E-Commerce Letters 电子商务评论, 2025, 14(5), 3195-3203
Published Online May 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.1451633

Hans汉斯

基于文本分析——LSTM的量化选股模型

路　优

贵州大学经济学院，贵州　贵阳

收稿日期：2025年4月11日；录用日期：2025年4月26日；发布日期：2025年5月31日

摘　要

股票市场的波动深刻影响投资者决策，凸显量化选股的关键意义。本研究创造性地将文本分析和长短期记忆神经网络(LSTM)相结合，构建了选股模型。模型先借助TF-IDF和TextRank剖析相关财经新闻及行业简报，挖掘热门利好行业及股票，再运用LSTM模型预测所选股票最低价走势，为投资者决策提供参考依据。经过模型评估，其均方误差(MSE)为0.02037，平均绝对误差(MAE)为0.09182，决定系数(R^2)达到0.9754，这些指标显示模型的预测性能较为出色。通过文本分析确定科技行业为热门利好行业，并选取该行业的九洲集团、焦点科技两支股票展开案例分析，验证了引入滑动窗口的LSTM模型在预测精度方面优于未引入滑动窗口的LSTM模型。此研究成果为量化选股开辟新方向，有望助力投资者决策更科学。

关键词

量化选股模型，词频逆向文件频率(TF-IDF)，TextRank算法，长短期记忆神经网络(LSTM)

Quantitative Stock Selection Model Based on Text Analysis-LSTM

You Lu

School of Economics, Guizhou University, Guiyang Guizhou

Received: Apr. 11th, 2025; accepted: Apr. 26th, 2025; published: May 31st, 2025

Abstract

The fluctuations in the stock market have a profound impact on investors' decisions, highlighting the crucial significance of quantitative stock selection. This study innovatively combines text analysis with the Long Short-Term Memory neural network (LSTM) to construct a stock selection model. The model first utilizes TF-IDF and TextRank to analyze relevant financial news and industry briefs, unearthing popular and favorable industries and stocks. Then, it employs the LSTM model to predict the lowest price trends of the selected stocks, providing a reference basis for investors' decisions. Through model evaluation, its Mean Squared Error (MSE) reaches 0.02037, the Mean Absolute Error

文章引用: 路优. 基于文本分析——LSTM 的量化选股模型[J]. 电子商务评论, 2025, 14(5): 3195-3203.
DOI: 10.12677/ecl.2025.1451633

(MAE) is 0.09182, and the Coefficient of Determination (R^2) is 0.9754, indicating that the model has a good prediction effect. Through text analysis, it is determined that the technology industry is a popular and favorable industry, and two stocks, Jiuzhou Group and Focus Technology, in this industry are selected for case analysis, confirming that the LSTM model with a sliding window is superior to the LSTM model without a sliding window in prediction accuracy. The results of this study open up a new direction for quantitative stock selection and are expected to help investors make more scientific and rational decisions and make their investment behaviors more forward-looking and accurate.

Keywords

Quantitative Stock Selection Model, Term Frequency - Inverse Document Frequency (TF-IDF), TextRank Algorithm, Long Short-Term Memory Neural Network (LSTM)

1. 引言

随着全球金融市场的持续发展，量化投资在金融领域的重要性愈发凸显。近年来，人工智能技术的快速进步推动了机器学习在量化投资中的广泛应用，为投资者带来了更高效、更精准的投资决策支持工具。在中国，金融市场规模持续扩大，截至 2023 年底，A 股市场总市值已超过 80 万亿美元，投资者对于科学有效的选股策略需求愈发迫切。与此同时，全球范围内，各国金融市场也在不断探索创新，寻求提高市场效率和投资收益的方法，量化选股策略的研究与应用成为共同关注的焦点。

