厦門大學

本 科 毕 业 论 文

基于随机森林回归的 Black-Litterman 模型

A Black-Litterman Model with Views Generated by Random Forest Regression

姓　　名：冯超

学　　号：15220182202396

学　　院：经济学院

专　　业：金融学

年　　级：2018 级

校内指导老师：陈坚　教授

二〇二二年五月四日

摘　要

传统的均值方差模型使用历史收益率数据刻画资产的预期收益与风险，许多研究发现其具有倾向于放大估计错误的弊端。Black-Litterman 模型通过引入投资者主观观点对收益率估计进行修正，在资产配置效率上优于均值方差模型。本文基于随机森林回归算法，以若干技术指标为特征，对美国两大权益指数和债券指数在未来一个月的收益率进行预测，将预测结果作为收益率观点输入到 Black-Litterman 模型中并构造投资组合。实证检验表明：随机森林回归对收益率的预测准确率高于 50%；基于随机森林回归的 Black-Litterman 模型能够取得比均值方差模型更优的绩效表现，前者的平均确定性等价收益比后者高出约 3%，且这一优势在较长的投资期限下也存在。稳健性分析表明，基于随机森林回归的 Black-Litterman 模型对机器学习算法的参数并不敏感，且能够在交易成本小于 0.6% 的情况下取得比均值方差模型更高的确定性等价收益。

关键词：Black-Litterman 模型；机器学习；收益率预测；资产配置模型

Abstract

The classical mean-variance portfolio selection model uses historical return to estimate the expected return and risk of assets. Many researchers have found that the mean-variance model is inclined to amplify the estimation error of expected return and risk. The Black-Litterman model allows investors to incorporate subjective views to revise the estimation, which shows an advantage in terms of asset allocation efficiency. This paper implements the random forest regression algorithm to predict the next-month returns of two major ETFs in the United States, with several technical indicators as inputted features. This paper designs a multi-period asset allocation strategy using the Black-Litterman model incorporated with subjective views generated by random forest regression. The empirical research suggests that: Random forest regression shows a predicting accuracy higher than 50%; The Black-Litterman model based on random forest regression outperforms the mean-variance model, with the former gains certainty equivalent return of around 3% more than the latter. This advantage also applies when the investment horizon is long. The robustness analysis shows that the Black-Litterman model based on random forest regression is not sensitive to the algorithm parameter and can achieve a higher certainty equivalent return when transaction cost is lower than 0.6%.

Key Words: Black-Litterman Model; Machine Learning; Return Prediction; Asset Allocation Model

目 录

Contents

第一章　绪论

1.1 研究背景和意义

Markowitz（1952）提出的均值方差模型（简称 MV 模型）使用历史收益率均值和协方差刻画资产的收益和风险，并通过最优化投资者的效用函数得到最优的投资组合，这一经典的模型标志着现代投资组合理论的开端。在随后的几十年中，有许多学者基于 MV 模型在实证检验和模型改进上进行研究。

许多研究发现，MV 模型在投资实践中暴露出对估计参数敏感和倾向于放大估计错误等问题，这使得 MV 模型的运用受到了一定的限制。针对 MV 模型具有的内在缺陷，Black 和 Litterman（1991，1992）提出了著名的 Black-Litterman 模型（简称 BL 模型）。BL 模型允许投资者提供针对资产预期收益的主观观点，结合贝叶斯理论对预期收益率和波动率进行修正，从而改进投资组合的绩效表现。许多实证研究表明，BL 模型在一定程度上缓解了 MV 模型对估计参数敏感的问题，其也因此在资产管理领域得到了广泛的应用。

然而，投资者依靠自身经验判断而主观生成的观点可能并不准确，这对 BL 模型指导投资实践造成了一定的困难。如何定量地给出准确的投资者观点成为应用 BL 模型的重要现实问题。近年来，随着机器学习技术的快速发展，利用机器学习算法对金融资产相关变量进行预测已经成为一个重要的研究方向。机器学习算法能够凭借计算机强大的计算能力和不受投资者情绪和经验影响等优点，在许多实证研究中被用于预测资产的收益率，这为生成 BL 模型的投资者观点提供了较好的参考。此外，机器学习算法在训练和预测的过程中能够用客观的量化指标对模型的优劣进行评价，这可以用于表达 BL 模型中投资者观点的信心程度。以上分析表明，BL 模型十分适合与机器学习算法生成的投资者观点相结合，在指导资产配置实践上具有较大潜力。

基于以上研究背景，本文使用随机森林回归这一机器学习算法对美国两大权益指数和债券指数的月度收益率进行预测，将预测的收益率作为主观观点与 BL 模型相结合，基于 BL 模型求解各资产权重并构建多期投资组合并进行实证检验。

本文的余下内容按如下结构组织：第二章梳理了资产配置模型和机器学习算法预测收益率两个方面的文献；第三章介绍了本文使用的 BL 模型和机器学习

算法的核心原理，并说明基于机器学习算法构建资产配置策略的详细流程；第四章对本文提出的资产配置策略进行了实证检验，比较了其与基准策略的绩效表现，分析了其在较长投资期限下的绩效表现以及对机器学习算法参数和交易成本的敏感程度；第五章对全文进行总结，并分析了本文的不足之处与改进方向。

第二章　文献综述

2.1 资产配置模型

资产配置的本质是分散投资多种资产的过程，它能够通过调整投资组合中不同资产的比例来平衡投资回报与风险。Markowitz（1952）提出的 MV 模型首次将资产的收益和风险进行量化，开创了量化资产配置领域的先河。MV 模型的核心思想是：在预期风险恒定时最大化预期收益，或者在预期收益恒定时最小化预期风险，得到资产配置的有效前沿曲线。基于对收益率风险的偏好，投资者可以最大化其效用函数，从而构建最优的投资组合。以预期风险恒定时最大化预期收益为例，将收益的波动率固定为 σ^2 且最大化预期收益时，MV 模型可以表达为公式(2-1)的形式。

$$\begin{aligned} &\max w'\mu \\ &\text{s.t. } w'\Sigma w = \sigma^2 \\ &\textstyle\sum_{i=1}^{N} w_i = 1 \end{aligned} \tag{2-1}$$

公式(2-1)中各符号的含义如表2-1所示：

表 2-1: **均值方差模型的符号含义**

符号	含义
N	投资组合中的资产数量。
μ	预期收益向量（$N \times 1$）。μ_i 为第 i 个资产的预期收益，即 $\mu_i = \mathrm{E}[r_i]$。
Σ	资产收益的协方差矩阵（$N \times N$）。$\Sigma_{i,j} = \mathrm{Cov}(r_i, r_j)$。
w	资产权重向量（$N \times 1$）。w_i 为第 i 个资产在投资组合中的权重。

