分类号_____ 密级 _____

U D C _____ 编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于机器学习的多因子量化投资策略研究

研 究 生 姓 名：欧阳飞

指导教师姓名、职称：陈芳平 教授

学 科、专 业 名 称：应用经济学、金融工程

研 究 方 向：金融投资

提 交 日 期：2022 年 6 月 5 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：欧阳飞　　签字日期：2022.6.1

导师签名：　　签字日期：2022.6.1

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：欧阳飞　　签字日期：2022.6.1

导师签名：　　签字日期：2022.6.1

A study of multi-factor quantitative investment strategies based on machine learning

Candidate :OuYang Fei

Supervisor:

摘 要

本文将建立在多因子理论模型的基础上利用线性回归模型、LSTM 神经网络、LightGBM 模型构建量化投资策略。即对市场众多影响股票收益率的因子分别运用传统的 IC 分析法以及随机森林机器学习算法进行有效因子以及重要因子筛选，将筛选过后的因子作为 LSTM 神经网络模型和 LightGBM 模型两种机器学习算法的输入特征，进行模型训练并进行预测股票涨跌，构建投资组合回测得到收益率曲线，并与多因子线性回归选股模型构建的投资组合收益率曲线进行分析对比。

本文基于 2015-01-01 至 2021-11-30 所有 A 股股票共 64 个影响股票收益率的因子数据和相应的月收益率进行建模分析，构建六组量化投资策略与一组改进后的量化投资策略，并基于相同的环境回测，研究发现：回测期间内（1）IC 分析法和随机森林机器学习算法两种方法都能有效的对初始因子库因子进行筛选，但两种方法筛选后的因子存在较大差异；（2）基于线性回归模型、LightGBM、LSTM 模型构建的七组量化投资策略在回测时期内，均获得了超过基准收益率的收益，并都获得较高夏普比率；（3）在基于线性回归、LSTM 模型构建的量化投资策略，将 IC 分析法作为因子筛选方法，获得更好的回测收益。基于 LightGBM 模型构建的量化投资策略，将随机森林作为因子筛选方法，获得更好的回测收益；（4）将 IC 分析法作为因子筛选方法，仅基于 LSTM 模型构建的量化投资策略回测收益与改进后的基于 LightGBM 构建的量化投资策略回测收益优于基于线性回归模型构建的量化投资策略回测收益；将随机森林方法作为因子筛选方法，基于 LSTM 和 LightGBM 机器学习模型构建的量化投资策略回测收益优于线性模型构建的量化投资策略回测收益。较基于线性回归模型构建的量化投资策略，机器学习模型构建的量化投资策略稳定性、风险控制等指标也都有所提升；（5）在七组量化投资策略模型中，将 IC 分析法作为因子筛选方法，基于 LSTM 模型构建的量化投资策略模型在回测期间获得了最佳收益，同时，在回测期间也拥有最大的收益回撤。

关键词：机器学习；多因子；量化策略

Abstract

This paper will build a quantitative investment strategy using linear regression model, LSTM neural network, and LightGBM model on the basis of multi-factor theory model. That is, the traditional IC analysis method and the stochastic forest machine learning algorithm are used to screen the effective factors and important factors respectively for many factors affecting the stock returns in the market, and the filtered factors are used as the input characteristics of the two machine learning algorithms of the LSTM neural network model and the LightGBM model, the model is trained and the stock rise and fall is predicted, the portfolio back test is constructed to obtain the yield curve, and the portfolio yield curve constructed by the multi-factor linear regression stock selection model is analyzed and compared.

Based on the modeling and analysis of a total of 64 factor data and corresponding monthly returns of all A-share stocks from 2015-01-01 to 2021-11-30, six sets of quantitative investment strategies and a set of improved quantitative investment strategies are constructed, and based on the same environmental back test, the study finds that the two methods of (1) IC analysis method and random forest machine learning algorithm during the back testing period can effectively filter the initial factor library factors, but there are large differences in the factors after the screening of the two methods ;(2) Seven groups of quantitative investment strategies based on linear regression models, LightGBM, and LSTM models have obtained returns that exceed the benchmark rate of return and all obtained higher Sharpe ratios during the back testing period; (3) In quantitative investment strategies based on linear regression and LSTM models, IC analysis methods are used as factor screening methods to obtain better back testing returns. Based on the lightGBM model, the

