 31 日买入这只股票是否合适。

2.2. 简单移动平均线

要判断某只股票近期是否处于平稳增长的状态，简单移动平均线(SMA)是一个非常实用的工具，它可以帮助我们平滑股价的短期波动，从而更清晰地观察股票的长期趋势。SMA 是通过计算股票收盘价在一定时间内的平均值得到的，n-SMA 表示计算周期为 n 的 SMA，即前 n 个交易日的收盘价的平均值。为了更全面地分析股价的短期波动和长期趋势，选择 $n_1 = 10$ 和 $n_2 = 60$ 两条 SMA 并给出以股票福莱新材为例的可视化结果。

从图 1 可以看出该股票的 60-SMA 在最后 2 个月呈现出平滑、明显的上升趋势，最后 2 个月的收盘价始终在 60-SMA 上方且 60-SMA 的波动较小，这说明福莱新材的收盘价在 2024 年 11 月 1 日至 2024 年 12 月 30 日这段较长的时间内处于平稳增长的状态。另外发现在 2024 年 10 月 10 日附近，10-SMA 向上穿过 60-SMA 并形成了“金叉”，这表明股价的短期走势已经强于长期走势，是收盘价即将上涨的信号，这之后收盘价的一路上涨也说明了这一论断的可靠性。

2.3. 相对强弱指数

虽然 SMA 是一个简单有效的技术指标，但其在单独使用时可能会受到市场噪音的影响。为了提高

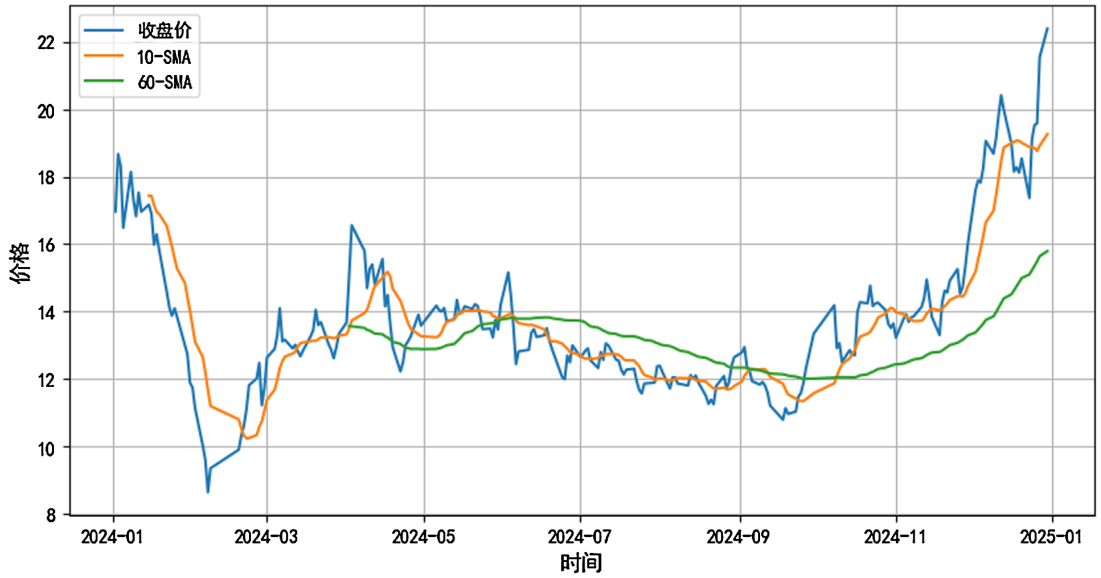

Figure 1. Simple moving average
图 1. 简单移动平均线

分析的准确性，需要结合相对强弱指数(RSI)来进行综合判断。RSI 是一种动量指标，其将收盘价视为体现买方和卖方之间力量大小的度量，并可以以此衡量股票是否处于超买或超卖状态[3]，计算公式如下所示，其中 RS 为在指定的时间范围(一般定为当前交易日及之前 8 个交易日)内股价上涨时段的平均收益除以股价下跌时段的平均损失。

$$\mathrm{RSI} = 100 - \frac{100}{1+\mathrm{RS}}$$

从公式可以看出当股价较为稳定时 RS 的值会在 1 附件波动，此时 RSI 的值会在 50 附近波动；当买方力量占据较大的优势时，收盘价上涨的倾向会大于下跌的倾向，这时 RSI 的值会明显大于 50，但普遍认为当 RSI 超过 70 时股票处于超买状态，股价很可能会马上下调，不建议在此时买入股票；反之，当卖方力量占据较大优势时收盘价下跌的倾向会大于上涨的倾向，此时 RSI 会明显小于 50，普遍认为当 RSI 小于 30 时股票处于超卖状态，股价很可能会马上上调，可以考虑在此时买入股票。下面以股票福莱新材为例，给出其 RSI 曲线的可视化结果。

从图 2 可以看出福莱新材在 2024 年 12 月 30 日处于超买状态，故虽然 2.2 节中的 SMA 显示其在近两个月处于稳定增长状态，但仍不建议马上买入该股票，可以在其收盘价适当下跌(RSI 为 60 左右)时再买入，这可以使收益最大化。

3. 收盘价的预测

通过第 2 节的相关理论，本次研究得到了 7 只优质股票，分别为新炬网络、万和电气、信德新材、祖名股份、时空科技、长华集团和迈信林。下面以新炬网络为例，根据其 2024 年 1 月 1 日至 2024 年 12 月 24 的数据预测其将来的收盘价。

3.1. 选择预测收盘价的方法

收盘价的预测作为整套股票选购理论的核心，许多相关领域的专家和学者为此提出过许多研究方法，这些方法可以分为线性预测类的和非线性预测类的[4]。收盘价作为一个波动性和随机性都较强的时间