从以上最优化的过程来看，MV 模型给出的资产权重可以认为是理性投资者的最优资产配置策略。然而，许多实证研究发现，基于 MV 模型构建的投资组合在样本外的绩效表现并不理想（Michaud，1989）。一个被学者们广泛认可的解释是，MV 模型对输入的估计参数十分敏感，在求解最优化问题的过程中倾向于放大对预期收益和风险的估计误差（Lim 等，2011）。Broadie（1993）指出，传统的 MV 模型使用资产的历史收益率均值和方差代表预期收益和风险可能并不是最合适的选择：在样本量较小的情况下，参数估计误差将十分严重；在样本量较大的情况下，由于参数的非平稳性，即历史收益率并不能准确地代表资产未来的绩效表现，MV 模型对预期收益和风险的估计也可能存在误差。

针对 MV 模型存在的对收益和风险估计不准确的问题，一些研究对其估计输入参数的过程进行了改进。

一些学者从优化估计参数的算法这一角度对 MV 模型进行改进。Ledoit 和 Wolf（2003，2004）提出一种对收益率样本协方差矩阵进行收缩的方法，发现收缩后的协方差矩阵可以用于改进对预期风险的估计。他们将改进后的协方差矩阵作为 MV 模型的输入，在美国指数市场的投资中取得了更稳健的样本外投资表现。这一做法被许多文献所采用（Donthireddy，2018），因此本文对于收益率协方差矩阵的估计也是基于 Ledoit 和 Wolf（2004）提出的收缩算法。

亦有学者从引入新的收益风险量化指标的角度对 MV 模型进行改进。Rom 等（1994）认为资产收益率并不服从对称的正态分布，并使用收益率下行风险来替代收益率方差。Joro 和 Na（2006）通过添加资产收益偏度这一维度，考虑投资者对资产收益正偏度的偏好，在 MV 模型的基础上提出均值-方差-偏度模型，从而改进 MV 模型对于风险的估计。

与替换或添加统计指标来改进 MV 模型的参数估计不同的是，Black 和 Litterman（1991，1992）在贝叶斯框架下将投资者对资产预期收益的主观观点与 MV 模型结合，提出了更加符合直觉的 BL 模型。几十年来，BL 模型经过了许多学者的检验和改进。Satchell 和 Scowcroft（2000）对 BL 模型的公式推导进行了详细的介绍，并展示了丰富的数值案例以说明 BL 模型的优势；Caliskan（2012）比较了 MV 模型和 BL 模型，发现基于 BL 模型构造的投资组合具有更低的系统性风险；Idzorek（2019）针对观点不确定性难以度量的问题，提出了一种更加易于实操的方法，投资者只需对主观观点提供一个介于 0 到 1 之间的信心程度指标，就可以应用 BL 模型给出权重。BL 模型凭借其符合直觉、易于计算和能够有效改善资产配置效率等优点，在学界和业界都得到了广泛的应用。

2.2 机器学习算法预测收益率

近年来，随着大数据技术的发展和计算机运行效率的提升，各种机器学习算法的理论和应用吸引着众多领域的学者进行相关的研究。在金融领域，以金融资产价格预测方面的研究最具代表性（赵琪等，2020）。与投资者凭借其主观经验对未来价格趋势做出的判断不同的是，机器学习算法是从历史数据中挖掘规律并对未来价格趋势给出定量预测，其训练与预测的过程不受人的认知或情绪变化的影响，因此具有客观性的优势。此外，机器学习算法能够挖掘出各种数据之间的非线性关系，与传统的计量和统计模型相比更具优势（Gu 等，2020）。

学术界关于机器学习预测收益率方面已有丰富的讨论。本文主要从各研究

使用的算法、特征变量的选取和模型评价指标这三个方面展开综述。

2.2.1 机器学习算法

从已有研究使用的机器学习算法来看，监督学习是最常被用于预测收益率的。Kumbure 等（2022）对近 20 年的相关研究进行整理和回顾后发现，有约 43% 是基于分类算法进行预测的，约 55% 是基于回归算法进行预测的。

基于分类算法展开的研究大多是预测资产价格的涨跌方向。例如：方匡南等（2010）使用多种方法预测我国基金超额收益率方向，发现基于随机森林的交易策略表现较好。Khaidem 等（2016）使用随机森林算法对股票价格的涨跌进行预测。Liew 和 Mayster（2017）使用支持向量机、随机森林和深度神经网络这三种机器学习算法，对美国主要的 ETF 在长期和短期的价格涨跌进行预测，并发现在 1 至 3 个月的预测区间内预测效果较好。Yuan 等（2020）使用支持向量机和随机森林算法对中国 A 股市场进行涨跌趋势预测。

基于回归算法展开的研究大多是定量地预测资产的价格或收益率。例如：Polamuri 等（2019）使用多元线性回归、支持向量回归和随机森林回归对股票价格进行预测，发现随机森林回归的预测效果较好。Huang 和 Tsai（2009）使用支持向量回归对台湾股票在单日的具体价格进行预测。亦有许多学者使用了 LSTM 和 CNN 等神经网络算法对资产价格进行预测，如 Liu 和 Long（2020）、Lu 等（2020）和 Jing 等（2021）。

2.2.2 特征变量

从机器学习算法的输入特征数据来看，各研究最常使用的是技术指标、宏观经济指标和资产的基本面指标，近年来新闻舆情指标等另类数据也被越来越多的学者研究（Kumbure，2022）。使用技术指标作为特征变量的研究有李斌等（2017）、Lu 等（2020）和林耀虎等（2022），他们大多将基本的量价指标和技术分析理论结合，构建复杂的技术指标作为特征变量。使用宏观经济指标和基本面指标的研究也有很多，如 Barak 和 Modarres（2014）和 Yuan 等（2020），他们利用 CPI、失业率、汇率和进出口数据等作为宏观经济指标，用市盈率、速动比率和 ROA 等财务数据作为基本面指标，以对收益率进行预测。

2.2.3 模型评价指标

从机器学习算法预测收益率的相关评价指标来看，基于准确度和错误率的评价指标最常被使用（Kumbure，2022），如均方根误差（RMSE）、精确度（Accuracy）、平均绝对百分比误差（MAPE）等。对收益率涨跌方向做分类预测的算法通常使

用精确度这一类评价指标，而使用回归预测收益率具体数值的算法通常使用误差这一类评价指标。考虑到 BL 模型需要输入的观点信心程度介于 0 到 1 之间，本文选取 R 方这一通常介于 0 到 1 之间的指标对训练的模型进行评价，并将其作为观点信心程度。

第三章　理论模型

3.1 Black-Litterman 模型

BL 模型优化资产配置的过程与 MV 模型大致相同，但在估计参数的选取上进行了改进。BL 模型使用市场隐含的均衡收益率作为先验收益率，加入主观观点后对预期收益和风险进行估计。先验收益率和后验收益率分别如公式(3-1)和公式(3-2)所示：

$$\Pi = \lambda \Sigma w_{mkt} \tag{3-1}$$

$$\mathrm{E}[R]_{BL} = \left[(\tau\Sigma)^{-1} + P^T\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + P^T\Omega^{-1}Q\right] \tag{3-2}$$

