Operations Research and Fuzziology 运筹与模糊学, 2024, 14(2), 599-609
Published Online April 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.142163

Hans汉斯

基于机器学习的股票收益预测与投资组合研究

陈　欣

贵州大学经济学院，贵州　贵阳

收稿日期：2024年2月17日；录用日期：2024年3月8日；发布日期：2024年4月19日

摘　要

计算机和互联网的高速发展使得量化投资在全球逐渐兴起。笔者将机器学习模型和多因子模型相结合构建量化选股模型，并使用上证50指数成分股2016年到2022年的日频数据进行模型训练和样本外预测，结果发现：1) 以随机森林、支持向量机、XGBoost三个模型进行选股构建的投资策略能够战胜市场；2) 投资收益受市场行情影响巨大，在下跌行情中，主动型投资策略即使能够战胜市场，也不能保证获得超过无风险收益率的收益。

关键词

机器学习模型，量化投资，多因子模型

Machine Learning Based Stock Return Prediction and Portfolio Research

Xin Chen

School of Economics GuiZhou University

Received: Feb. 17th, 2024; accepted: Mar. 8th, 2024; published: Apr. 19th, 2024

Abstract

The rapid development of computers and the Internet has led to the gradual rise of quantitative investment worldwide. This author combines machine learning models and multi-factor models to construct a quantitative stock selection model, and uses the daily frequency data of the constituents of the SSE 50 index from 2016 to 2022 for model training and out-of-sample prediction, and finds that 1) The investment strategy constructed by stock selection with the three models of Random Forests, Support Vector Machines, and XGBoost is able to the market; 2) The investment return is affected by the market sentiment greatly, and it is difficult to get more than the risk-free rate of return in the falling market.

文章引用: 陈欣. 基于机器学习的股票收益预测与投资组合研究[J]. 运筹与模糊学, 2024, 14(2): 599-609.
DOI: 10.12677/orf.2024.142163

Keywords

Machine Learning Models, Quantitative Investing, Multi-Factor Models

1. 引言

股票市场是进行已发行股票的买卖、转让和流通的场所。有一句俗语称股市为经济的晴雨表，然而，自股票问世以来，其复杂的变化一直难以准确预测。目前我国金融体系仍处于快速发展阶段，我国的证券交易所成立时间尚短，相较于国外发达国家仍有差距。传统的分析方法在我国这样一个走势复杂，比较容易受到诸多因素干扰的证券市场时常无法发挥出应有的效果。随着电子信息技术的迅速发展，量化投资逐渐崛起。量化投资通过使用计算机强大的算力，能够对市场上海量的信息进行准确筛选，其可以有效防止情绪化行为的不利影响。随着人工智能时代的兴起，机器学习在多个领域取得广泛应用，而在量化投资领域，机器学习正逐渐受到机构投资者的青睐。股票数据中包含着丰富的线性和非线性复杂信息，而传统的时间序列方法通常基于线性或非线性的特定假设。尽管这些方法在一定程度上能够对股票走势进行预测，但其准确度并不十分理想。相较之下，机器学习算法通常是非线性且非参数化的，能够自主学习底层数据中的高级特征，更好地拟合股票数据之间的复杂关系。此外，多因子选股模型和机器学习算法多数源自西方国家，在国内市场的应用相对较少，这有助于有效避免量化策略同质化问题，为投资者提供更为稳定的超额收益。当然，这也要求投资者随着市场变化不断更新已有的量化模型。因此，进行基于机器学习方法和多因子模型相结合的选股研究显得非常必要。在过去，学者们通常采用线性模型研究传统因子，而本文在经过多类因子的筛选后，采用 CART 决策树、随机森林、XGBoost 和支持向量回归等四种不同的机器学习算法构建回归选股模型。通过尝试发现有效因子与股票未来收益之间的关系，本文致力于拓宽多因子模型的理论框架，推动量化投资理论的不断更新和发展。同时，通过实证分析，本文研究了选取的有效因子与多种机器学习回归算法在过去几年上证 50 指数成分股中的可行性，并详细介绍了量化投资过程中多种机器学习算法的建模步骤，有助于投资者更灵活地应对市场的变化和风险。本文对于提高我国证券市场的投资研究水平，推动我国证券市场的良性发展，具有重要的实际意义。