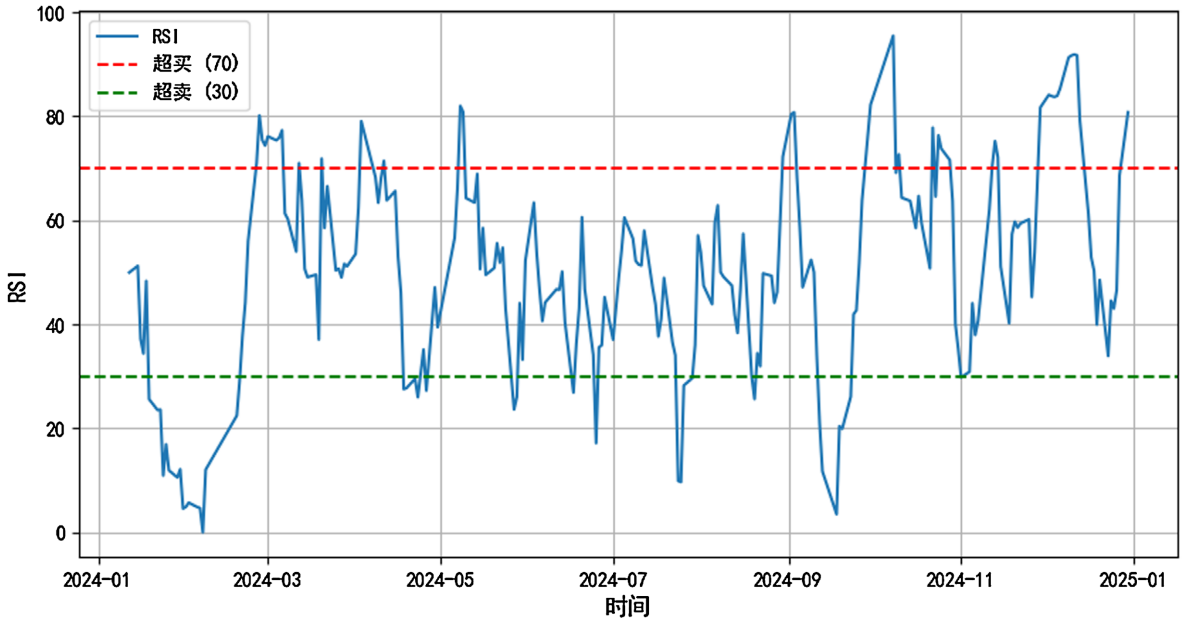

Figure 2. RSI curve
图 2. RSI 曲线

序列，采用非线性的方法对其进行预测已成为研究者们的共识，这是因为这类方法在处理数据中的噪音和捕捉各影响因素之间的交互效应等方面更具优势。在众多非线性的预测方法，随机森林模型因其较强的泛化能力和抗过拟合能力而脱颖而出[5]，Python 中的 RandomForestRegressor 函数可以直接用来构建随机森林模型，使得该方法易操作实现而适合新手投资者掌握运用。

3.2. 构建随机森林模型

通过数据挖掘得到了新炬网络 2024 年 1 月 1 日至 2024 年 12 月 24 的数据，包括开盘价、收盘价、最低价、最高价、成交量、成交金额和换手率。将收盘价作为目标变量，其余 6 个指标作为特征变量。模型参数的调优、特征变量的选择和模型拟合效果的评估是构建随机森林模型的 3 个最重要的步骤，下面分别介绍这三个步骤的操作流程。

3.2.1. 模型参数的调优

RandomForestRegressor 函数中最重要的两个参数是 n_estimators (随机森林中决策树的数量)和 max_features (在寻找最佳分裂时最多使用几个特征变量)。n_estimators 和 max_features 过小都会使模型的稳定性和预测能力较差，前者过大会增加模型的计算成本，后者过大会导致过拟合而降低泛化能力，下面给出将参数调至最优的步骤。

1) 将数据划分为训练集和测试集，本次研究把 2024 年 10 月 1 日之前的数据作为训练集，将其及之后的数据作为测试集。

2) 将除 n_estimators 之外的参数设置为默认值，n_estimators 的值则依次取为 50, 150, 250, ..., 2050，每取一个不同的 n_estimators 就计算出模型训练后的 OOB 误差，OOB 误差是利用那些训练集中由于 Bootstrap 采样而未被选中的本应用来训练决策树的数据(袋外样本)来评估模型性能的指标，可以由训练后的模型直接获得，OOB 误差达到最小时对应的 n_estimators 的值即为其最优取值[6]。

3) 将 n_estimators 的最优值代入 RandomForestRegressor 函数，max_features 依次取 1, 2, ..., n (特征变

量的总个数)，每选取一个不同的 max_features 就通过测试集计算各个节点的均方误差，均方误差达到最小时对应的 max_features 就是其最优取值。