公式(3-1)和(3-2)中各符号的含义如表3-1所示。

表 3-1: **Black-Litterman 模型的符号含义**

符号	含义
Π	先验收益率向量（$N \times 1$），用市场隐含均衡收益率衡量。
λ	风险厌恶系数。
Σ	超额收益的协方差矩阵（$N \times N$）。
w_{mkt}	各资产的市值权重向量（$N \times 1$）。
$\mathrm{E}[R]_{BL}$	结合主观观点的后验收益率向量（$N \times 1$）。
K	观点个数。
P	观点矩阵（$K \times N$）。每一行代表一个观点涉及的资产。
Q	观点矩阵（$K \times 1$）。每一行代表一个观点对应的收益率。
Ω	观点不确定性矩阵（$K \times K$）。
τ	观点权重标量，用于衡量各信息的权重。

对于 BL 模型公式的理解，Idzorek（2019）在研究中指出：直觉上，BL 模型得到的后验收益率可以认为是将先验收益 Π 和主观观点 Q 进行复杂的加权平均后的结果，其中的权重是关于观点不确定性矩阵 Ω 和观点权重标量 τ 的函数。图3-1展示了 BL 模型的计算过程。

在公式(3-2)中，标量 τ 被用来衡量观点的不确定性。τ 的值越大，说明观点的可靠程度越高。但学术界对 τ 的具体取值并没有定论，如 Satchell 和 Scowcroft（2000）认为 τ 通常应该取 1，而 Lee（2000）将 τ 设置在 0.01 到 0.05 之间。Idzorek（2019）提出的方法是使用一个介于 0 到 1 之间的信心程度指标，使得最终配置的权重在观点完全无效和完全准确之间进行放缩，免去了讨论 τ 的取值问题这一过

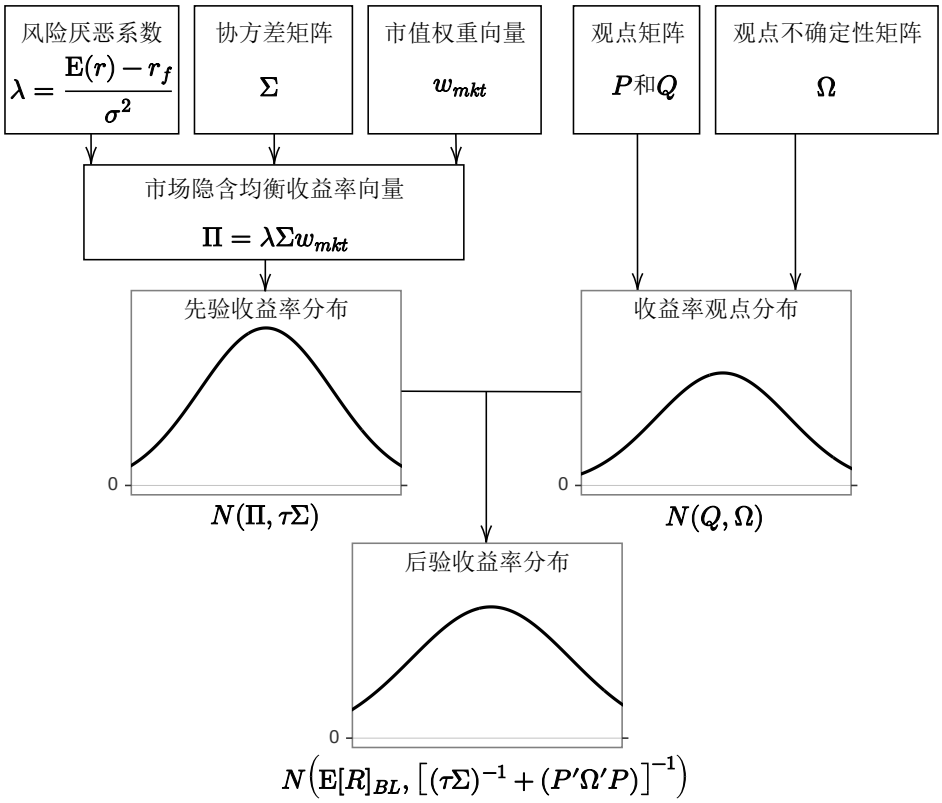

图 3-1: **Black-Litterman 模型计算示意图**

程，因此可以有效地解决观点不确定性难以度量的问题。由于 Idzorek（2019）提出的方法的简便性和易于实操性，本文使用的 BL 模型均是基于 Idzorek（2019）改进后的形式，并使用机器学习算法中产生的 R 方这一通常介于 0 到 1 之间的模型评价指标作为观点信心程度。

3.2 随机森林回归

随机森林（Random Forest，简称 RF）由 Breiman（2001）提出，是一种以回归树作为基预测器的集成学习算法。RF 的核心思想是，利用 Bootstrap 方法随机抽取 K 个特征变量作为一个训练样本，对每一个样本用一个回归树进行拟合，最后将所有回归树的拟合结果取平均值作为预测值。

在实证检验中，本文使用 50 个回归树作为基预测器，并在稳健性分析部分考察预测效果对基预测器个数的敏感程度。衡量每一个回归树决策质量的指标为均方误差（MSE，如公式3-3所示），即每次分枝时以最小化子结点的均方误差为目标。

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3-3}$$

随机森林算法在本文的收益率预测应用上有一定的优势：第一，相比于传统的线性回归算法，随机森林回归可以基于回归树捕捉特征与标签之间的非线性关系。第二，相比于深度神经网络等复杂的非线性模型，随机森林回归在样本量并不庞大的情况下（如本文的 100 个交易日）不容易出现过拟合的问题。第三，随机森林回归能够基于信息增益对每个特征变量的重要程度进行量化，增强了模型的可解释性。

3.3 特征工程

参照李斌等（2017）的研究中对技术指标重要性的排序，本文使用 OBV、CMO、K 值、J 值、MACD 和 SMA 这六个技术指标，并生成滞后 0、1、2、3、5、10、15 和 30 个交易日的数据作为特征变量。各技术指标的构建公式在附录中。

3.4 资产配置策略

以技术指标为特征变量，以资产在一段时间的收益率作为标签，随机森林回归算法可以对资产收益率进行预测。参照 Liew 的研究，机器学习算法对资产在未来 1 至 3 个月的预测区间内预测效果较好，因此本文使用 22 个交易日（约一个月）作为预测区间。

为了与现实中的投资决策过程相符合，本文采用滚动窗口的方式构建投资组合，即在每月底对未来 22 个交易日的收益率进行预测，并将预测值作为 BL 模型的输入，以确定各资产在下一个月的初始权重。