2. 文献综述

2.1. 多因子模型相关研究

在众多量化投资策略中，利用多因子模型对证券市场进行预测是其中的主流方法之一。1952 年，美国经济学家马科维茨(Markowitz) [1]提出了投资组合理论，该理论通过组合收益率的方差来衡量风险，解决了在获取收益与控制风险之间如何平衡投资组合的问题，标志着现代投资组合理论的开端。在这一经典理论的基础上，1964 年，夏普(Sharpe) [2]提出了资本资产定价模型(CAPM)，认为资产的期望收益等于无风险利率与风险报酬之和。风险报酬又可分解为市场组合的风险报酬和特定证券的风险贝塔系数，指出为获取更高回报，投资者必须承担更多波动和风险。资本资产定价模型将理论研究从定性转向定量，并大大简化了模型的复杂性，对整个金融理论和实际操作都产生了巨大影响。1976 年，罗斯(Ross) [3]提出了套利定价模型(APT)，认为除了市场系统风险和个股非系统性风险因素影响股价外，市场还存在其他

多种影响股票价格的因子。同时，他们使用数学表达式说明了股票收益率和因子敞口之间的关系，为多因子模型理论奠定了基础。1993 年，法马和弗伦奇(Fama 和 French) [4]共同提出了三因子模型，解决了 APT 模型未能回答的问题，即哪些因子能解释回报率。通过对美国证券市场中不同股票收益率差异的研究，他们发现仅依靠贝塔系数带来的波动无法解释这些差异。而上市公司的市值、市盈率以及账面市值比可以充分解释收益率的差异。三因子模型为市场中各类股票的收益提供了合理的解释。然而，随着市场的不断变化，学者们发现有许多情况仅凭三因子模型难以解释，因此尝试将更多因子引入其中，不断丰富和发展多因子模型的理论。2015 年，法马和弗伦奇[5]对模型进行了更新，提出了五因子模型，增加了公司的投资和盈利水平以及两个因子。

在国内市场，已有多位学者证实了多因子模型的有效性。王伟(2007) [6]以三因子模型为基础，对上海证券交易所的 A 股数据进行了研究。通过对 2001 年至 2006 年五年内所有 A 股的月度数据进行分析，他证实了三因子模型在中国证券市场的有效性。研究结果显示，在中国证券市场，规模因子、市场因子和账面价值比这三个因子对股票表现有显著影响。另一方面，王涛(2012) [7]基于 2004 年至 2011 年的沪深两市剔除 ST 股后的所有 A 股数据，对中国市场上三因子模型的适用性进行了实证研究。他提出了账面市值因子对大、小账面市值比组合显著，对中账面市值比组合则不显著的观点。同时，规模因子和市场因子对我国股票收益率的影响较为显著。文章还将中国证券市场分为牛市、震荡市和熊市三类走势，并对它们进行了比较，指出规模因子在震荡市中可能失效。何路(2020) [8]选取了 2009 年至 2019 年沪深 300 指数成分股的月度数据，通过 IC 法和单调性检验筛选有效因子。他基于等权重打分法构建了多因子选股模型，其收益表现显著超过了指数，但风险指标略次于指数。随后，他利用消费者信心指数等指标构建了投资者情绪因子，并将其作为择时策略加入模型。结果显示，这显著优化了策略的超额收益、最大回撤和夏普比率等指标，从而证实了投资者情绪因子对市场的影响。

2.2. 机器学习相关研究

在国外研究中，Cao 等人(2001) [9]使用神经网络和支持向量机两种方法，选择五组股票和国债指数数据进行预测。他们发现支持向量机在多个指标上的预测结果优于 BP 神经网络，并强调参数选择对预测结果的重要影响。Kim (2003) [10]则采用神经网络、支持向量机和案例推理三种方法，通过技术分析指标对韩国股指趋势进行预测，指出支持向量机在股指价格方面的预测能力较为出色。Kwon (2005) [11]等人运用遗传算法对股票历史数据进行分析，得到各公司财务指标之间的相关性，通过在输入变量中添加高相关性的公司股票走势，提高了 RNN 模型对目标股票走势预测的效果，并成功建立了超越原有买入并长期持有策略的交易策略。