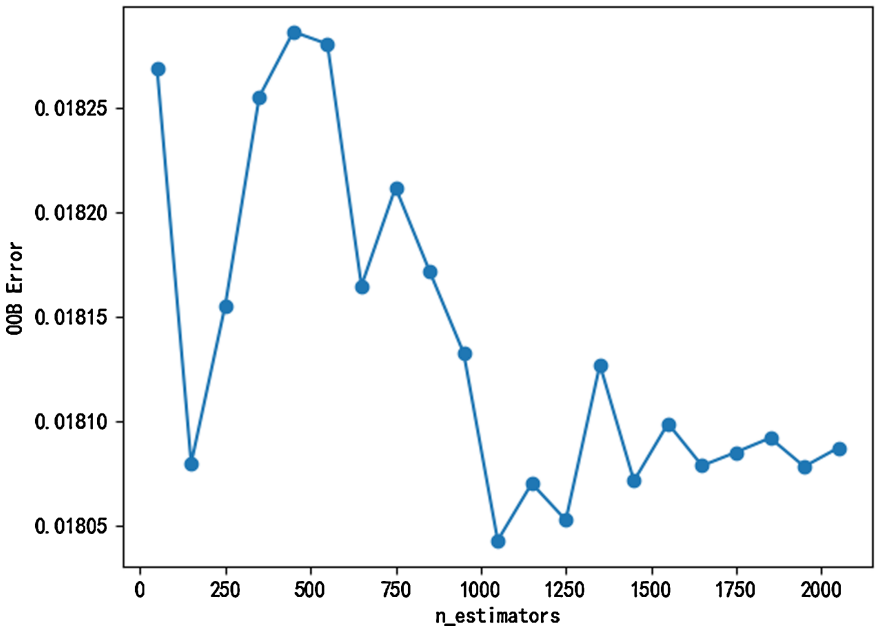

Figure 3. OOB error
图 3. OOB 误差

图 3 是以选取新炬网络全部 6 各特征变量为例给出的 OOB 误差的可视化结果，可以看出 n_estimators 的最优取值在 1000 附近，通过进一步细化得到其最优取值为 1050，将这个值带入 RandomForestRegressor 函数后 max_features 的值依次取 1、2、3、4、5、6，计算出随机森林模型训练后在测试集上的均方误差(MSE)和决定系数(R^2)，从表 1 可以看出 max_features 的最优取值为 5。

Table 1. Select the optimal value of max_features
表 1. 选取 max_features 的最优取值

max_features	MSE	R^2
1	0.410	0.832
2	0.343	0.859
3	0.312	0.872
4	0.288	0.882
5	0.272	0.888
6	0.275	0.877

3.2.2. 特征变量的选取

在构建随机森林模型时选取适当数量的特征变量是十分重要的。选取的数量过少会因欠拟合而导致模型的稳定性和预测性能较差；选取的数量过多则会因过拟合而导致模型的泛化能力较差，此时需要剔除重要性较低的冗余特征变量来提高模型的运行效率[7]，因此需要找到合适的指标来刻画其重要性。

随机森林模型中每棵决策树在构建时会通过分裂特征节点来降低不纯度，在回归问题中通常选择均

方误差(MSE)来刻画不纯度，降低的不纯度越大说明该分裂对降低误差的贡献越大，该特征变量就越重要，这就是可以用均方误差的减少百分比(%IncMSE)来刻画特征变量重要性的原理。同理，特征变量在这个随机森林中对不纯度降低的平均贡献，即其在分裂节点时所导致的不纯度减少量的平均值(IncNodePurity)也可以用来衡量特征变量的重要性。%IncMSE 和 IncNodePurity 的值越大，表明该特征变量在减少节点不纯度方面的能力越强，它就越重要。这两个指标中 IncNodePurity 侧重于特征变量在减少节点不纯度方面的作用，%IncMSE 侧重于特征变量对预测误差的影响。由训练后的随机森林模型可以直接获得所有特征变量的这两个指标的值，下面给出选取特征变量的具体步骤。

1) 将根据 3.21 节计算得到的 n_estimators 和 max_features 的最优取值带入随机森林模型的。

2) 用训练集训练过程中节点不纯度的减少量计算得到 IncNodePurity，由测试集上的初始 MSE 和置换特征变量后的 MSE 增加百分比计算出%IncMSE。

3) 将步骤 2)的计算结果可视化可以更直观地对比所有特征变量的重要性，视图中特征变量的位置越偏向于右上角，就可以初步认为该特征变量越重要。

4) 考虑到%IncMSE 的计算结果会受到随机性的影响，为了使其具有统计学意义，需要通过置换检验对其进行显著性检验，即对每次置换特征变量后的数据重新计算每个特征变量的%IncMSE，通常将置换的次数设为 1000。将共计 1000 次置换后计算出的%IncMSE 大于等于原始%IncMSE 的比例作为该显著性检验的 P 值。

5) 将计算出的 P 值与显著性水平($\alpha = 0.05$)进行比较，如果 P 值小于 α 则认为该特征变量的重要性很强且该结果不是由于随机波动导致的，而是具有统计学意义上的显著性。如果有特征变量不显著，则将它们剔除后根据 3.21 节的理论重新计算已经选取了适当特征变量后的随机森林模型的最优参数。