quantitative investment strategy uses random forest as a factor screening method to obtain better back testing returns; (4) Using ic analysis as a factor screening method, the back testing return of quantitative investment strategy based only on LSTM model and the back testing return of the improved quantitative investment strategy based on LightGBM are better than the back testing return of quantitative investment strategy based on linear regression model; the back testing return of quantitative investment strategy based on LSTM and LightGBM machine learning model is superior to the back testing return of quantitative investment strategy constructed by linear model. Compared with the quantitative investment strategy based on the linear regression model, the quantitative investment strategy stability and risk control indicators constructed by the machine learning model have also been improved; (5) In the seven sets of quantitative investment strategy models, the IC analysis method is used as a factor screening method, and the quantitative investment strategy model based on the LSTM model obtains the best return during the back testing period, and at the same time, it also has the largest draw down during the back testing period.

Keywords：Machine Learning；Multi-factor；Quantify the strategy

目 录

1.绪论

1.1 研究背景

随着计算机算力的大幅增长和海量的数据来源，使得机器学习算法能应用到许多领域，其中在量化投资领域有着广泛深刻的应用。有学者将机器学习定义为是一种高维模型，模型常常用于进行预测的工作，机器学习的本质是预测。而在多因子模型理论中，该理论认为资产的超额收益率是由许多因子暴露所推动，即资产的超额收益率可以得到解释。因此机器学习与经典因子投资理论的结合，成为了量化投资的热门研究领域。

任何投资决策的做出，都是为了获得超过资产平均收益率的部分，也即资产超额收益率，股票市场在漫长的发展中，催生出了两种经典的投资分析方法，从股票背后的公司基本面角度以及股票交易走势技术方面分析，基本面分析即根据当前国内国际宏观形势、行业发展趋势以及公司财务信息等，判断当前公司股价的高低；而技术面则是根据股票交易过程中所形成的各类技术指标，对未来股价涨跌进行预测。无论是公司基本面分析还是股票交易走势技术方面分析，都是基于决策人主观的进行分析决策，对公司股票价格进行判断，进而做出决策。将机器学习算法应用到多因子模型上，其主要思想是在大量的数据分析基础上，通过机器学习算法的运用，构建出模型来描述因子与资产超额收益率之间的关系，进而通过因子数据以及机器学习算法训练出来的模型对资产的超额收益率进行预测，再帮助决策人进行投资决策。该投资分析方法一定程度上杜绝了投资者心态对投资决策的影响，因而量化投资的出现，使得投资人的决策更加理性客观。

机器学习算法训练出来的模型有着较强的分类预测能力，因此在量化研究领域运行机器学习算法，可很大程度上提高投资决策的准确性，将基于多因子模型和机器学习算法量化投资策略作为论文的选题有很强的理论和实践研究价值。

1.2 研究意义

理论意义方面，马科维茨 1952 年在他的论文中提出了资产组合的相关理论，至此开始进入现代金融学阶段，之后 CAPM、APT、多因子模型等理论成果相继

出现，这些理论将资产收益率归因于因子暴露，构建线性关系模型描述资产收益率与因子暴露之间关系，学者们在资产定价实证研究的文章中，大多都延续了线性模型的假设。但股票市场是极其复杂的市场，因此在资产定价实证研究时适当考虑非线性关系，将会是不错的选择，机器学习算法能较为准确的描述非线性关系，因此本文将为资产定价的实证研究提供一些结论参考。

实践意义方面，有效市场理论将股票市场分为三类，分别是强有效市场、半强有效市场以及弱有效市场，而对于市场非有效的解释大多来自于行为金融学，一方面是人们的预期不一样，一方面是对于信息或者事件的过度反映造成了市场的非有效。相对来讲，国外的股票市场较为成熟，有效性比我国股票市场要强，资产的定价常常偏离其真实的市场价值，市场中的投资者的行为常常是不理性的。通过量化投资策略提高广大投资者的纪律性、准确性、系统性以及科学性，帮助投资者构建系统科学的投资策略，对促进我国股票市场的长期稳定发展有着重要的实践意义。

1.3 文献综述

本文的国内外文献综述将集中在两个方面，一方面是国内外展开对因子模型的研究，其中研究的因子模型包括单因子模型和多因子模型；另外一方面是国内外展开对机器学习算法方面的量化投资应用研究，包括单一分类的机器学习算法量化投资应用研究和集成类的机器学习算法量化投资应用研究。