尽管机器学习算法主要起源于西方国家，国内也已有许多使用机器学习预测股票的研究。徐国祥和杨振建(2011) [12]以沪深 300 指数的交易数据为基础，通过主成分方法降维，再利用遗传算法对 SVM 模型进行参数调优，成功对指数短期走势进行了滑动预测，准确率在不同长度的滑动预测中多数达到 80%。韩燕龙(2015) [13]以上证 50 指数为研究对象，运用随机森林算法进行指数化投资中成分股的筛选，通过解释随机森林算法在选股问题中的适用性，以及与其他方法的对比研究，证明了随机森林在整个指数化投资过程中的有效性。李想(2017) [14]广泛研究了股票的财务、红利、规模等因子以及宏观、债券、楼市等因子的近 300 个数据，通过 XGBoost 算法建模，构建的选股模型在走势上大幅超越了同期沪深 300 指数，取得了超额收益。他还对 XGBoost、随机森林和支持向量机三种算法在分类准确率等方面进行了对比，认为 XGBoost 的预测效果明显优于其他模型。贺隆超(2020) [15]以沪深 300 成分股为研究对象，通过逻辑回归、决策树、随机森林、支持向量机和 K 近邻五种机器学习方法进行对比，效果最好的决策树模型在 2020 年回测中取得了两倍于同期指数的收益。吕子夷(2020) [16]以沪深 300 股指期货一分钟高频

数据为样本，采用支持向量机、BP 神经网络等五种机器学习方法构建回归预测模型，通过建立买涨卖跌的交易策略，支持向量机模型在 2019 年回测中取得了年化 40%的收益率，验证了机器学习模型在预测股指期货价格上的可行性。

3. 模型介绍

3.1. 随机森林

随机森林(Random Forest)是一种集成学习(Ensemble Learning)方法，它通过构建多个决策树来完成任务。随机森林在每个决策树的训练过程中引入了随机性，以提高模型的性能和泛化能力。随机森林具有以下优点：

1. 高准确性

随机森林通过构建多个决策树，并结合它们的预测结果，显著提升了模型的准确性。每个决策树的训练过程采用自助采样，确保每棵树所学到的模式略有不同，从而提高了整体模型的泛化能力。通过多数投票机制，随机森林有效地综合了各个子树的预测结果，进一步增强了整体模型的准确性。

2. 强鲁棒性

随机森林在面对数据中的噪声和异常值时表现出色，得益于其对决策树进行集成的特性。由于每个决策树都在随机子集上进行训练，模型对于个别树的错误预测具有一定的容忍性。这种鲁棒性使得随机森林在处理真实世界中复杂、嘈杂的数据时表现突出，减少了过拟合的风险。

3. 适用性广泛

随机森林不仅可以用于分类问题，还可应用于回归任务，展现了其在不同领域的广泛适应性。无论是医学、金融还是其他领域，随机森林都表现出色，成为解决多种机器学习问题的强大工具。其对大规模数据集和高维特征空间的良好适应性，使其在实际应用中备受青睐。

3.2. XGBoost

XGBoost 是一种梯度提升框架，被广泛应用于机器学习和数据挖掘任务。它在梯度提升算法的基础上进行了优化和改进，具有卓越的性能和高效的计算速度。XGBoost 算法通过运用泰勒展开式计算模型损失的残差的近似值，结合以树的复杂度构建的正则化项来拟合数据残差。这一方法允许控制模型的复杂度，减少模型方差。在优化过程中，XGBoost 实施预剪枝，从而提高预测结果的准确性。此外，XGBoost 采用了预排序、加权分位数、缓存识别等技术，显著提升了算法的运算速度。

3.3. 支持向量回归

支持向量回归(Support Vector Regression)是一种基于支持向量机(Support Vector Machine, SVM)的回归算法。与传统的线性回归方法不同，SVR 的目标是建立一个能够在预测中保持一定误差容忍度的模型。支持向量回归(SVR)的主要优点包括：

1. 非线性关系拟合：通过使用核技巧，SVR 能够有效地拟合非线性关系。
2. 避免过拟合：通过间隔最大化和正则化，SVR 有效地避免了过拟合问题。
3. 稀疏性：由于只有少数样本被选为支持向量，模型具有稀疏性，减少了存储和计算成本。

4. 因子选取

4.1. 数据来源

本研究选取上证 50 指数及其成分股作为研究对象，主要因其市值总额占比较高，并且这些成分股均属

于上海证券市场中流动性较佳、规模较大的股票，具有经营情况稳健、抗风险能力强以及投资价值较高等特点。具体而言，本研究以 2022 年 12 月上证 50 指数成分股为样本，获取了该指数及其成分股在 2016 年至 2022 年间的日度数据，包括开盘价、收盘价、最高价、最低价等，以进行多因子模型的构建与分析。