下面给出新炬网络根据步骤 1)、2)、3)得到的以 IncNodePurity 为横坐标、%IncMSE 为纵坐标的 6 个特征变量重要性评估的可视化结果。

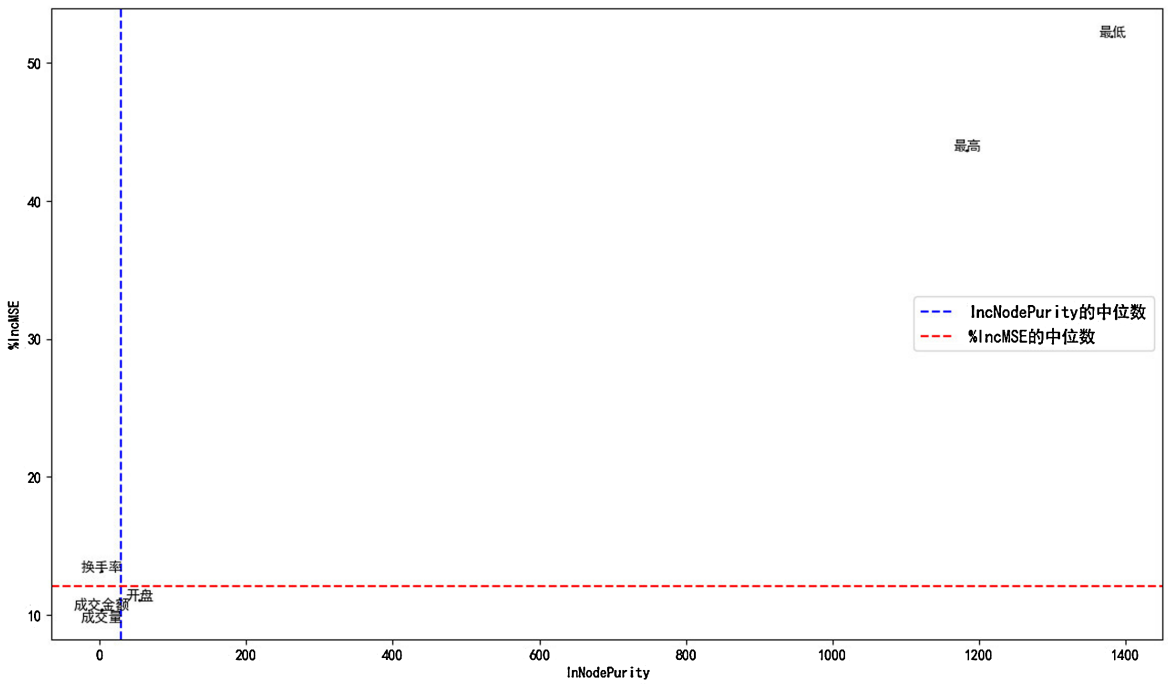

Figure 4. Importance assessment of characteristic variables
图 4. 特征变量的重要性评估

由图 4 可以初步看出最低价和最高价这两个特征变量的重要性明显强于其余 4 个，为了判别另外 4 个特征变量的重要性在统计意义上是否显著，下面给出了全部 6 个特征变量的%IncMSE 显著性检验的 P 值。从表 2 可以看出 6 个特征变量的重要性都是显著的，之前构建的随机森林模型不需要剔除特征变量，其最优参数就是 3.21 节算出来的 n_estimators = 1050，max_features = 5。

Table 2. Significance tests for significance of characteristic variables
表 2. 特征变量重要性的显著性检验

特征变量	IncNodePurity	%IncMSE	%IncMSE 显著性检验的 P 值	重要性是否显著
开盘价	54.717682	11.02987	0.012	显著
最低价	1381.841716	51.90498	0.000	显著
最高价	1184.017559	43.63387	0.000	显著
成交量	3.349442	10.30823	0.012	显著
成交金额	3.493216	10.34906	0.017	显著
换手率	3.386955	13.12956	0.003	显著

3.2.3. 模型评估

将测试集特征变量的数据代入由 3.21 节和 3.22 节得到的已经训练好的随机森林模型，即可得到新炬网络 2024 年 10 月 1 日之后收盘价的预测值，下面先给出预测值与真实值之间的残差密度图，由图 5 发现残差大致服从正态分布并进一步计算出残差的偏度为−0.515、峰度为 0.828。Kline (1998)提出的标准认为如果偏度的绝对值在 3 以内且峰度的绝对值在 8 以内则可认为数据服从近似正态分布。故计算出的残差服从近似正态分布，模型误差主要集中在较小的范围内且没有明显的系统性偏差，预测效果较好。

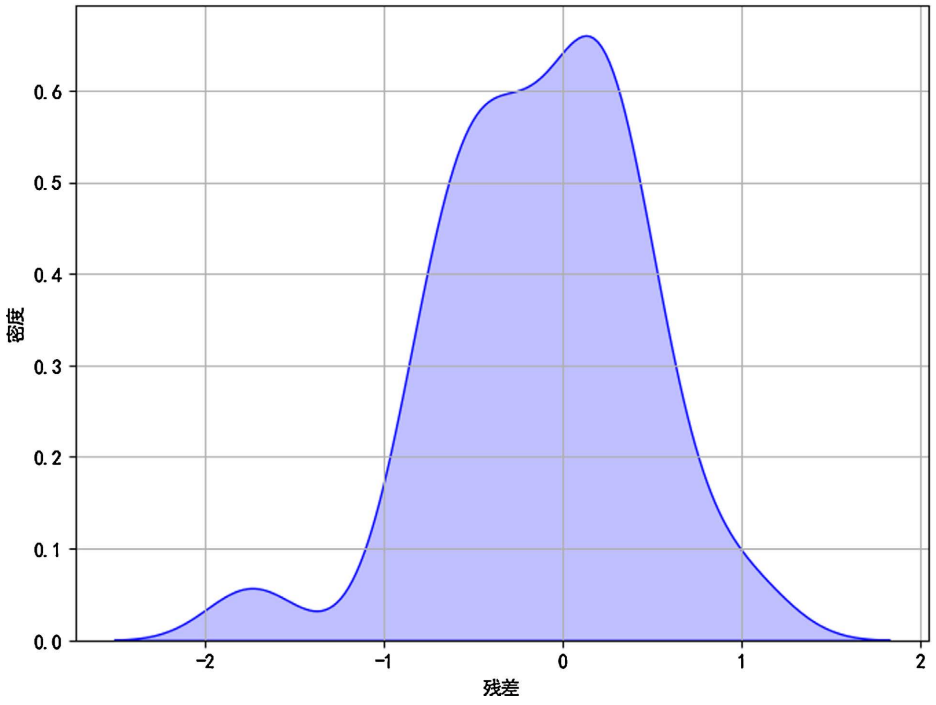

Figure 5. Residual density map
图 5. 残差密度图

为了更直观地展示预测效果，下面给出预测值和真实值之间比较的可视化结果，从图 6 可以看出拟合效果较好，进一步计算出均方误差(MSE) = 0.272，平均绝对误差(MAE) = 0.410，决定系数(R^2) = 0.888。其中 R^2 趋近于 1，说明该随机森林模型通过特征变量预测出的目标变量(收盘价)较为准确。