1.3.1 国外相关文献综述

单因子模型研究方面，Markowitz（1952）发表的论述资产组合理论的学术论文，开创了研究现代金融学的先河。随后 Sharp 等人（1964）年将资产组合理论作进一步延伸，用一个一元的线性模型解释资产的收益率，这个一元线性模型也叫做资本资产定价模型（CAPM），该理论将资产的收益率和市场超额收益率建立起一元的线性关系，即市场风险因子的暴露是资产获得收益率的原因。并将此模型用于给资产定价，自此学界开始对资产的预期收益率有所清晰认识，也拉开了多因子定价模型的研究。在之后学者展开的研究中，发现不能由单一的因子解释不同资产的收益率。

多因子模型研究方面，学者们发现不能仅仅使用一元线性的数学模型构建资产收益率和因子之间关系，Ross（1976）提出了的套利定价理论（APT），不再使用 CAPM 理论的相关假设，将一元线性模型扩充成立多元的线性模型，增加了多个因子来共同对资产收益率进行解释，其中增加的因子有 GDP 增长率因子、通货膨胀率因子但其并未给学界带来解释力强的多因子模型。直到 Fama 和 French（1992）提出线性三因子模型，和套利定价模型相同的是，数学模型是多元线性模型，比套利定价理论更为进步的是，将三个解释收益率的因子确定下来，它们是市场因子，企业规模因子以及企业价值因子。将这三个因子和资产收益率建立模型，解释资产收益率。Fama-French 三因子模型也成为多因子模型的开山鼻祖。在二人后续的研究成果中，Fama 和 French（2015）又加入盈利因子、风格因子等进一步将三因子模型拓展到了五因子模型，用于解释资产收益率。随着市场行为被不断的挖掘研究，基于行为金融学的因子以及技术层面的因子也逐渐被加入到多因子模型中。Stambaugh 和 Yuan（2017）在市场因子和企业规模因子的基础上增加了企业管理因子和股价表现因子，增加行为金融学的角度对资产收益率进行解释。Liu et al.（2019）将换手率因子加入到多因子模型中，构建了四因子模型。

单一分类的机器学习算法量化投资应用研究方面，单一分类算法的代表算法有决策树算法、支持向量机、BP 神经网络等。kimk 等人（2003）把支持向量机用来对股价进行预测，而且实证分析认为支持向量机应用在股票价格预测时，能取得很好的预测结果。Huang（2012）将遗传算法与支持向量回归结合构建 GA-SVA 模型，取得了不错的预测效果。但随着数据结构的复杂、数据量大问题的突出。

集成类机器学习算法量化投资应用研究方面，单一的分类算法已经不能很好的满足量化研究需要，学者们开始展开以 LSTM、XGBoost、LightGBM、随机森林等集成类算法为主的量化研究。Maragoudakis 和 Serpanos（2010）运用随机森林数据挖掘技术，解决了股票预测等波动性复杂领域中的高维度问题。Ladyzynski 等（2013）将标普 500 成分股作为研究数据，用随机森林机器学习算法进行股票价格走势的预测，预测准确性很高。Manojlovic 和 Stajdubar（2015）

利用随机森林构建了提前 5 天和 10 天的股票预测模型，平均分类准确率为 76.5% 和 80.8%。Khan 等人（2008）研究发现相对于反向传播神经网络，基于遗传的反向传播神经网络更加适用于进行股票价格的预测。Hadavandi（2010）基于人工神经网络和遗产模糊系统的股价预测专家系统，对股价具有较好的预测能力。Tinor（2013）基于贝叶斯正则化的人工神经网络提高了证券价格的预测能力和泛化能力。Xiong（2016）研究发现长短期记忆网络（LSTM）在预测标普 500 指数的波动率时，具有很强的预测能力。Fscher 和 Krauss（2017）使用 LSTM 神经网络模型对标普 500 指数的成分股进行选股研究分析，分析研究发现基于 LSTM 的选股模型效果优于支持向量机、随机森林和逻辑回归模型。Sun 和 Zhao（2015）将 AdaBoost 模型及其改进模型应用于中国股票市场构建选股模型，研究结果表明发现采用 AdaBoot 模型及其改进模型的选股策略均取得了强于基准的投资收益。Chen 和 Guestrin（2016）年提出 XGBoost 模型，随后微软研究亚洲研究院（2017）年在提出与较 XGBoost 性能有所提升的 LightGBM 模型，但因为两个模型提出日期较近，在多因子量化投资研究领域研究论文还较少，但是在其他领域有许多有效应用。Chen 和 Fu 等（2017）通过 XGBoost 加权分类器，对复杂的雷达信号进行有效的分类。Wang 和 Zhang 等（2017）分别以随机森林、XGBoost、LightGBM 模型对乳腺癌中的 miRNA 目标进行识别和和分类，研究结构现实 LightGBM 模型的分类能力较其他两种模型相比较强。Zhong 等人（2021）分别以线性回归模型、logistic 回归模型、XGBoost 模型，基于多因子对沪深 300 成分股进行选股策略研究，研究发现 logistic 回归模型选股策略具有最高的年化夏普比率。