为了使预测过程更加科学可靠，我们将数据划分为训练区间、验证区间、暂停区间和预测区间。各区间上的运行过程和意义为：利用训练区间上的特征和标签对随机森林回归的参数进行拟合；用拟合好的模型在验证区间上做预测，并根据验证区间上标签的预测值和真实值计算 R 方。R 方的值一般在 0 到 1 之间，因此十分适合作为投资者观点的信心程度；设置暂停区间是为了防止预测过程中用到未来的收益率信息而高估模型的准确率，可以保障预测的合理性；预测区间为每月的最后一个交易日，我们使用这一天的特征变量作为输入，用拟合好的模型预测各资产在未来 22 个交易日的收益率，以此作为 BL 模型的投资者收益率观点。

本文选取各区间的规则为：预测区间为每月的最后一个交易日；暂停区间为预测区间之前的 22 个交易日，这与标签数据中收益率对应的区间长度一致，可以防止训练区间使用到未来的收益率信息；训练区间和验证区间分别为预测区间之前的 100 个交易日中的随机 70 天和 30 天，即训练区间和验证区间长度

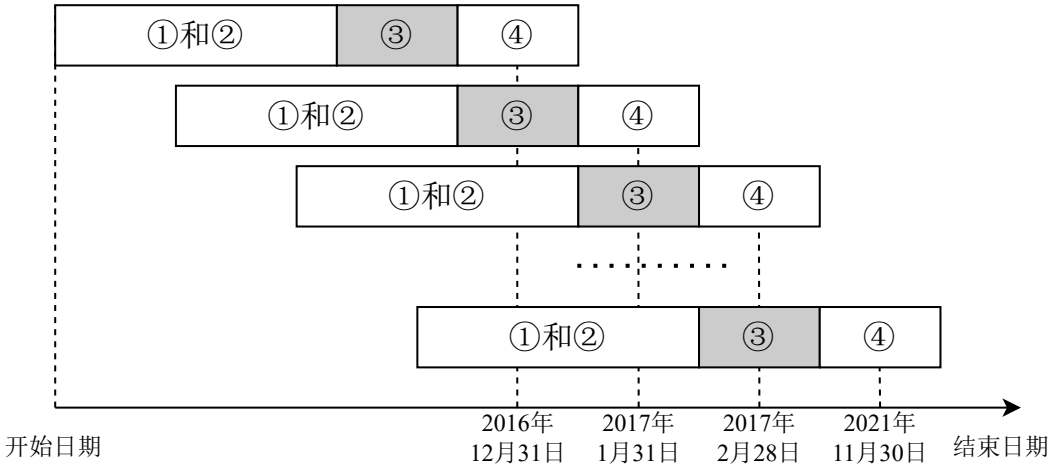

图 3-2: **滚动窗口及各区间的划分情况**

之比为 7 比 3。图3-2展示了滑动窗口下的模型训练与预测过程。其中，①、②、③和④分别代表训练区间、验证区间、暂停区间和测试区间。

图3-3是实证检验过程中预测收益率与资产配置的流程图。我们对于每一个月底，先按照滑动窗口的划分方法，结合特征和标签数据在各区间上进行训练与预测，得到收益率观点及其信心程度，再将收益率观点输入到 BL 模型中，得到各资产在下一个月的初始权重。

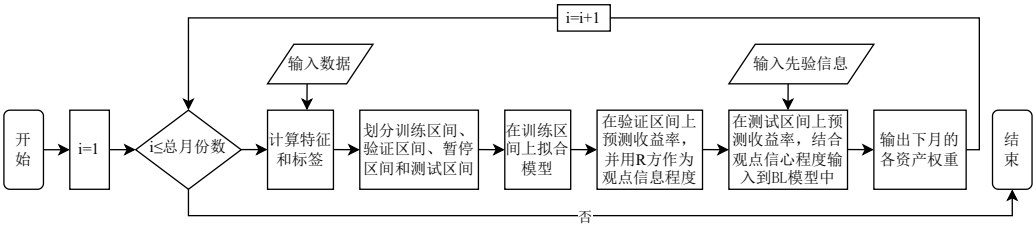

图 3-3: **预测收益率与资产配置流程图**

第四章　实证检验

4.1 数据

考虑到美国市场数据的可获得性较好，且宽基指数相关数据的信噪比较高，本文选取标普 500ETF（证券代码为 SPY）和 iShares 7-10 年期国债指数（证券代码为 IEF）作为投资标的，在 2017 年 1 月 1 日至 2021 年 12 月 31 日这 5 年进行回测检验。相关数据均来自雅虎财经。各特征和标签变量的描述性统计如表4-1和表4-2所示。

表 4-1: **特征和标签变量描述性统计-SPY**

变量	样本数	平均值	标准差	最小值	中位数	最大值
MACD	1238	1.45	3.34	-22.85	1.85	8.99
OBV	1238	6.67e+08	1.14e+09	-1.69e+09	6.03e+08	3.00e+09
CMO	1238	16.79	22.70	-64.66	20.20	74.38
K	1238	-109.12	176.94	-828.12	-50.64	91.41
J	1238	-108.23	191.21	-1134.49	-46.07	143.49
SMA	1238	295.60	67.60	206.17	274.27	466.09
Return	1238	17.98%	55.65%	-383.08%	24.79%	288.38%

表 4-2: **特征和标签变量描述性统计-IEF**

变量	样本数	平均值	标准差	最小值	中位数	最大值
MACD	1238	0.10	0.41	-1.00	0.05	1.96
OBV	1238	1.39e+08	1.23e+08	-2.77e+07	1.26e+08	3.70e+08
CMO	1238	4.21	23.45	-67.47	5.20	78.36
K	1238	-285.67	268.42	-1085.37	-197.45	70.80
J	1238	-284.55	289.97	-1269.86	-190.58	182.83
SMA	1238	106.47	8.90	95.34	105.58	120.87
Return	1238	3.70%	16.98%	-40.93%	2.75%	91.37%

4.2 基准策略

为了更好地评价基于随机森林回归的 BL 模型的绩效表现，本文使用传统的均值方差模型和完美预测收益率下的 BL 模型作为对比，共回测三个资产配置策略。各策略的英文简称及描述如表4-3所示。

表 4-3: 三个资产配置策略

英文简称	描述
MV	传统的均值方差模型
BL-RF	基于随机森林回归的 BL 模型
BL-Perfect	完美预测收益率的 BL 模型

传统的均值方差模型以历史收益率均值和收缩协方差（时间窗口为 125 个交易日（约半年），下同。）作为输入；基于随机森林回归的 BL 模型以市场均衡收益率和收缩协方差作为输入，并结合了随机森林回归预测的收益率观点及其信心程度；完美预测收益率的 BL 模型加入的收益率观点是真实的次月收益率，且观点信心程度为 100%。完美预测收益率的 BL 模型是以一种事后的视角，考察当投资者能够完全精准地预测次月收益率时 BL 模型的资产配置效果。实际上，投资者并不能完全精准地预测收益率，即完美预测收益率的 BL 模型并不能指导现实投资实践，但可以为 BL 模型的有效性提供参考。