在这样的现实背景下，如何提高选股策略的准确性和稳定性成为了投资者和研究者面临的重要问题。选股策略的质量直接关系到投资组合的收益水平，对于投资者的财富增长和金融市场的稳定发展具有至关重要的意义。然而，当前的选股策略面临着诸多挑战，如市场的复杂性、数据的高维度和非线性等，传统的选股方法难以充分挖掘数据中的有效信息，导致选股效果不尽如人意。因此，迫切需要探索新的方法和技术来提升选股策略的性能。

量化选股的研究历史悠久，众多学者从不同视角进行了深入研究。早期研究多聚焦于传统统计分析方法和基本面分析。随着技术发展，机器学习算法逐渐被引入该领域。其中，神经网络算法凭借其强大的非线性映射能力备受关注。长短期记忆网络(LSTM)作为特殊神经网络，在处理时间序列数据时表现出独特优势，能有效捕捉数据中的长期依赖关系。但目前将文本分析与 LSTM 相结合应用于量化选股策略的研究相对较少，多数研究仅侧重于单一数据类型或传统模型，缺乏对多源数据融合及先进神经网络模型综合应用的深入探讨。

后文将按照以下结构展开论述。第二章进行文献回顾，第三章详细阐述文本分析与 LSTM 的相关理论基础，包括文本分析的方法、LSTM 的工作原理。第四章介绍数据的来源与处理过程，以及构建基于文本分析——LSTM 的量化选股模型，包括模型的架构设计、参数选择、训练过程。第五章对研究成果进行了总结，剖析了研究存在的局限性，并对未来的研究方向进行了展望。

2. 文献综述

在量化选股领域的早期，机器学习方法开始逐渐被引入，但应用相对较为简单。研究主要集中在利

用传统机器学习算法，如支持向量机(SVM)、决策树等，对股票市场数据进行分析和预测。邓晶[1]改进TWSVM 算法用于量化投资，评估核函数性能，将股票预测模型与多模型对比，引入函数提出 L1FTWSVM 模型并验证，探究其量化投资适用性。TWSVM 模型在高维含噪股票数据上预测性能好，L1FTWSVM 模型分类正确率高，对应量化投资策略收益率高且风险指标良好。梁馨月[2]构建多因子选股模型，选因子并检验剔除，用多种回归方法建模回测，选宏观经济指标构建并分析其与股市关系，依此制定择时策略并分析效果。逐步回归法因子构建策略优，宏观经济因子与股价正相关，择时策略可增收益、胜率，降回撤。李馨蕊[3]研究 RF-SVM 算法多因子选股，介绍相关原理算法与选股流程，以沪深 300 成分股为样本，筛选因子、优化参数、回测对比，变换条件考察策略优势。RF 能筛因子，SVM 适合股价预测，RF-SVM 策略短期收益高但随调仓间隔增加而递减，降维效果和回测收益优于对比策略，泛化能力良好。

随着大数据技术的蓬勃发展以及计算能力的持续增强，机器学习算法在量化选股领域的应用呈现出快速推进的态势。研究者们开始探索更为复杂的机器学习模型，例如神经网络(涵盖多层感知机、卷积神经网络等)以及集成学习方法(如随机森林、梯度提升树等)。杨淼杰，王彩凤，武辰华[4]构建了基于随机森林算法的多因子选股模型。从 6 个维度选取 12 个因子构建因子池，预处理因子数据，确定随机森林函数参数。用处理后数据预测单个股票，决定买卖并计算收益率，构建多因子选股模型，以沪深 300 成分股为股票池，选 12 个因子和股票上涨情况数据，设交易成本和滑点，回测模型，根据随机森林算法特征重要性和相关性分析改进因子和模型。随机森林算法预测有一定效果，多因子选股模型年化收益率 25.5%。张儒森，黄外斌[5]研究基于集成学习算法的指数增强策略。以沪深 300 指数为研究对象，选取其成分股相关数据，包括收盘价、成交量等，计算技术指标和基本面指标，预处理后构建数据集。运用随机森林(RF)、梯度提升决策树(GBDT)和极端梯度提升(XGBoost)这三种集成学习算法搭建指数增强模型，以均方误差(MSE)作为损失函数来训练模型，并借助网格搜索与交叉验证对参数进行优化，对比不同模型的性能，挑选出最优模型构建投资组合，进而与沪深 300 指数展开对比分析。XGBoost 模型在预测准确性和稳定性上表现最优，基于 XGBoost 模型构建的投资组合年化收益率达 12.5%，夏普比率为 0.65，优于沪深 300 指数。该指数增强策略能有效跟踪指数走势并获取超额收益，为投资者提供了投资参考，但在市场波动较大时超额收益可能受影响。赵俊茹[6]以沪深 300 成分股作为样本，选取 2016 年至 2023 年的数据，挑选出 36 个因子构建多因子交易策略的备选因子池。经过数据预处理以及有效性检验(包括 IC 检验、MIC 检验)，剔除那些无效和冗余的因子，最终保留了 15 个因子构建了图神经网络模型。