1.3.2 国内相关文献综述

单因子模型研究方面，刘霖（2001）在进行资本资产定价模型的实证研究中，发现用中国 A 股市场的数据进行研究，不管采用何种研究方法，资本资产定价模型无法得到证明。并且还发现，和资本资产定价模型不同的是，股票收益率和因子之间的关系呈现的也不是线性的关系。而且股票收益率不仅仅由 β 因子决定，还与 β 之外的因子有关。陈浪南和屈文洲（2000）同样也开展了资本资产定价模型的实证研究，选用的也是中国股票的数据，研究发现股票收益率和 β 因子之间

存在不稳定的关系。虽然选取中国股票市场数据进行资本资产定价模型的实证研究，并不能利用中国股票市场得到有效验证，也即呈现出较强的不适用性。

多因子模型研究方面，利用中国股票市场实证 Fama-French 三因子模型的研究中，取得了良好的实证分析效果。余世典（2002）在论文中针对中国股票市场提出了三因子模型，研究发现这个三因子模型基本可以对资产收益进行解释，与 Fama-French 三因子模型不同的是，选取的因子不一样，该文章中选取的三个因子分别是公司规模因子、公司价值因子以及市场组合，杨炘和陈展辉（2003）发现中国 A 股市场有明显的企业规模效应和账面市值比效应，增加市场因子、企业规模因子、账面市值比因子的 Fama-French 三因子模型能有效说明中国股票市场收益率的截面差异。王源昌等人（2011）构建了一个改进的三因子模型，并且深入研究认为改进后的三因子模型样本中的资产收益率的解释力度较强，新的三因子模型中除了市场因子还增加了市盈率因子和换手率因子，王茵田和朱茵资（2011）与其他学者不同的是，构建的多因子模型中因子数量多达八个，并且深入研究发现八因子模型比三因子模型解释力更强，其中因子包括市场风险因子、投资资本比、账面市值比、盈利股票价值比等，欧阳志刚和李飞（2016）在进行多因子模型研究中发现在 Fama-French 三因子模型的基础上加入有 6 个月滞后期的动量因子的四因子模型，对中国股票市场的资产收益率解释，比 Fama-French 三因子模型和 CAPM 模型更有解释力。李志冰，杨光艺等（2017）在对 Fama-French 提出的五因子模型进行实证研究发现，选取的数据是中国股票市场数据，该五因子模型相比于资本资产定价模型、三因子模型等有更好资产收益率解释力度。

单一分类的机器学习算法量化投资应用研究方面，方匡南等人（2010）年将随机森林方法用于预测我国基金超额收益率，并与自回归移动平均方法、随机游走、支持向量机方法等加以对比，随机森林方法有很好的预测效果。曹正凤等人（2014 年）研究发现使用随机森林算法能获得正确率较高的股票分类。王淑燕等人（2016）构建了八因子选股模型，并利用随机森林算法对中国 A 股股票的涨跌情况进行了精准的预测。贾秀娟（2019）构建了基于随机森林的支持向量机选股模型，先对输入因子变量进行降维处理并与主成分降维方法进行比较，结果表明基于随机森林挑选出来的因子作为输入变量，有更好的预测效果。闫政旭等

人（2021）在随机森林方法的基础提出了一种基于 Pearson 系数的随机森林组合模型，最后发现改进后的随机森林算法比传统随机森林算法的股票预测模型效果有所提高。除了随机森林集成算法。