在资产配置的过程中需要最优化投资者的效用函数，本文将投资者的风险厌恶系数设为 3，即投资者的效用函数如公式4-1所示：

$$U(\mathrm{E}(r), \sigma(r)) = \mathrm{E}(r) - \frac{3}{2}\sigma^2(r) \tag{4-1}$$

4.3 收益率的预测准确率

基于随机森林回归的 BL 模型本质上是通过提高对资产收益率的预测准确率，从而提高投资组合的绩效。因此，本文首先考察随机森林回归在每月底预测收益率时相比 MV 模型有多大的改进。由于回归预测的结果是精确数值，难以直接判断预测准确率，因此本文按照如下方法定义收益率预测的准确与否：若随机森林回归预测的收益率在 MV 模型预测的收益率的基础上，向真实的收益率方向偏移，则该预测为准确，反之则不准确。在数学上，这一定义也可表示为公式4-2：

$$\text{Accuracy} = \begin{cases} 1 & \text{若 } (r_{true} - r_{MV})(r_{BL} - r_{MV}) \geq 0 \\ 0 & \text{若 } (r_{true} - r_{MV})(r_{BL} - r_{MV}) < 0 \end{cases} \tag{4-2}$$

按照公式4-2计算发现，随机森林回归对 SPY 和 IEF 收益率的预测准确度分别为 56.67% 和 55%，均大于 50%。因此，可以认为随机森林回归在一定程度上提高了对收益率的预测准确率。

4.4 特征变量的权重排序

回归树是随机森林回归中的基预测器，它根据特征变量对模型带来的信息增益大小进行分枝，各变量带来的信息增益大小可以表示该变量的重要程度。随机森林回归是基于多个回归树的集成学习模型，对每一个基回归树中的特征变量权重求平均值，可以得到各特征变量的平均权重。本文以各特征变量的平均权重分析各特征变量的重要程度，以增强模型的可解释性。

为举例说明，表4-4展示了 SPY 和 IEF 在最后一个调仓日期（即 2021 年 11 月 30 日）的训练区间中前十大权重的特征变量。各特征变量的右下角标代表滞后的交易日天数。从表中可以看出：对 SPY 来说，近期的 MACD 和 OBV 对随机森林回归结果的影响程度较大；对 IEF 来说，MACD 和 SMA 对随机森林回归结果的影响程度较大。

表 4-4: 2021 年 11 月 30 日对应训练区间中前十大权重的特征变量

SPY			IEF		
特征变量	平均权重	标准差	特征变量	平均权重	标准差
OBV_3	43.52%	28.87%	$MACD_{30}$	19.81%	25.65%
$MACD_3$	12.05%	12.97%	SMA_{15}	19.63%	25.88%
$MACD_2$	6.42%	15.64%	SMA_{10}	12.41%	23.84%
OBV_2	5.50%	15.38%	SMA_1	8.56%	20.75%
OBV_0	5.38%	16.00%	OBV_0	6.35%	10.65%
$MACD_1$	4.23%	10.62%	SMA_5	4.88%	15.65%
OBV_1	2.86%	12.34%	SMA_0	4.77%	15.51%
$MACD_0$	2.73%	6.17%	OBV_{10}	4.62%	10.56%
K_{30}	2.39%	3.26%	SMA_2	2.60%	12.42%
SMA_5	1.86%	6.18%	CMO_{30}	2.33%	8.77%

4.5 投资组合绩效分析

本文首先在不考虑交易成本的情况下，对本文4.2部分中涉及的三种资产配置策略在 2017 年 1 月 1 日至 2021 年 12 月 31 日进行回测，并计算各年的年化收益率（Return）、年化波动率（Volitility）、确定性等价回报（CER），最大回撤（MD）、夏普比率（Sharpe）、卡玛比率（Calmar）和换手率（Turnover），取 5 年的平均值汇总如表4-5所示。其中，确定性等价回报是根据公式4-1求出的在无风险情况下能够与各策略获得同等效用的年化收益率。

从绩效指标中可以看出：BL 模型-RF 在平均年化收益率指标上高出 MV 模型约 3%；在风险控制上，BL 模型-RF 的平均波动率和最大回撤较 MV 模型略大；

在经风险调整后的绩效指标上，BL 模型-RF 的平均确定性等价回报、夏普比率和卡玛比率都更高。作为对比，BL 模型-Perfect 的各项绩效指标都表现最优，这也说明了 BL 模型在投资者观点正确时是十分有效的。

表 4-5: **各投资组合的绩效指标**

	Ret	Vol	CER	MD	Sharpe	Calmar	Turnover
MV	10.64%	11.23%	8.75%	8.84%	1.10	2.71	4.15
BL-RF	14.05%	11.43%	12.09%	9.17%	1.57	3.77	8.92
BL-Perfect	39.57%	10.63%	37.87%	4.44%	3.68	9.28	10.11

MV 模型和 BL 模型-RF 的归一化净值和 SPY 的权重变化分别如图4-1和图4-2所示。可以看出，BL 模型-RF 在近三年的表现比 MV 模型更好。从两个策略中各资产的权重变化情况来看，在多数时间段内，MV 模型和 BL 模型-RF 对权益资产和债券资产配置了同等的权重，但在 2019 年和 2020 年的一些区间配置了相反的权重。结合单个资产的绩效表现看：债券资产在 2019 年的表现更好；权益资产在 2020 年初虽有大幅下跌，但随后上涨较大。MV 模型仅根据历史收益率判断未来预期收益率，而 BL 模型-RF 可以基于特征变量的变化调整资产的预期收益率，因此在波动较大的时期取得了更好的绩效表现。