4.2. 数据预处理

通常情况下，原始数据往往因为量纲不同、数据缺失以及数据异常等问题不能直接使用，因此数据预处理在建模过程中至关重要，它通过一系列操作解决原始数据存在的问题，从而提高数据质量和可用性。有效的数据预处理可以为后续因子筛选和模型建立奠定基础，有助于投资者挖掘有价值的投资标的。

1. 缺失值处理

缺失值的处理方法主要有不处理、数据填充以及直接删除三种。不处理只适用于数据量较小、缺失值较少的情况，否则会影响建模的效果；直接删除适用于数据量较大而缺失值较少的情况，在数据量少时不适用，并且只要删除数据都会导致数据中原有的信息被删除；数据填充则是使用其它数据补全原始数据中的缺失值，常用的填充方法有很多，包括向前后填充、移动平均法、非参数方法、中位数填充、矩阵分解、插值填充、人工填写等等。本文根据实际情况采用插值填充法解决原始数据缺失值问题。

2. 异常值处理

异常值是指偏离数据整体趋势的极端值，异常值的存在会对数据分析、建模和预测等产生不利影响，因为它们会显著地扰乱数据的分布，使得模型偏离真实情况，因此在使用之前必须对数据中的异常值进行处理。判断异常值的方法包括 3σ 原则、观察数据的分布图、使用箱线图检测异常值等。本文根据 3σ 原则删除了在 3σ 之外的数据。

3. 标准化

由于数据的量级不同，若不对数据进行标准化，将对构建的模型造成重大的负面影响。而标准化数据可以降低模型对噪声的敏感性、加速模型的训练速度、提高模型的预测精度。数据标准化中最常用的方式有 Box-Cox 标准化、Min-Max 标准化和 Z-score 标准化三种，他们各有其优劣势，本文根据数据特征选择使用 Min-Max 方法。

4.3. 因子介绍

本文选取的因子包括技术面因子和基本面因子共 19 个；其中技术面因子分别是开盘价、收盘价、最高价、最低价、成交量、成交额、日振幅以及日换手率；基本面因子是估值类因子市盈率、市现率、市销率、市净率，盈利类因子净资产收益率、总资产收益率、销售净利率，成长类因子营业收入同比增长率、净利润同比增长率、经营性净现金流同比增长率、、净资产收益率同比增长率等。

4.4. 因子有效性检验

在量化投资领域，常用信息系数(IC)和信息比率(IR)来评估因子对未来收益率的影响。当 IC 值和 IR 值的绝对值较大时，表明该因子与未来收益率存在显著关联，暗示其在选股方面具备较为优越的能力。本研究主要运用因子 IC 值和 IR 值来评估因子在 2016 年至 2021 年期间的有效性和稳定性，通过进行单因子检验对股票数据进行了全面的考察，其结果见表 1。

鉴于上述结果中 IC 值小于 0.01 的因子效果较为不显著，本文将予以剔除，并以剩余的 12 个因子构建新的因子组合。

5. 实证研究

本文以精选的因子作为输入特征，以股票收益率作为输出特征。基于随机森林、XGBoost 和支持向

量回归等机器学习算法，通过在训练集上进行滑动交叉验证和调参，比较各模型的效果。随后，进行量化选股，最终挑选出月收益率前 10 名的股票。在此基础上，进行等权重组合构建，计算投资组合的收益指标，为投资者提供参考。

Table 1. Results of factor validity test
表 1. 因子有效性检验结果

因子	IC	IR
开盘价	0.002648	0.022098
收盘价	0.018409	0.153612
最高价	0.011822	0.098645
最低价	0.010185	0.084989
交易量	0.043889	0.366220
交易额	0.063450	0.529445
日振幅	0.094075	0.784987
日换手率	0.059237	0.494286
市盈率	0.020596	0.171860
市现率	0.004131	0.034471
市销率	0.005799	0.048389
市净率	0.006887	0.057466
净资产收益率	0.007198	0.060063
总资产收益率	0.016948	0.141420
销售净利率	0.008936	0.074562
营业收入同比增长率	0.017223	0.143717
净利润同比增长率	0.023563	0.196614
经营性活动现金流量净额增长率	0.006109	0.050973
净资产收益率同比增长率	0.020795	0.173519