近年来，深度学习技术的不断进步为量化选股领域注入了新的活力。像长短期记忆网络(LSTM)、门控循环单元(GRU)以及 Transformer 等深度学习模型，在处理时间序列数据以及复杂的非线性关系时展现出了出色的能力。何开元[7]构建选股与择时量化交易策略。选股策略上，用 Stacking 集成学习模型预测沪深 300 成分股日未来收益率，评估多种单一机器学习模型后构建多种 Stacking 集成模型，根据预测结果选股回测。择时策略上，提出 Bayes-CNN-LSTM-Attention 模型预测沪深 300 指数日未来趋势，引入注意力机制改进特征提取和时序预测，用贝叶斯算法优化 LSTM 参数，进行消融和对比实验回测。Stacking 集成模型比单一模型预测精度和策略收益率显著提升，Stacking 模型二表现最佳，策略累计收益率 17.78%，Bayes-CNN-LSTM-Attention 模型预测精度优于基准模型，准确率、精确率和召回率分别达 70.25%、72.97%和 65.85%，加入择时策略后选股策略收益率增加、最大回撤减少，证明择时策略有效。徐舫[8]提出两阶段投资组合选择方法。第一阶段，用 CEEMDAN 分解沪深 300 成分股收益率时间序列，以排列熵重构，用 LSTM 和 GRU 分别预测高频和低频分量，引入注意力机制构建组合预测模型，以遗传算法优化超参数，对比其他模型。第二阶段，用第一阶段预测结果替换 Markowitz 模型历史收益率构建投资组合，设计两种调仓频率方案并回测。组合预测模型在多种损失函数上对沪深 300 成分股收益率预测精度更高。基于该模型和 Markowitz 模型构建的投资组合，策略收益和年化收益更高，一天调仓频率方案表现更好，

证实投资组合选择方法可提升预测精确度和投资组合整体收益。

机器学习在量化选股领域的应用经历了从传统机器学习算法到深度学习技术的发展过程。传统机器学习模型在早期为量化选股提供了有效的工具，但随着数据量的增加和计算能力的提升，模型的复杂度和预测精度相对有限，复杂模型和深度学习模型逐渐展现出更强的性能和优势。深度学习模型有自动学习数据深层次特征和模式的能力，以及在处理时间序列数据和复杂非线性关系方面的卓越表现，更好地适应了股票市场的动态变化，显著提升了量化选股策略的准确性和稳定性，成为当前机器学习量化选股研究的热点方向。