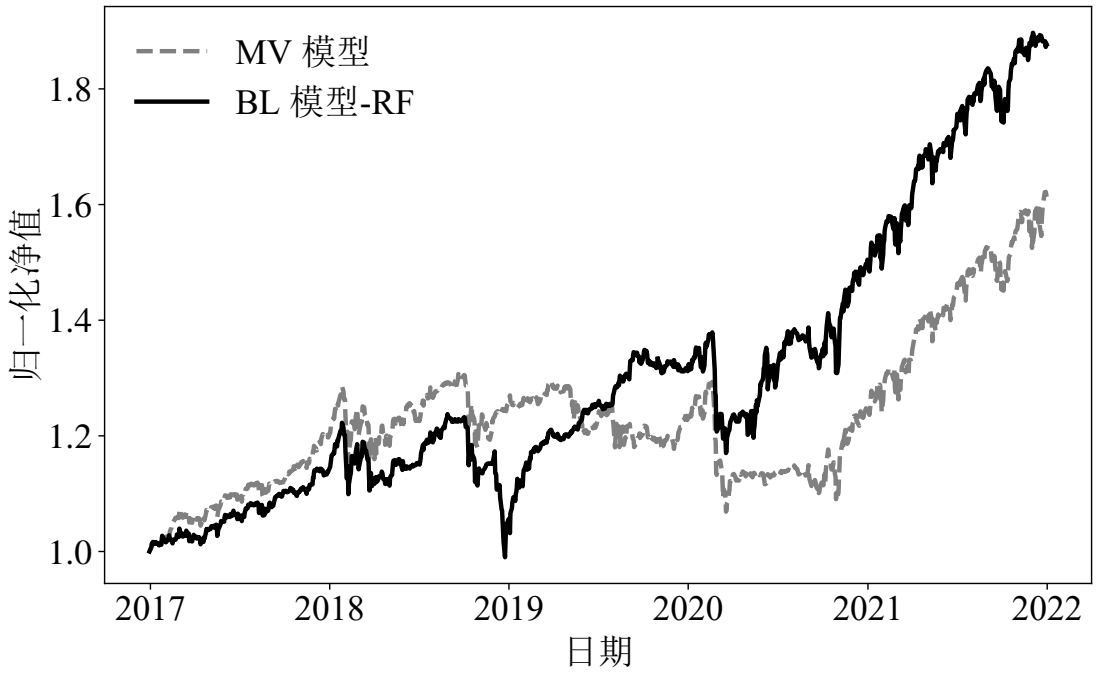

图 4-1: **MV 模型和 BL 模型-RF 的归一化净值**

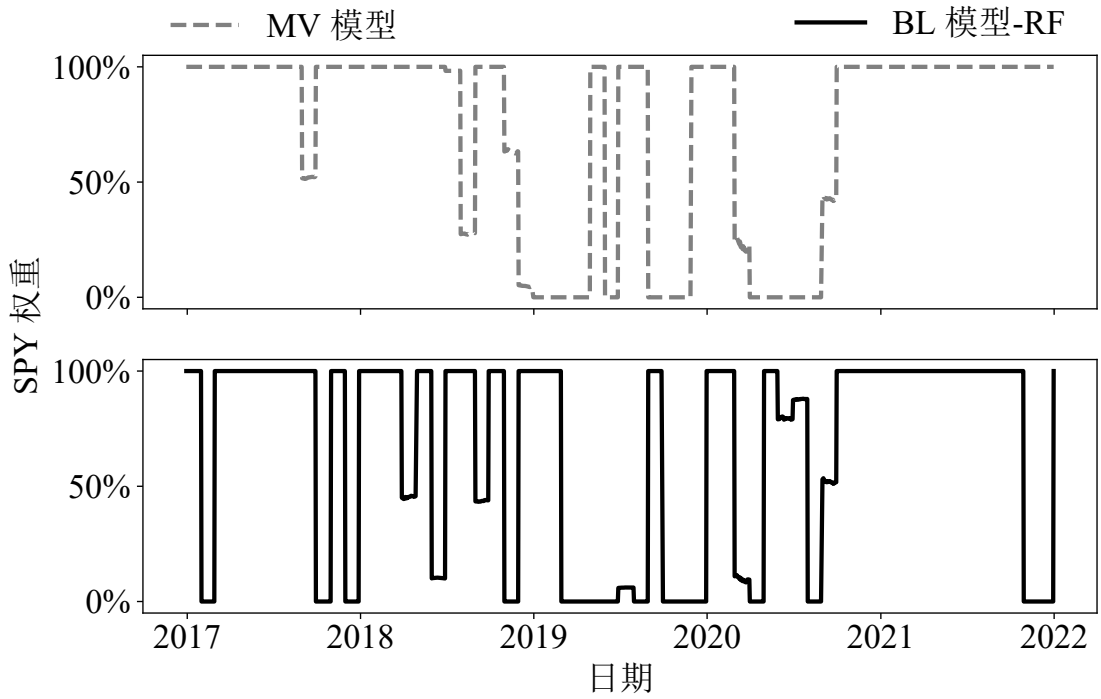

图 4-2: **MV 模型和 BL 模型-RF 的 SPY 的权重变化**

同时我们也可以注意到，MV 模型和 BL 模型-RF 在大部分时期都是全部投资于某一个资产。集中投资于某一个资产在大量资金的投资实践中是一个弊端，它本质上是由公式2-1中求解最优化权重过程的固有性质造成的，因此在两个策略中都无法避免。若投资者希望避免过度集中投资的情况发生，可以在目标函数中加入对权重集中的惩罚项（$\gamma > 0$），如公式4-3所示。

$$\max w'\mu - \gamma w'w \tag{4-3}$$

4.6 稳健性分析

4.6.1 投资期限

上文应用 BL 模型-RF 构造投资组合并持有 5 年，本节对投资期限进行延伸，考察过去 10 年内 BL 模型-RF 和 MV 模型的回测表现。表4-6展示了从 2012 年 1 月 1 日至 2021 年 12 月 31 日各投资组合的绩效指标。从数值上可以看出，在较长的投资期限下，BL 模型-RF 的绩效表现仍比 MV 模型更好，其中前者的确定性等价收益比后者高出约 2%。

表 4-6: 10 年投资期限下各投资组合的绩效指标

	Ret	Vol	CER	MD	Sharpe	Calmar	Turnover
MV	10.69%	10.62%	9.00%	8.43%	1.08	2.39	4.61
BL-RF	12.76%	10.69%	11.04%	8.00%	1.35	2.79	10.35

4.6.2 基预测器的个数

基预测器的个数是随机森林回归中的主要参数之一。基预测器的个数会影响模型的偏差和方差：过少的基预测器组成的随机森林回归模型较简单，可能难以取得较好的预测准确度；过多的基预测器组成的随机森林回归模型又可能导致过拟合的问题，使得预测结果较不稳定。本文4.5部分将基预测器的个数设为50，本节为了考察随机森林回归对基预测器个数的敏感性，对 50 至 500 个基预测器的情形均进行检验，统计预测准确率及其绩效结果如图4-3所示。