5.1. 样本内训练

在使用机器学习模型之前，通常需要对模型进行交叉验证和参数调优。交叉验证是一种模型评估的技术，它通过将数据集划分为训练集和测试集的多个子集，并多次训练模型，以更准确地评估模型的性能。交叉验证有助于避免模型在特定训练集或测试集上的性能波动，并提高对模型泛化性能的信心。常见的交叉验证方法包括 k 折交叉验证、留一法交叉验证、固定窗口滑动交叉验证等。股票数据属于时间序列数据，其关联性具有一定的时效性。股票走势受到较近期数据的较大影响，而过去较远的数据影响较小。为了更贴近实际情况，本文采用了滑动预测方法。在滑动预测的过程中，通过设置滑动窗口对数据集进行划分，使得训练数据不断向前滑动。与传统的留出法和交叉验证法相比，滑动预测方法更接近实际情况，具有更强的实用意义。另外，机器学习模型通常有多个超参数，超参数将影响模型的性能，而且是在模型训练之前由人为设定的，无法通过训练数据学习得到，因此为了优化模型的性能，我们需要先调参。常见的调参方法有强化学习调参、网格搜索调参、粒子群优化算法、随机搜索调参等，本文将采用网格搜索法进行调参。具体而言，我们首先把数据分为 2016~2021 年和 2022 年两部分，前者为训

练集，后者为测试集；然后将训练集进行固定窗口滑动交叉验证，即每次将前 4 年的数据作为训练集，下一个月的数据作为验证集，在每一个窗口下进行网格搜索调参，找到每个模型在此窗口下的最优参数组合，并以最优参数组合下的模型在验证集上进行性能评估，输出评估结果，最后根据每个窗口下得到的最优参数组合得到全局的最优参数组合。本文通过固定窗口滑动交叉验证法进行调参，随机森林模型参数对比如表 2 所示：

Table 2. Comparison of parameters of random forest model before and after optimization
表 2. 随机森林模型参数调优前后对比

参数名	默认值	优化后
n_estimators	100	50
max_depth	None	10
learning rate	2	10

5.2. 样本外测试

使用训练所得的随机森林、XGBoost 和支持向量回归模型，在 2022 年的数据集上进行滑动回归预测，并以均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)作为模型评价指标，模型在 2022 年 1 月调参前后的回归表现如表 3 所示：

Table 3. Evaluation indicators
表 3. 评价指标

评价指标	参数优化情况	随机森林	支持向量机	XGBoost
MSE	优化前	0.013224	0.013213	0.010886
	优化后	0.013207	0.013046	0.010175
RMSE	优化前	0.114997	0.114948	0.104338
	优化后	0.114920	0.114219	0.103189
MAE	优化前	0.086189	0.085961	0.077201
	优化后	0.086054	0.085143	0.072910

可以观察到，XGBoost 模型在回归任务中表现最佳，其次是支持向量机模型，而随机森林回归模型的效果相对较差。具体来看，在调参之前 XGBoost 的 MSE 比随机森林和支持向量机分别的 MSE 分别减小 17.68%和 17.61%；XGBoost 的 RMSE 比随机森林和支持向量机分别的 RMSE 分别减小 9.27%和 9.23%；XGBoost 的 MAE 比随机森林和支持向量机分别的 MAE 分别减小 10.43%和 10.19%。在经过调参之后，XGBoost、随机森林以及支持向量机三个模型的三个预测指标都略有改善，但提升效果并不显著，同时 XGBoost 的各项指标仍旧优于随机森林和支持向量机；从 XGBoost 模型上看，优化后三个指标的性能提升分别为 6.53%、1.1%和 5.56%。这可能是因为默认参数已经对数据有了较好的拟合效果，导致后续调参对性能提升的贡献比较有限。这也提醒我们，仅仅通过参数调整未必能够显著提高模型性能，数据预处理阶段的质量同样至关重要。