3. 相关理论和模型介绍

3.1. 文本分析

本文使用 TF-IDF 和 TextRank 对财经新闻进行文本分析。TF-IDF 值是词频和逆文档频率的乘积，即 $\text{TF-IDF}(t,D)=\text{TF}(t,D)\times\text{IDF}(t)$，这个值综合考虑了一个词在特定文档中的出现频率和它在整个语料库中的普遍程度，从而能够更准确地衡量一个词对于某一文档的重要性。词频(TF)指的是某个词在一篇文档里出现的次数占该文档总词数的比例。它能在一定程度上体现一个词在特定文档中的关键程度，通常情况下，词在文档中出现得越频繁，可能就越关键。逆文档频率(IDF)则是衡量一个词在众多文档中普遍重要性的一种指标，要是某个词在大量文档中都有出现，它的独特性就比较差，重要性也就相对低一些；反之，若一个词仅在少数文档中出现，它就具有较高的独特性，可能更具重要性。TextRank 是一种借鉴了 PageRank (网页排序算法)思想的基于图的文本排序算法。它把文本中的句子或单词当作图中的节点。如果两个句子(或单词)之间有某种关联(例如，在一定的窗口范围内共同出现)，则在它们之间建立一条边。边的权重表示两个节点之间关联的强度。每个节点(句子或单词)的重要性得分是通过迭代计算得到的。初始时，每个节点都被赋予一个相同的初始得分(例如，1/n，其中 n 是节点的总数)。在每次迭代中，一个节点的得分是根据与其相连的其他节点的得分以及边的权重来更新的。具体公式为：

$$S(V_i)=(1-d)+d\times\sum_{j\ni(\ln(v_i))}\frac{w_{ji}}{\sum_{k\in\text{Out}(v_j)}w_k}S(V_j)$$

其中，$S(V_i)$ 是 V_i 节点的分数，d 是阻尼系数，$\ln(V_i)$ 是指向节点 V_i 的节点集合，w_{ji} 是从 V_j 节点到 V_i 节点的边权重，$\text{Out}(v_j)$ 是从节点 V_j 出发的边连接的节点集合。这个迭代过程会一直进行，直到节点的得分收敛(即前后两次迭代的得分变化小于某个阈值)。

3.2. LSTM

LSTM (长短期记忆网络)是循环神经网络(RNN)的一种改进架构，主要用于处理序列数据，例如时间序列、文本序列等。LSTM 单元由输入门(Input Gate)、遗忘门(Forget Gate)、输出门(Output Gate)以及记忆单元(Cell)构成。

输入门决定了哪些新的信息可以进入记忆单元。它通过一个 sigmoid 函数和一个点乘操作来控制输入的信息流。例如，输入门的值为 0 时，新的信息完全无法进入记忆单元；值为 1 时，新的信息可以完全进入。

遗忘门用于决定从记忆单元中丢弃哪些信息。它同样通过 sigmoid 函数产生一个介于 0 和 1 之间的值，来控制上一时刻记忆单元中的信息保留多少。例如，如果遗忘门的值为 0，就会把上一时刻记忆单元中的所有信息都丢弃；如果为 1，则全部保留。

输出门控制记忆单元中的哪些信息可以作为当前时刻的输出。通过 sigmoid 函数和 tanh 函数的组合，

输出门决定了输出的信息流。

记忆单元(Cell)是 LSTM 的核心部分，它存储了长期的状态信息。新的输入信息经过处理后与记忆单元中的旧信息进行组合，从而更新记忆单元的状态。

4. 模型构建

4.1. 数据来源及处理

本文的数据获取自 tushare 平台，具体的数据采集过程如下：首先，从该平台分别抓取同花顺、新浪财经以及东方财富的新闻各 1000 条；随后，进一步筛选出利好行业中的利好股票，并提取其 2021 年 12 月 17 日至 2024 年 12 月 17 日的日度数据，这些数据涵盖了开盘价、收盘价、最低价和最高价等关键特征。在进行数据分析前，本文对基础数据实施了一系列适当的预处理操作，旨在提升数据质量并优化其可用性。第一步，针对数据中存在的缺失值，本文采用插入中值的方法予以补齐，确保数据的连续性和完整性。第二步，鉴于不同价格特征的数据量级存在差异，为了使各类价格数据能够处于同一量级范围，从而更有利于后续 LSTM 模型高效地学习数据内在特征，本文使用 Min-Max 归一化方法处理数据，将其映射到[0, 1]区间。(Min-Max 归一化，即假设最低价序列为 $P_{\text{low}} = [p_1, p_2, \cdots, p_n]$，归一化后的序列计算公式为 $P'_{\text{low}} = \frac{P_{low} - \min(P_{\text{low}})}{\max(P_{\text{low}}) - \min(P_{\text{low}})}$)。这样可以使不同价格数据在同一量级上，有助于 LSTM 更好地学习数据的特征。通过这一标准化处理，有效消除了不同价格特征因量级差异可能带来的不利影响，为后续的深入分析和模型构建奠定了坚实基础。