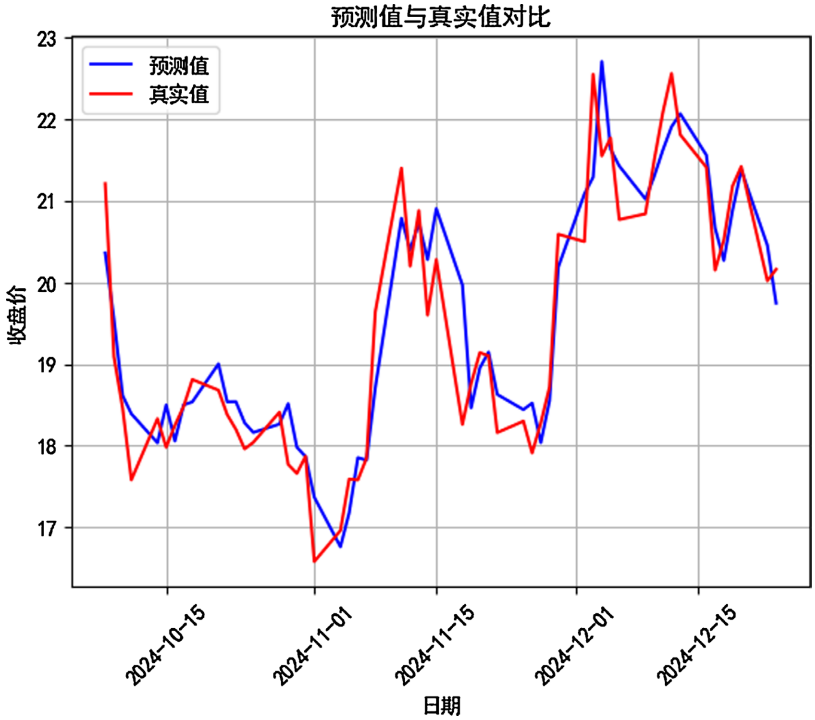

Figure 6. Comparison of predicted and true values
图 6. 预测值与真实值的比较

3.3. 检测随机森林模型的泛化能力

由第 2 节的相关理论得到的优质股票除了新炬网络之外，还有万和电气、信德新材、祖名股份、时空科技、长华集团和迈信林，为了检测 3.2 节构建的随机森林模型是否有较强的泛化能力，用这 6 只股票 2024 年 1 月 1 日至 2024 年 12 月 24 日的数据分别构建随机森林模型，最终得到了这 6 个随机森林模

Table 3. A test of die generalization ability
表 3. 模泛化能力的检验

股票	重要性显著的特征变量	n_estimators 的最优取值	max_features 的最优取值	MSE	MAE	R^2
万和电气	[开盘，最低，最高，成交量，成交金额，换手率]	1345	6	0.021	0.113	0.865
信德新材	[开盘，最低，最高]	225	3	0.687	0.580	0.904
祖名股份	[开盘，最低，最高]	1228	2	0.078	0.203	0.927
时空科技	[开盘，最低，最高]	250	3	0.168	0.331	0.891
长华集团	[开盘，最低，最高，成交金额，换手率]	660	4	0.031	0.129	0.957
迈信林	[开盘，最低，最高，成交金额]	925	4	0.741	0.665	0.868

型各个参数和指标的值。从表 3 可以看出由这 6 只股票构建的预测收盘价的随机森林模型的拟合效果都较好，说明随机森林模型的泛化能力较强，为该模型的广泛应用提供了理论依据。

3.4. 预测未来的收盘价

下面将以新炬网络为例预测其 2024 年 12 月 24 日之后的收盘价，在 3.2 节构建随机森林模型的过程中发现，虽然该模型的预测效果较好，但需要先预测出将来特征变量的值才能预测出收盘价。在众多预测特征变量的方法中 ARIMA 模型具有许多优势，包括运行简便(可以直接调用 python 中 statsmodels 包的 tsa.arima.model 模块中的 ARIMA 函数)，可以通过调整模型的参数来灵活地适应不同特征的时间序列，且即使数据中存在一些噪声，ARIMA 模型也能够提供相对稳定的预测结果。于是本次研究选择使用 ARIMA 模型来预测新炬网络未来 6 个特征变量的值，下面给出建模的具体步骤。

3.4.1. 构建 ARIMA 模型

1) 将原始的时间序列 TS_0 转化为平稳的时间序列。先对 TS_0 进行单位根检验(通常是 ADF 检验)，如果 P 值小于显著性水平($\alpha = 0.05$)则认为 TS_0 平稳，反之对 TS_0 进行一阶差分得到 TS_1 并对 TS_1 进行 ADF 检验，如果 TS_1 仍不平稳则再对 TS_1 进行一阶差分得到 TS_2，为了防止损失过多的信息，应尽量避免差分次数 d 大于 2，最终可以得到平稳时间序列 TS_d。

2) 对 TS_d 进行白噪声检验。如果 P 值小于显著性水平($\alpha = 0.05$)，则认为 TS_d 不是白噪声且其中包含可提取的信息，可以进行下一步 ARIMA 模型的参数定阶；反之说明 TS_d 是白噪声序列，因其没有可提取的规律性信息而使进一步建模失去意义。

3) ARIMA 模型的参数为(p, d, q)，其中 d (差分次数)已经确定了，p 是回归(AR)部分的阶数，q 是移动平均(MA)部分的阶数，可以先通过 TS_d 的自相关(ACF)图和偏自相关(PACF)图初步确定参数(p, q)的大致取值。如果 PACF 图在某个滞后阶数 k_1 之后的系数突然落入两倍标准差中(截尾)，可以认为 $p = k_1$；如果 ACF 图在某个滞后阶数 k_2 之后的系数突然落入两倍标准差中(截尾)，可以认为 $q = k_2$，TS_0 符合 ARIMA (p, d, q)模型。

4) 仅通过 ACF 图和 PACF 图确定的参数(p, q)依赖于观察者的主观判断且缺乏严格的统计检验支持，为了减小模型误差提高拟合效果，通过网格搜索的方式选择 AIC (赤池信息量准则)最小的 ARIMA 模型并对其进行模型检验，如果检验不通过则选择 AIC 第二小的 ARIMA 模型并再次进行模型检验，直到检验通过并得到最优的 ARIMA (p, d, q)模型。