5.3. 模型选股

为了客观地比较三种机器学习预测模型的预测表现，本文采用滑动窗口的方法，基于调参后的模型，

在 2022 年的数据上进行滑动预测，逐月分别输出收益率前十的股票。三个模型 2022 年 1 月份选股结果如表 4 所示：

Table 4. Stock selection results
表 4. 选股结果

模型	股票代码	实际收益率
随机森林	601669	0.304218
	600048	0.045727
	601919	0.032725
	603986	0.018376
	601668	0.014240
	600436	0.013734
	600111	0.010293
	601857	0.008639
	603260	0.006786
	603259	0.006542
支持向量机	601669	0.268342
	600048	0.102032
	600406	0.075619
	601633	0.071614
	601919	0.058356
	600346	0.048626
	600111	0.033893
	600031	0.019637
	600690	0.016237
	603259	0.008031
XGBoost	601669	0.311754
	600031	0.087143
	601919	0.071634
	603501	0.053806
	600346	0.027613
	601995	0.026301
	600436	0.024859
	601012	0.016778
	603259	0.011126
	601668	0.007074

从表 4 的结果可以看出，入选的股票 2022 年 1 月均获得了正向收益率，而上证 50 在 2022 年 1 月的

收益率为−6.7%，说明三个模型选股都跑赢了上证 50 指数。

5.4. 等权重组合

在金融投资领域，等权重投资组合是一种广泛应用的策略，通常被用作基准策略，以便比较其他投资组合的表现。本文利用不同模型逐月进行选股，并构建等权重投资组合，以对比该投资组合与相关指数的走势。在评价投资策略时，运用多个评价指标有助于对投资策略进行全面和客观的评估。在本文中，我们选择了年化收益率和夏普比率这两个指标，用以评估投资组合的效果。年化收益率是用来表示一个投资组合或资产在一年内的平均收益的指标。这是一种用百分比表示的数值，通常用于衡量投资的盈利能力和效果。年化收益率可以帮助投资者更好地理解他们的投资在长期内的表现。而夏普比率是用于衡量投资组合或资产相对于无风险利率的超额收益与风险之比的指标，其作为一种常用的衡量风险调整后收益率方法，用于评估投资组合的性能。三个模型按照以上选股策略构建的等权重投资组合的评价指标如表 5 所示：

Table 5. Portfolio performance
表 5. 投资组合表现

模型	年化收益率	夏普比率
随机森林	0.64%	−0.36
支持向量机	1.14%	−0.20
XGBoost	0.84%	−0.31

从上表中不难发现，首先三个模型构建的投资组合年化收益率都为正，而 2022 年上证 50 指数的年收益率为−19.5%，说明三种投资组合策略均战胜了市场，取得了超额收益；其次，从夏普比率来看，几个投资策略的夏普比率均为负值，说明个组合的绝对收益均没有超过无风险收益率。所以，在市场行情下跌的时候，即使战胜了市场也并不代表就能获益。

6. 结论与展望

6.1. 结论

本文以 2016 年 1 月年到 2022 年 12 月上证 50 指数及其成分股日交易数据为基础构建机器学习量化选股模型，首先对所选因子库里的因子进行了有效性检验，挑选出效果较好的因子作为输入变量，其次选择随机森林、支持向量机和 XGBoost 三个机器学习方法，使用上证 50 指数成分股 2016~2022 年间数据，将筛选后的因子作为输入变量来对收益率进行预测，并使用固定窗口滑动交叉验证和网格搜索调参方法调整参数，最后以 RMSE，MSE，MAE 评价模型性能改进情况，结果显示调参有一定的效果。随后，我们以参数调优之后的模型对 2022 年逐月进行选股预测，选取当月收益率前十的股票构建等权重投资组合，并每月调整组合，结果发现：1) 三个组合年化收益率都高于上证 50 指数当年的收益率，战胜了市场；2) 三个组合的绝对收益率均低于无风险收益，表明投资收益受市场行情影响巨大，想要在下跌行情中取得收益非常困难。

6.2. 不足与展望

本文虽然从多个维度对因子选取和机器学习模型选股进行了研究和对比，得出了一些结论，但是由于作者在文献查阅和科研实践方面的不足，文章在因子筛选、机器学习模型构建和实证上仍然存在一些