集成类机器学习算法量化投资应用研究方面，邓凤欣和王洪良（2018）利用 LSTM 神经网络对友邦保险、长和、微软以及亚马逊的收盘价进行预测，发现 LSTM 神经网络模型有较高的精度和较为稳定的预测效果。陈佳等人（2018）利用聚类分析、主成分分析法对输入参数进行优化，应用 LSTM 模型预测纳斯达克指数、标普 500 指数，结果和准确度都有显著提升。彭燕等人（2019）利用苹果公司的量价指标作为基础数据，发现引入正则化项和 Dropout 机制的 LSTM 神经网络模型对苹果公司股价预测的准确率提高了 30%。冯宇旭和李裕梅（2019）经过对比 SVR、AdaBoost 模型，LSTM 模型在沪深 300 指数期货变化趋势的预测能力优于其他两个模型。裴大卫和朱明（2019）将多因子模型与 LSTM 神经网络模型结合，将多因子作为输入特征，发现多因子模型的引入，提升了 LSTM 股票价格的准确率和带来更好的模型鲁棒性。欧阳红兵等人（2020）将小波分析和联合 LSTM 神经网络模型结合，对比支持向量机、多层感知机、K 近邻、GARCH 等四种模型，发现 LSTM 神经网络对于金融时间序列数据的预测能力强于其他模型。区别于 LSTM 神经网络集成学习算法，国内应用 XGBoost 和 LightGBM 两种算法的量化研究丰富程度不及 LSTM 神经网络模型，但是 XGBoost 和 LightGBM 在量化投资领域和其他领域得到了有效验证。黄卿和谢合亮（2018）利用沪深 300 股指期货 1 分钟的数据为研究对象，发现对比于神经网路模型、支持向量机，XGBoost 对下 1 分钟的股指期货价格变动方向有较强的预测能力。王燕和郭元凯（2019）利用网格搜索算法对 XGBoost 模型进行参数优化并构建 GS-XGBoost 预测模型，发现 XGBoost、GBDT、SVM 模型，GS-XGBoost 模型在 MSE、RMSE、MAE 评价指标上表现出好的预测效果。葛橹漠和周显（2020）构建了一个机遇 XGBoost 的多因子选股模型，研究发现基于 XGBoost 机器学习模型选出的股票组合相对于等权重的多因子模型有明显的效果提升。比 XGBoost 计算速度提高的 LightGBM 模型在国内量化研究不大丰富，但在其他领域有许多有效应用。牛雪琪（2018）利用 LightGBM 对美国 P2P 平台的违约风险进行预测，

相比于其他预测算法，LightGBM 算法的预测效果更好。李泽远（2021）对比与 logistic 回归、卷积神经网络，LightGBM 模型根据客户信息和贷款信息对贷款违约预测分类的效果最优。

1.3.3 文献评述

对上述文献进行总结发现，自多因子模型被广泛研究以来，市场已经挖掘出许多不同因子可以用于解释不同资产收益率的差异，这些因子涵盖了宏观层面、公司层面以及技术层面等，并被应用于多因子模型的构建。多因子模型也都得到了国内外股票市场数据的实证证明，将这些多因子模型中包含的因子作为输入特征并利用机器学习算法进行股票价格进行预测，有充分的金融学理论支持。

关注机器学习算法量化投资应用层面，不管从单一的分类算法还是集成类的算法，都表现出对股票价格优秀预测能力，但不同的算法之间存在着些许差异，我们同时可以看到的是，大多数基于机器学习算法的股票预测模型，都是将股票交易形成的量价指标以及技术指标作为输入特征变量，而鲜有将多因子模型中包含的因子数据作为输入特征变量进行股票价格预测。在仔细研究相关多因子模型和机器学习算法结合的文献中，基于机器学习算法的多因子选股模型，相比也表现出很好预测效果。从具体的研究内容上讲，将因子数据作为输入特征，由于因子数量众多，需要进行有效的因子筛选，从而提高预测精度以及计算速度，大多数论文对因子筛选采用的是 IC 分析法分析因子有效性，达到筛选因子的目的，除此之外有少量论文采用主成分分析法和随机森林算法对因子数据进行降维处理，随机森林算法不仅能对股票价格进行预测，还能对输出输入特征的重要性排序，从而达到因子筛选的目的。然而鲜有论文将 IC 分析法和随机森林算法一起应用于因子的筛选并进行比较。通过了筛选的因子，作为机器学习算法的输入特征进行股票价格预测，作为集成学习算法的两个典型算法 LSTM 和 LightGBM 在量化研究领域有着强势的预测能力，还未有将多因子模型和这两种算法结合并进行比较分析的尝试，因此将这两种算法和经过 IC 分析法、随机森林算法筛选的因子模型进行结合，是一个比较新的尝试，可以为量化投资研究领域提供一些策略参考。