4.2. 文本分析-LSTM 量化选股模型

Figure 1. TextRank word cloud (based on aggregated financial news from three platforms)
图 1. TextRank 词云图(综合三个平台财经新闻结果)

Figure 2. TF-IDF word cloud (aggregated from three financial news platforms)
图 2. TF-IDF 词云图(综合三个平台财经新闻结果)

首先，本文运用 TF-IDF (词频－逆文档频率)和 TextRank 这两种先进的文本分析技术，对从各大平台抓取而来的财经新闻进行深度剖析，旨在精准筛选出其中具有关键影响力的词汇。在此过程中，本文

不仅精确识别出这些关键词，还依据科学的算法计算出每个关键词的重要性得分，进而根据这些关键词及其对应的得分成功绘制出直观且富有信息量的词云图，如图 1 和图 2 所示。通过对综合同花顺新闻、新浪财经新闻以及东方财富新闻这三类财经新闻所生成的词云图进行观察与解读，可以清晰地发现，在当前的财经领域中，科技行业成为了备受瞩目的焦点，诸多与科技相关的词汇在词云图中占据显著位置，据此，获取科技行业简报进行文本分析，如图 3 和图 4，有比较大的“科技”、“涨停”字眼，有力地表明科技行业是近期热门的利好行业，其在市场动态和经济发展趋势中展现出强劲的发展势头和巨大的潜力。

Figure 3. TextRank word cloud (tech industry brief)
图 3. TextRank 词云图(科技行业简报)

Figure 4. TF-IDF word cloud (tech industry brief)
图 4. TF-IDF 词云图(科技行业简报)

其次，抓取科技行业中的部分股票数据，用 python 进行建模。其步骤如下；

步骤 1：抓取数据及数据预处理。使用 tushare 平台的 Python 接口，调用相应的数据获取函数抓取九洲集团，焦点科技两支股票 2021 年 12 月 17 日至 2024 年 12 月 17 日的日度数据，采用中值插补的方法补齐缺失值，使用 MinMaxScaler 类，将数据缩放到 0 到 1 的范围。最后，将处理好的数据存储到 Excel 表格中。

步骤 2：建立不带滑动窗口的 LSTM 模型。首先将数据集按 4:1 划分为训练集和测试集，并分别存储在 train 和 test 变量中。接着进行模型构建，使用 numpy 的 reshape 函数对训练集和测试集数据重塑为(样本数，1，特征数)的形状。采用 Sequential 模型构建 LSTM 架构：先添加 units 为 50 的 LSTM 层，其 input_shape 为(1, X_train.shape [2])；再添加单元数量为 1 的 Dense 层作为输出层。模型构建完成后进行编译，使用 adam 优化器和 mean_squared_error 损失函数。最后使用 fit 函数训练模型，输入为重塑后的 X_train 和 Y_train，epochs 设为 100，batch_size 设为 1，verbose 设为 2 以输出每轮训练详细信息，便于监控和调整。

步骤 3：建立带滑动窗口的 LSTM 模型，设置滑动窗口长度为 5，用前 5 天的数据预测第六天的数据。首先定义 sliding_window 函数，该函数用于创建滑动窗口数据。通过循环遍历数据集，将数据按照