5) 将 TS_0 代入步骤(4)得到的最优 ARIMA 模型并进行模型训练，考虑到 ARIAM 模型只适用于短期预测且较为精确，于是选择用该模型预测未来 3 天的相关数据。

3.4.2. 以新炬网络的最低价为例运行 ARIMA 模型

先对新炬网络 2024 年 1 月 1 日至 2024 年 12 月 24 日最低价的原始数据 TS_0 进行 ADF 检验，检验的 P 值为 0.0316 < 0.05，故认为 TS_0 平稳(可以确定 d = 0)，再对 TS_0 进行白噪声检验(滞后数 lags 依次取 1, 2, ..., 12)，发现该检验的 P 值都小于 0.05，故认为 TS_0 不是白噪声，可以进行 ARIMA 模型的参数定阶，下面先给出 TS_0 的自相关图和偏自相关图。

从图 7 可以看出 PACF 在 1 阶之后突然落入两倍标准差中，可以初步认为 p = 1；而图 8 中的 ACF 是慢慢落入两倍标准差中的(ACF 图呈现拖尾现象)，按理应认为 $q = 0$，但发现 ACF 保持着较高的值，这表明可能存在周期性的自相关而无法直接确定 ARIMA 模型中参数 q 的值，之后通过网格搜索找到了 AIC 最小(为 544.506)的参数组合(p = 1, q = 1)。下面对 ARIMA (1, 0, 1)进行模型检验，主要包括对 AR (ar.L1)和 MA (ma.L1)的 t 检验以及对残差的白噪声检验，具体结果如下。

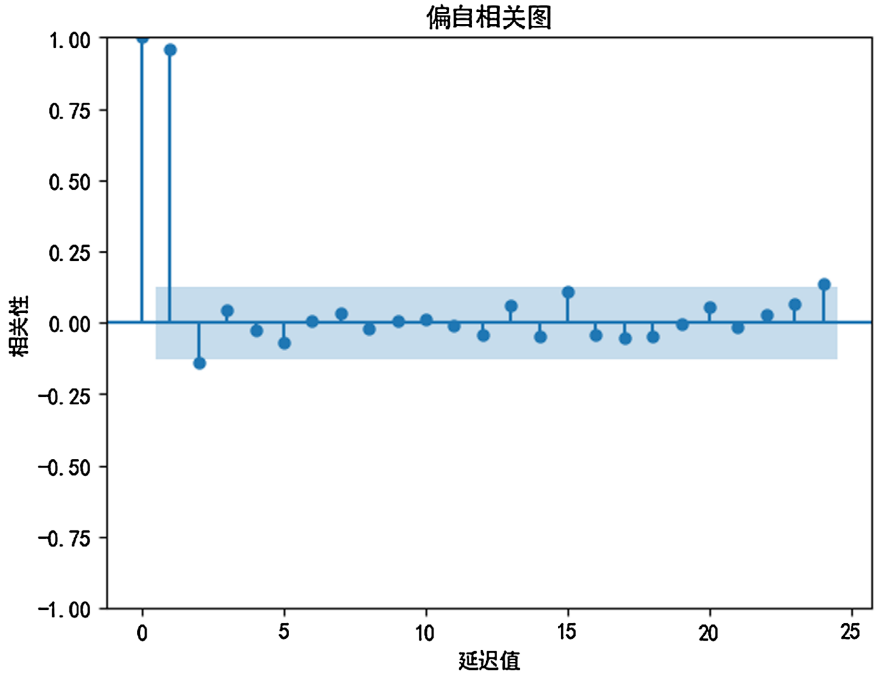

Figure 7. Partial autocorrelation diagram
图 7. 偏自相关图

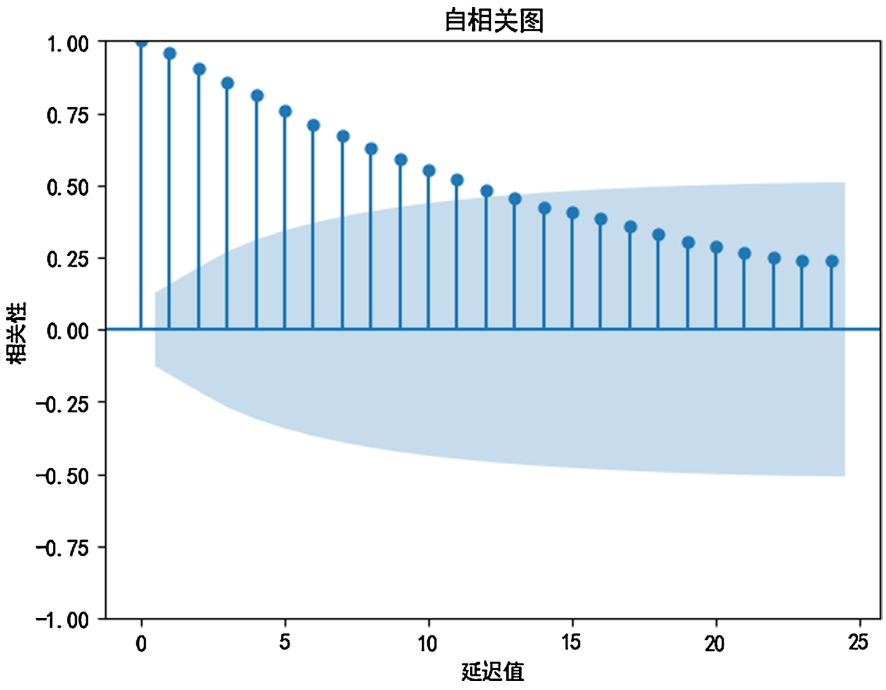

Figure 8. Autocorrelation diagram
图 8. 自相关图

从表 4 可以看出 ARIMA (1, 0, 1)模型通过了模型检验，下面给出 2024 年 12 月 24 日之后 3 个交易日新炬网络最低价的预测值和真实值，从表 5 可以看出 ARIMA (1, 0, 1)模型的预测效果较好，尤其对未来第一天(2024 年 12 月 25 日)预测得较为准确。