1.4 研究方法与思路

本文采用的研究方法为文献研究法、比较分析法以及实证分析法。文献研究法，通过阅读大量文献，对国内外的多因子模型以及机器学习量化策略的相关文件进行梳理和总结，对现阶段研究进展进行归纳总结，确定本文的研究思路和研究方向；比较分析法，在实证部分，分别对三种策略模型利用中国的股票进行实证分析，进行回测分析师，对构建策略的收益及风险指标等指标进行分析对比，得出一组最优的策略效果；实证分析法，实证研究侧重于对经济变量之间的内在数量逻辑的探究，旨在揭露内在经济的规律性。本文运用实证分析方法，通过对我国 2015 年 6 月至 2021 年 11 月的股票因子数据和股票收益率数据进行分析和预测，从而寻求各个模型对股票收益率的解释。

研究思路方面，如下图中的流程图来表示：

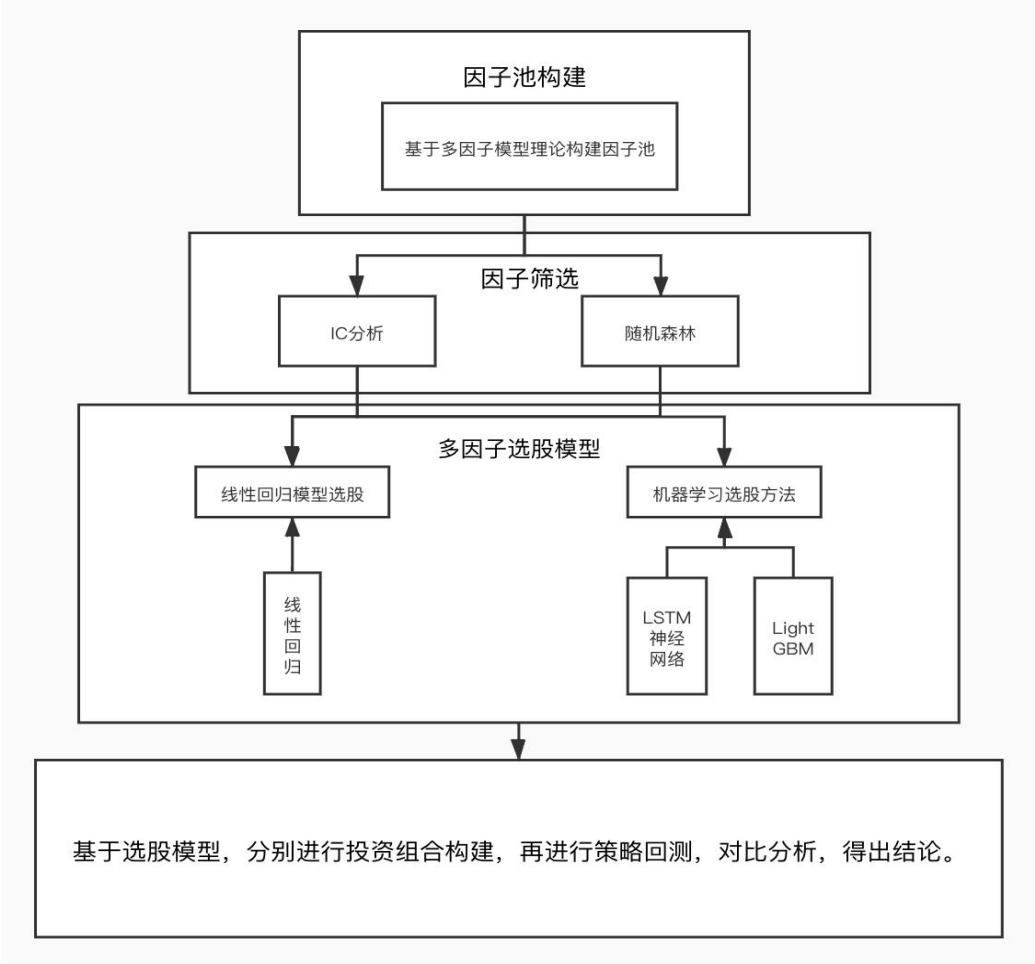


图 1.1 研究思路

1.5 创新与不足

本文的创新点主要体现在两个方面，首先，分别利用 IC 分析法和随机森林

算法对因子进行筛选并比较分析，其次基于多因子模型利用两种典型算法 LSTM 和 LightGBM 构建量化策略，并与传统线性回归模型进行比较分析。

本文在现有的研究基础上进行了一些改进，仍存在一些不足：本文虽然参考众多学者研究影响股票收益率的因素，但由于数据的可获取性，对候选因子库的候选因子指标有所取舍，导致最终的量化投资效果有所局限性，对于中国股票市场的适应性有待提高；构建回测策略进行模型分析时，仅仅考虑了手续费等，并未考虑可以购买的股数是 100 的倍数，回测环境与中国股票市场真实交易环境存在差异，因此若将模型投入使用，在买卖股票数量的需要加以限制。

2.相关理论介绍

2.1 多因子模型

2.1.1 资产组合理论

资产组合理论由马科维茨于1952年在其论文中所提到，研究的是怎样实现合理的资产配置问题，用风险资产收益率的平均值作为期望回报，用风险资产收益的标准差来作为风险，基本假设是，市场上有n种风险资产，收益率为$r_1, r_2, r_3, \cdots, r_n$投资者进行资产配置，在各类风险资产中配置不同的比例，比例分别为$w_1, w_2, \cdots, w_n$，投资组合的收益率假设为r_p，$r_p = \sum_{i=1}^{n} w_i r_i$，其中$\sum_{i=1}^{n} w_i = 1$。投资组合的期望收益率可以表示为：

$$E(r_p) = \sum_{i=1}^{n} w_i E(r_i) \#(2.1)$$

方差为：