滑动窗口的方式添加到 X 和 Y 列表中，设置滑动窗口和步长参数定义 window_size 为 5，表示滑动窗口大小为 5。定义 stride 为 1，表示每次滑动的步长为 1。其次，划分训练集和测试集，训练集占 80%，测试集占 20%。将 X 和 Y 数据划分为训练集 X_train、Y_train 和测试集 X_test、Y_test。使用 numpy 的 reshape 函数对训练集和测试集数据进行重塑。重塑后的形状为(样本数，时间步长，特征数)，其中时间步长由滑动窗口大小决定。然后再构建 LSTM 模型。

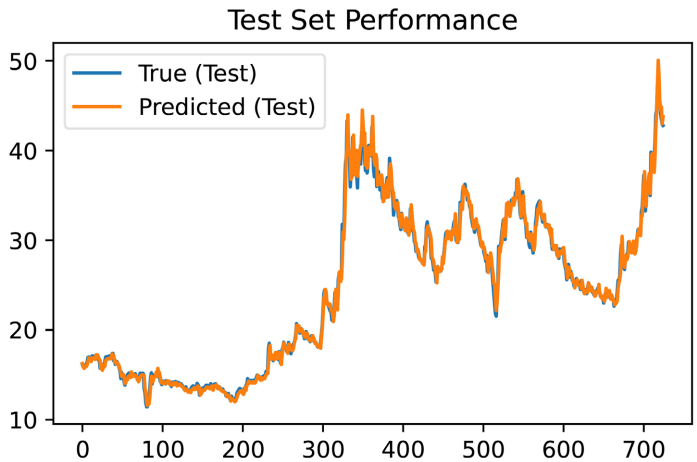
步骤 4：模型评估。均方误差(MSE)是预测值和真实值差值的平方和的均值，它反映了模型预测值和真实值之间平均的平方误差。MSE 值越小，表明模型的预测效果越出色。平均绝对误差(MAE)是预测值与真实值差值的绝对值的均值，它衡量了模型预测值与真实值之间的平均绝对误差。MAE 值越小，说明模型预测的准确性越高。决定系数(R^2)用于表示模型对数据的拟合优度，其取值在 0 到 1 之间。R^2 越接近 1，意味着模型对数据的拟合效果越好；R^2 越接近 0，表示模型对数据的拟合效果越差。如表 1，带滑动窗口的 LSTM 模型均方误差(MSE)是 0.02037，平均绝对误差(MAE)是 0.09182，决定系数(R^2)是 0.9754，说明该模型预测效果良好。且带滑动窗口的 LSTM 模型对比不带滑动窗口的 LSTM 模型决定系数(R^2)较大，均方误差(MSE)和平均绝对误差(MAE)较小，说明带滑动窗口的 LSTM 模型对比不带滑动窗口的 LSTM 模型拟合效果好。

Table 1. Model evaluation results
表 1. 模型评估结果

模型	均方误差(MSE)	平均绝对误差(MAE)	决定系数(R^2)
不带滑动窗口的 LSTM 模型	1.5994	0.4174	0.9648
带滑动窗口的 LSTM 模型	0.02037	0.09182	0.9754

步骤 5：可视化。如图所示，这是采用带滑动窗口的 LSTM 模型，依据九洲集团和焦点科技的日度数据所训练得出的预测值与实际值随时间变化的趋势图。从图中可以清晰地看到，两张图中的两条曲线大致重合，这一现象再次有力地证明了该模型具有良好的拟合效果(图 5，图 6)。

进一步观察九洲集团和焦点科技的曲线走势，能够发现九洲集团在第 75 时间点时达到了最低点，焦点科技在 460 时间点达到了最低。从投资的角度来看，这可能是一个买入的信号。



横轴代表时间(单位：天，从 2021 年 12 月 17 日计数为 0，依次递推)。
纵轴代表股票的最低价(单位：元)。

Figure 5. Trend of predicted and actual values over time for Jiuzhou Group
图 5. 九洲集团预测值和实际值随时间变化趋势图