不足，未来有待改进。

1. 没有考虑宏观和情绪等因子。宏观经济和投资者情绪，作为极其重要的影响因素，对股票收益的演变具有不可忽视的影响。当前模型的研究虽然涵盖了基本面和技术面因子，但却未考虑到宏观层面的经济状况以及投资者的情绪因素，这可能导致对股票市场行为的理解不够全面。在宏观经济方面，考虑到国家和全球范围内的宏观经济指标，如 GDP、通货膨胀率和利率水平，对于准确预测股票收益至关重要。这些宏观经济因素能够揭示经济体整体的健康状况，对企业盈利和市场流动性等方面产生深远影响，从而直接影响股票价格的波动。投资者情绪方面的考量同样重要，因为市场参与者的情感状态直接关系到他们的交易决策。投资者的信心、恐慌和情绪传导在市场中产生连锁反应，从而影响股票价格的波动。考虑这些情绪因素有助于更好地理解市场行为背后的驱动力，提高对未来股票走势的准确性。

2. 在股票选取的范围方面，目前的研究局限于选择上证 50 指数的成分股作为主要研究对象，着眼于大盘股的特性。因此，在提高研究的广度和代表性方面存在一定的不足。扩展研究对象至 A 股市场更为广泛的股票类型，可能会为股票选择的研究提供更具代表性和普适性的结论。首先，考虑到 A 股市场的多样性，包括小盘股、中盘股以及创业板股票等不同类别，这些股票所处的市值、行业分布和成长阶段差异显著。通过将更多类型的股票纳入研究，可以更全面地了解不同市值和行业的股票在市场中的表现，进而为投资者提供更为丰富的投资选择。其次，大盘股和小盘股在市场行为和价格波动方面往往呈现出不同的特征。通过扩大研究范围，研究者可以更好地理解不同市值股票之间的相关性和相互作用，为投资组合的构建提供更具洞察力的信息。这有助于投资者更有效地管理风险和实现长期的资本增值。

3. 在模型选择方面，当前的研究未充分应用近年来深度学习领域日益广泛应用的先进模型。深度学习模型，如长短时记忆网络(LSTM)和递归神经网络(RNN)，在处理时序性数据和复杂非线性关系方面具有显著的优势。对于股票市场这样的高度动态和非线性的金融领域，引入这些深度学习模型可能为模型的预测性能和泛化能力带来显著的提升。具体而言，LSTM 作为一种能够捕捉时间序列数据长期依赖关系的深度学习模型，适用于分析股票价格的历史走势。其能够有效地学习并记忆时间序列数据中的复杂模式，对于捕捉股票市场中的短期和长期趋势提供了更为强大的建模能力。类似地，RNN 模型也能够处理时序数据，对于捕捉时间序列中的动态关系具有独特的优势。

参考文献

[1] Markowitz, H. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

[2] Sharpe, W.F. (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, **19**, 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

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

[4] 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

[5] Fama, E.F. and French, K.R. (2015) A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, **116**, 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

[6] 王伟. 三因素模型在中国资本市场的有效性研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2008.

[7] 王涛. Fama-French 三因子模型及其添加市盈率因子模型在中国股市的适用性研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2012.

[8] 何路. 多因子量化选股及投资者情绪择时策略的实证检验[D]: [硕士学位论文]. 南京: 南京大学, 2020.

[9] Tay, F.E.H. and Cao, L. (2001) Application of Support Vector Machines in Financial Time Series Forecasting. *Omega*, **29**, 309-317. https://doi.org/10.1016/S0305-0483(01)00026-3

[10] Kim, K.J. (2003) Financial Time Series Forecasting Using Support Vector Mechines. *Neurocomputing*, **55**, 307-319. https://doi.org/10.1016/S0925-2312(03)00372-2

[11] Kwon, Y.K., Choi, S.S. and Moon, B.R. (2005) Stock Prediction Based on Financial Correlation. *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, Washington DC USA, 25-29 June 2005, 2061-2066. https://doi.org/10.1145/1068009.1068351

[12] 徐国祥, 杨振建. PCA-GA-SVM 模型的构建及应用研究——沪深 300 指数预测精度实证分析[J]. 数量经济技术经济研究, 2011, 28(2): 135-147.

[13] 韩燕龙. 基于随机森林的指数化投资组合构建研究[D]: [硕士学位论文]. 广州: 华南理工大学, 2015.

[14] 李想. 基于 XGBoost 算法的多因子量化选股方案策划[D]: [硕士学位论文]. 上海: 上海师范大学, 2017.

[15] 贺隆超. 多因子量化选股与机器学习量化择时投资策略研究[D]: [硕士学位论文]. 乌鲁木齐: 新疆财经大学, 2020.

[16] 吕子夷. 基于机器学习算法的股指期货价格预测与比较研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2020.