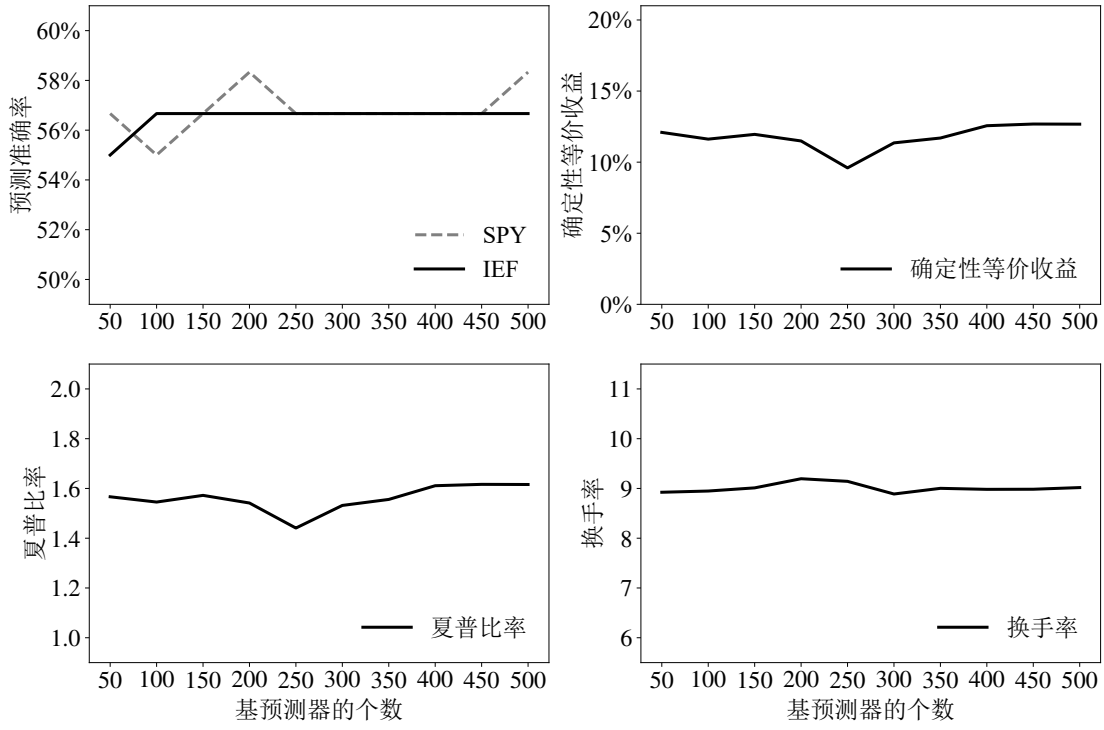

图 4-3: 不同基预测器个数下的预测准确率与绩效指标

从图中可以看出，基预测器的个数在 50 至 500 的范围内，预测准确率和绩效指标都比较稳定，可见随机森林回归对基预测器个数的敏感性并不高。当基预测器个数为 250 时，BL 模型-RF 的绩效表现相对较差，但仍比 MV 模型的绩效表现更好。

4.6.3 交易成本

在实际投资过程中，交易成本是投资者需要考虑的一个重要指标。交易成本不仅包括执行交易时所需要付出的手续费用这类显性交易成本，也包括冲击成本等隐形交易成本。由于后者较难以准确衡量，本节以不同程度的显性交易成本代表整体的交易成本。考虑到美国 ETF 的交易费率约为 0.1%，本节设置了 0% 至 0.3% 的 7 种双边交易费率并分别进行回测。图4-4展示了 MV 模型和 BL 模型-RF 在不同交易成本下确定性等价收益的变化情况。

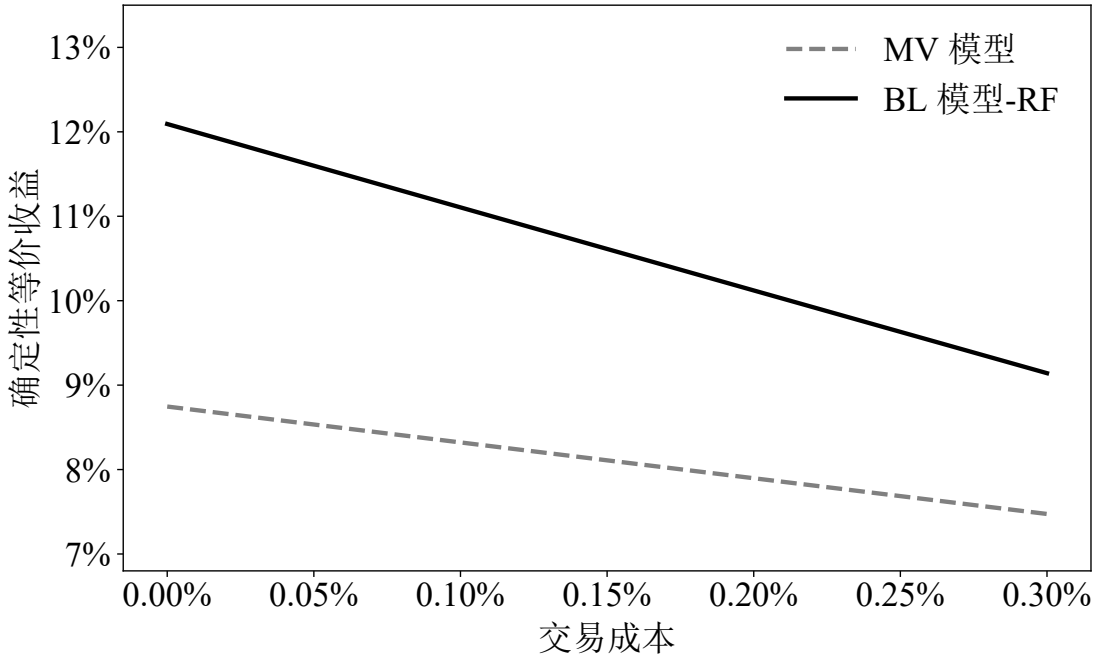

图 4-4: 不同交易成本下的确定性等价收益

从确定性等价收益随着交易成本的增加而下降的幅度来看，BL 模型-RF 对交易成本更加敏感，这本质上是由于其调仓过程中换手率较高导致的。从 MV 模型和 BL 模型-RF 在确定性等价收益表现的比较上来看，BL 模型-RF 的确定性等价收益在交易成本小于 0.6% 时始终比 MV 模型高，这也说明 BL 模型-RF 在投资实践中比 MV 模型更具有优势。

第五章　结论

本文基于随机森林回归算法，以若干技术指标为特征，对美国两大权益指数和债券指数在未来一个月的收益率进行预测，将预测结果作为收益率观点输入到 BL 模型中并构造投资组合。实证检验表明：随机森林回归对收益率的预测准确率高于 50%，且基于随机森林回归的 BL 模型能够取得比均值方差模型更优的绩效表现，前者的平均确定性等价收益比后者高出约 3%，且在较长投资期限下也存在优势。针对随机森林回归中的基预测器个数进行的稳健性分析表明，BL 模型-RF 对基预测器个数这一参数的敏感性并不高，在多组参数下均能取得较稳定的预测效果。加入交易成本后，BL 模型-RF 的平均确定性等价收益相比 MV 模型下降得更快，但在市场合理交易成本下仍然优于传统的 MV 模型。

由于数据可获得性和研究时间有限等局限性，本文目前仍存在一些可以改进之处：第一，机器学习算法输入的特征数据均为技术指标，可以通过考虑纳入基本面数据和宏观经济数据等，进一步考察机器学习算法预测资产收益率的效果；第二，本文仅考虑了随机森林回归这一种机器学习算法，可以使用更多的机器学习算法对收益率预测效果和资产配置策略的绩效加以对比分析；最后，可以对预测收益率和资产配置流程中的更多参数做稳健性分析，更全面地分析和评价模型。

参考文献

[1] 方匡南, 朱建平, 谢邦昌. 基于随机森林方法的基金收益率方向预测与交易策略研究 [J]. 经济经纬, 2010 (2):61-65.

[2] 李斌, 林彦, 唐闻轩. ML-TEA: 一套基于机器学习和技术分析的量化投资算法 [J]. 系统工程理论与实践, 2017, 37(5):1089-1100.