Table 4. ARIMA model test results
表 4. ARIMA 模型检验结果

检验对象	检验方法	统计量的值	P 值	检验是否通过
ar.L1	t 检验	72.779	0.000	是
ma.L1	t 检验	4.562	0.000	是
残差	白噪声检验	0.23	0.63	是

Table 5. Predicted and true values of the minimum price
表 5. 最低价的预测值和真实值

交易日	实际的最低价	预测的最低价	误差	相对误差(%)
2024-12-25	19.35	19.468305	−0.118305	0.611394
2024-12-26	19.80	19.517783	0.282217	1.425339
2024-12-27	20.16	19.612354	0.614215	3.043684

3.4.3. 比较 ARIMA 模型和随机森林模型的预测效果

股票的超短线交易可以使投资者根据市场情况灵活地调整交易策略，从而减少长期风险并提供大量的实践机会，在新手股票投资者中这种方式十分流行。此外，由 3.42 节的理论和许多相关资料可以证实 ARIMA 模型对未来第一天的预测值十分精确，而随着预测天数的增加精度会急速下降。下面给出新炬网络 6 个特征变量通过 ARIMA 模型预测出的 2024 年 12 月 25 日的值，可以看出和真实值相比都较为精确。

Table 6. True and predicted values of characteristic variables
表 6. 特征变量的真实值与预测值

特征变量	预测值	真实值	误差	相对误差(%)
开盘价	20.141	20.05	−0.091	0.453865
最低价	19.468	19.35	−0.118	0.609819
最高价	20.217	20.28	0.063	0.310651
成交量	30266	27703	−2563	9.251706
成交金额	5729.651	5506.18	−223.471	4.058549
换手率	0.0176	0.0170	−0.0006	3.529412

显然可以通过 ARIMA 模型直接预测新炬网络 2024 年 12 月 25 日的收盘价，也可以通过将表 6 中 6 个特征变量的预测值带入 3.2 节构建的随机森林模型得到收盘机的预测值。最终得到 ARIMA 模型的预测值为 20.200，随机森林模型的预测值为 19.818，真实值 19.82，可以看出后者的预测结果明显更精确。为了检验随机森林模型预测出的第二天的收盘价比 ARIMA 模型更精确的这个性能是否对大多数股票都适用，表 7 给出表 3 中 6 只优质股票分别通过 ARIMA 模型和随机森林模型预测出的 2024 年 12 月 25 日的收盘价，可以发现随机森林模型在预测第二天的收盘价时都要更加准确，证实了它的这个性能对大多数股票都适用。

4. 股票的投资组合

4.1. “收益率－方差”模型

如果将所有资金只用于购买一只股票会导致风险过大，所以股票投资者们通常会选择适当的投资

Table 7. Comparison of closing price forecasts
表 7. 收盘价预测值的对比

股票	预测值		真实值
	ARIMA 模型	随机森林模型	
万和电气	10.077	10.276	10.21
信德新材	31.100	30.879	30.75
祖名股份	15.601	15.564	15.47
时空科技	14.540	13.987	14.04
长华集团	8.645	9.062	8.97
迈信林	33.217	31.893	32.40

组合方式来降低风险。这其中的方法也是数不胜数，对于新手投资者而言，由哈里·马科维茨(Harry Markowitz)提出的“均值–方差”模型是一个简便高效的投资组合方式，它也是现代投资组合理论的基础。但由于第 3 节的理论已经可以通过构建随机森林模型来预测出第二天收盘价的较为精确的值，于是本次研究对传统的“均值–方差”模型做出一些改进并得到“收益率–方差”模型，使得今天以收盘价购买股票后第二天以收盘价卖出时的收益率在控制风险的情况下达到最优。

本次研究通过第 2 节和第 3 节的理论，由 2024 年 1 月 1 日至 12 月 25 日的数据分别得到了 7 只优质股票(新炬网络、万和电气、信德新材、祖名股份、时空科技、长华集团和迈信林)经随机森林模型预测出的 2024 年 12 月 26 日的收盘价。下面以这 7 只股票为例，给出在 2024 年 12 月 25 日这天最优的投资组合方案。

Table 8. Projected rate of return
表 8. 预测收益率

股票	12 月 25 日的收盘价	12 月 26 日收盘价的预测值	预测收益率
新炬网络	19.82	20.19	1.87%
万和电气	10.21	10.3	0.88%
信德新材	30.75	30.22	-1.72%
祖名股份	15.47	15.91	2.84%
时空科技	14.04	13.88	-1.14%
长华集团	8.97	9.11	1.56%
迈信林	32.4	32.89	1.51%

4.2. 模型构建

设表 8 中 7 只股票的预测收益率依次为$k_1,k_2,...,k_7$，现有 1 单位的资产用于购买这些股票，假设最优的投资组合是 7 只股票购买的比重分别为$x_1,x_2,...,x_7$，显然有$\sum_{n=1}^{7}x_n=1$。此次投资的总收益率为$\sum_{n=1}^{7}k_nx_n$，如果将所有资产用于购买预测收益率最大的股票(祖名股份)，确实会让预测收益率达到最大，但这样没有考虑到风险因素。收益率的协方差矩阵是衡量投资组合中风险大小的一个直观工具，这是因为协方差矩阵显示了两只股票收益率之间的相关性，例如如果某两只股票之间收益率的协方差是正的，说明这两只股票很可能会一起涨或一起跌，同时购买这两只股票的风险就会比较大。下面给出 7 只股票 2024