横轴代表时间(单位：天，从 2021 年 12 月 17 日计数为 0，依次递推)。
纵轴代表股票的最低价(单位：元)。

Figure 6. Trend of predicted and actual values over time for Focus Technology
图 6. 焦点科技预测值和实际值随时间变化趋势图

5. 结论

本文创新性地将文本分析与 LSTM 相结合，构建了量化选股模型。通过 tushare 平台采集多平台新闻以及九洲集团、焦点科技的股票数据，对数据进行筛选、缺失值处理和归一化等预处理操作后，运用 TF-IDF 和 TextRank 技术深入分析新闻，成功绘制词云图，从而精准识别出科技行业为当前热门利好行业。在模型构建方面，分别建立了带滑动窗口和不带滑动窗口的 LSTM 模型，并对其进行全面评估。结果显示，带滑动窗口的 LSTM 模型在预测九洲集团和焦点科技股价走势时表现出色，其均方误差(MSE)为 0.02037、平均绝对误差(MAE)为 0.09182、决定系数(R^2)达到 0.9754，显著优于不带滑动窗口的模型，且预测值与实际值曲线高度吻合，在股价波动关键拐点的判断上具有较高准确性。

然而，本研究也存在一定局限性。一方面，数据采集范围相对狭窄，仅从同花顺、新浪财经、东方财富抓取各 1000 条新闻以及两支股票的数据，有限的数据量难以全面反映复杂多变的市场全貌，可能削弱模型对多样化市场状态和各类股票特性的精准捕捉能力。另一方面，股票市场受众多复杂因素交互影响，而本模型可能未能充分考量政策突变、重大突发事件等特殊情形对股价走势的潜在冲击，这在极端市场条件下可能导致模型预测能力和适应性的显著下降。

展望未来，研究方向可从以下几方面拓展：在数据层面，积极拓展数据来源的广度和深度，纳入更多元化的财经新闻渠道、社交媒体信息以及行业报告等，丰富数据类型，同时引入宏观经济指标、公司治理数据等更多维度的数据，为模型提供更全面的市场信息，以进一步提升选股的准确性。在模型优化方面，致力于增强模型的适应性和稳定性。通过优化模型结构和算法，使其具备更强的自适应能力，能够灵活应对不同市场环境和各类股票的独特特征。同时，引入更完善的风险控制指标和机制，全面评估模型在不同市场条件下的风险收益表现，确保在极端市场波动时仍能有效预测并精准控制风险。此外，结合强化学习等前沿技术，赋予模型动态优化选股策略的能力，不断提升其在复杂多变市场中的稳定性和盈利能力，为投资者提供更具价值的决策支持。

参考文献

[1] 邓晶. 基于孪生支持向量机的量化投资研究[D]: [硕士学位论文]. 上海: 上海工程技术大学, 2021.

[2] 梁馨月. 基于宏观经济因素下的多因子选股及实证分析[D]: [硕士学位论文]. 北京: 中央民族大学, 2022.

[3] 李馨蕊. 基于 RF-SVM 算法的多因子量化选股研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2022.

[4] 杨淼杰, 王彩凤, 武辰华. 基于随机森林算法的多因子选股模型研究(英文) [J]. 纯粹数学与应用数学, 2023, 39(4): 506-519.

[5] 张儒森, 黄外斌. 基于集成学习算法的指数增强策略研究[J]. 金融客, 2024(3): 27-29.

[6] 赵俊茹. 基于图神经网络的多因子选股策略研究[D]: [硕士学位论文]. 兰州: 兰州财经大学, 2024.

[7] 何开元. 基于机器学习的选股与择时量化交易策略研究[D]: [硕士学位论文]. 武汉: 武汉轻工大学, 2024.

[8] 徐舫. 基于机器学习预测的量化选股研究[D]: [硕士学位论文]. 武汉: 武汉轻工大学, 2024.