[3] 林耀虎, 刘善存, 杨海军. 一种基于机器学习和蜡烛图的股市投资策略研究 [J]. 计量经济学报, 2022, 2(1):126.

[4] 赵琪, 徐维军, 季昱丞, 等. 机器学习在金融资产价格预测和配置中的应用研究述评 [J]. 管理学报, 2020, 17(11):1716-1728.

[5] Barak S, Modarres M. Developing an approach to evaluate stocks by forecasting effective features with data mining methods [J]. Expert Systems with Applications, 2015, 42(3):1325-1339.

[6] Black F, Litterman R. Asset Allocation [J]. The Journal of Fixed Income, 1991, 1(2):7-18.

[7] Black F, Litterman R. Global Portfolio Optimization [J]. Financial Analysts Journal, 1992, 48(5):28-43.

[8] Breiman L. Random forests [J]. Machine learning, 2001, 45(1):5-32.

[9] Broadie M. Computing efficient frontiers using estimated parameters [J]. Annals of Operations Research, 1993, 45(1):21-58.

[10] Caliskan T. Comparing Black Litterman Model and Markowitz Mean Variance Model with Beta Factor, Unsystematic Risk and Total Risk [J]. Business and Economics Research Journal, 2012, 3(4):1-43.

[11] Donthireddy Pavan. Black-Litterman Portfolios with Machine Learning derived Views [J/OL]. 10.13140/RG.2.2.26727.96160, 2018.

[12] Gu S, Kelly B, Xiu D. Empirical Asset Pricing via Machine Learning [J]. The Review of Financial Studies, 2020, 33(5):2223-2273.

[13] Huang C, Tsai C. A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting [J]. Expert Systems with Applications, 2009, 36(2):1529-1539.

[14] Idzorek T. A Step-By-Step Guide to the Black-Litterman Model Incorporating User-specified Confidence Levels [J]. SSRN Electronic Journal, 2019.

[15] Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction [J]. Expert Systems with Applications, 2021, 178:115019.

[16] Joro T, Na P. Portfolio performance evaluation in a mean–variance–skewness framework [J].

European Journal of Operational Research, 2006, 175(1):446-461.

[17] Khaidem L, Saha S, Dey S R. Predicting the direction of stock market prices using random forest [J]. arXiv preprint arXiv:1605.00003, 2016.

[18] Kumbure M, Lohrmann C, Luukka P et al. Machine learning techniques and data for stock market forecasting: A literature review [J]. Expert Systems with Applications, 2022, 197:116659.

[19] Ledoit O, Wolf M. Honey, I Shrunk the Sample Covariance Matrix [J]. The Journal of Portfolio Management, 2004, 30(4):110-119.

[20] Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection [J]. Journal of Empirical Finance, 2003, 10(5):603-621.

[21] Lee W. Theory and methodology of tactical asset allocation [M]. New Hope, Pa: Fabozzi, 2000.

[22] Liew J, Mayster B. Forecasting ETFs with Machine Learning Algorithms [J]. The Journal of Alternative Investments, 2017, 20(3):58-78.

[23] Lim A, Shanthikumar J, Vahn G. Conditional value-at-risk in portfolio optimization: Coherent but fragile [J]. Operations Research Letters, 2011, 39(3):163-171.

[24] Liu H, Long Z. An improved deep learning model for predicting stock market price time series [J]. Digital Signal Processing, 2020, 102:102741.

[25] Lu W, Li J, Li Y et al. A CNN-LSTM-Based Model to Forecast Stock Prices [J]. Complexity, 2020, 2020:1-10.

[26] Markowitz H. Portfolio Selection [J]. The Journal of Finance, 1952, 7(1):77-91.

[27] Michaud R. The Markowitz Optimization Enigma: Is ‘Optimized’ Optimal? [J]. Financial Analysts Journal, 1989, 45(1):31-42.

[28] Polamuri S R, Srinivas K, Mohan A K. Stock Market Prices Prediction using Random Forest and Extra Tree Regression [J]. International Journal of Recent Technology and Engineering, 2019, 8(3):1224-1228.

[29] Rom B, Ferguson K. Post-Modern Portfolio Theory Comes of Age [J]. The Journal of Investing, 1994, 3(3):11-17.

[30] Satchell S, Scowcroft A. A demystification of the Black–Litterman model: Managing quantitative and traditional portfolio construction [J]. Journal of Asset Management, 2000, 1(2):138-150.

[31] Yuan X, Yuan J, Jiang T, et al. Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market [J]. IEEE Access, 2020, 8:22672-22685.

附 录

各技术指标的构建公式

On Balance Volume（OBV），能量潮指标是根据交易量对价格趋势进行判断的技术指标，具体的计算公式为：

$$\mathrm{OBV}_t = \mathrm{OBV}_{t-1} + \begin{cases} \mathrm{volume}, & \text{若 } \mathrm{close}_t > \mathrm{close}_{t-1} \\ 0, & \text{若 } \mathrm{close}_t = \mathrm{close}_{t-1} \\ -\mathrm{volume}, & \text{若 } \mathrm{close}_t < \mathrm{close}_{t-1} \end{cases}$$

Chande Momentum Osciliator（CMO），钱德动量摆动基于根据上涨和下跌日期的数量的技术指标，具体的计算公式为：

$$\mathrm{CMO} = \frac{\mathrm{S_u} - \mathrm{S_d}}{\mathrm{S_u} + \mathrm{S_d}} \times 100$$

其中，$\mathrm{S_u}$ 是最近个 14 交易日中上涨日的收盘价与前一日收盘价差值的和，$\mathrm{S_d}$ 是最近 14 个交易日中下跌日的收盘价与前一日收盘价差值的绝对值之和。

KDJ 的计算公式为：

$$\begin{aligned} \mathrm{K} &= (q-1)/q \times \mathrm{K}_{t-1} + 1/q \times \mathrm{RSV}; \\ \mathrm{D} &= (p-1)/p \times \mathrm{D}_{t-1} + 1/p \times \mathrm{K}; \\ \mathrm{J} &= 3\,\mathrm{K} - 2\,\mathrm{D}; \end{aligned}$$

Moving Average Convergence Divergence（MACD），指数平滑移动平均线的计算公式为：

$$\begin{aligned} \mathrm{EMA}(n) &= \mathrm{EMA}_{t-1}(n) * (n-1)/(n+1) + \mathrm{close} * 2/(n+1) \\ \mathrm{DIF} &= \mathrm{EMA}(q) - \mathrm{EMA}(p) \\ \mathrm{DEA} &= \mathrm{DEA}_{t-1} * (t-1)/(t+1) + \mathrm{DIF} * 2/(t+1) \end{aligned}$$

Simple Moving Average（SMA），简单移动平均线的计算公式为：

$$\mathrm{SMA}(t) = \frac{1}{n}\sum_{i=0}^{n-1} \mathrm{close}_{t-i}$$