年 1 月 1 日至 2024 年 12 月 25 日的每个交易日相较于上一个交易日收盘价的收益率的协方差矩阵 A。

$$A=\left(a_{ij}\right)_{7\times 7}=\begin{pmatrix} 0.001914 & 0.000438 & 0.000957 & 0.000648 & 0.001021 & 0.000784 & 0.000877 \\ 0.000438 & 0.000556 & 0.000496 & 0.000291 & 0.000222 & 0.000351 & 0.000461 \\ 0.000957 & 0.000496 & 0.002060 & 0.000650 & 0.000591 & 0.000753 & 0.000841 \\ 0.000648 & 0.000291 & 0.000650 & 0.000841 & 0.000592 & 0.000572 & 0.000496 \\ 0.001021 & 0.000222 & 0.000591 & 0.000592 & 0.002414 & 0.000635 & 0.000845 \\ 0.000784 & 0.000351 & 0.000753 & 0.000572 & 0.000635 & 0.000922 & 0.000629 \\ 0.000877 & 0.000416 & 0.000841 & 0.000496 & 0.000845 & 0.000629 & 0.002347 \end{pmatrix}$$

投资组合总收益率的方差为 $D\left(\sum_{n=1}^{7}k_n x_n\right)=\sum_{j=1}^{7}\sum_{i=1}^{7}x_i x_j a_{ij}$，在金融学中方差是衡量资产收益率波动性的常用指标。方差越大，表示资产收益率的波动性越大，风险也越高。由此可以将选择最优投资组合的问题转化为二次规划的优化问题。该优化问题有两种解决方案，一是使总收益率大于某个预先给定的值 k 并使总收益率的方差达到最小；二是使总收益率的方差小于某个预先给定的值并使总收益率达到最大。选择其中哪种方案取决于投资者的喜好，对于新手投资者而言，他们通常愿意适当减少预期收益率来最小化风险，下面给出该优化投资组合方案的数学模型。

$$\min \sum_{j=1}^{7}\sum_{i=1}^{7}x_i x_j a_{ij}$$

$$\begin{cases} \sum_{n=1}^{7}k_n x_n \geq k \\ \sum_{n=1}^{7}x_n = 1 \\ x_1, x_2, \ldots, x_7 \geq 0 \end{cases}$$

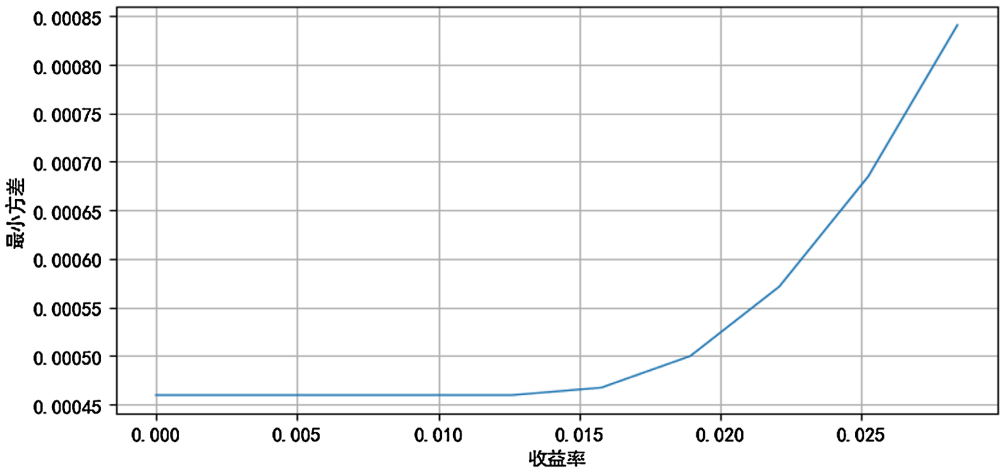

Figure 9. Yield-variance plot
图 9. 收益率－方差图

4.3. 模型运行

从表 8 发现最大收益率为 0.0284，这是 k 的上界，图 9 是总收益率与其最小方差之间的关系图(有效

前沿)。根据有效前沿理论，投资者应该选择位于有效前沿上的投资组合，例如当选择收益率 k 不小于0.018 时，通过 4.2 节的优化模型计算出的最优投资组合为：用 50.7%的资产购买万和电气、用 45.7%的资产购买祖名股份、用 3.1%的资产购买长华集团、用 0.5%的资产购买迈信林。

5. 结语

本次研究构建了全套股票选购的理论。先由第 2 节的理论通过 SMA 和 RSI 选出优质股票，再通过第 3 节构建的随机森林模型和 ARIMA 模型预测出这些优质股票将来一天的收盘价，最后根据第 4 节构建的“收益率–方差”模型计算出给定收益率下风险最小的投资组合方案。这套股票选购的方法操作流程简洁高效且通过大量的实践发现其实用性较好，为那些追求短期利润的投资者以及相关研究人员提供了参考依据。

参考文献

[1] 谢磊. 我国股票投资者投资风险管理研究[D]: [博士学位论文]. 长沙: 中南大学, 2006.

[2] 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, 16(2): 21-27.

[3] 朱宝宪, 潘丽娜. 对相对强弱指数与货币流量指数预测效应的实证研究[J]. 财经论丛(浙江财经学院学报), 2002(2): 45-50.

[4] 赵婷婷, 韩雅杰, 杨梦楠, 等. 基于机器学习的时序数据预测方法研究综述[J]. 天津科技大学学报, 2021, 36(5): 1-9.

[5] 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.

[6] 刘敏, 郎荣玲, 曹永斌. 随机森林中树的数量[J]. 计算机工程与应用, 2015, 51(5): 126-131.

[7] 林娜娜, 秦江涛. 基于随机森林的 A 股股票涨跌预测研究[J]. 上海理工大学学报, 2018, 40(3): 267-273+301.