$$Var(r_p) = \sum_{i=1}^{n} w_i^2 \, Var(r_i) + \sum_{i \neq j} w_i w_j \, Cov(r_i, r_j) \#(2.2)$$

投资者资产配置的前提条件是对市场未来情况形成一致预期，即$r_1, r_2, r_3, \cdots, r_n$的概率分布，从而确定自身的预期收益目标，之后再决定在各类风险资产的配置比例$w_1, w_2, \cdots, w_n$，达成收益目标。

2.1.2 资本资产定价模型

夏普等人进一步研究马科维茨的资产组合理论，并总结出了一个资产的定价模型，称为资本资产定价模型，也就是常说的CAPM模型，该理论将资产的收益率和市场超额收益率建立起一元的线性关系，即市场风险因子的暴露是资产获得收益率的原因。并将此模型用于给资产定价，该模型的数学表示形式：

$$\mathrm{E}(r_i) - r_f = \beta_i (E(r_m) - r_i) \#(2.3)$$

其中E()表示的是期望，r_i是资产i的收益率，r_f是无风险收益，r_m表示市场收益率，$\beta_i = cov(r_i, r_m)/var(r_m)$（2.4）用来描述资产收益率随市场收益率波动的敏感程度刻，学界也可将β_i称作是资产i暴露在市场风险中的程度。

通过将资产的预期收益和市场的收益及市场风险的暴露程度构建成一元线性的数学模型，资本资产定价模型在现代金融学的研究中是重要的基础性理论。

2.1.3 套利定价理论

Ross（1976）提出了的套利定价理论（APT），但不再使用 CAPM 理论的一些假设条件，将一元线性模型扩充成立多元的线性模型，增加了多个因子来共同对资产收益率进行解释，其中增加的因子有 GDP 增长率因子、通货膨胀率因子但其并未给学界带来解释力强的多因子模型。成为第一个提出多因子模型的理论，资产的预期超额收益率由多元线性模型决定：

$$\mathrm{E}(r_i^e) = e_i + \beta_i^{'}\lambda \#(2.5)$$

其中 $\mathrm{E}(r_i^e)$表示资产 i 未来的超额收益率，β_i表示资产 i 在市场风险上的因子暴露程度，λ是因子溢价，e_i是定价误差。

在套利定价理论的框架下，资产的收益率并不能由单一的风险因子决定而应该加入更多的因子共同解释资产的预期超额收益率，但套利定价理论并未给出具体有哪些因子决定了资产的预期超额收益率。

2.1.4 Fama-French 三因子模型

Fama-French（1992）的三因子模型和套利定价模型相同的是，数学模型是多元线性模型，比套利定价理论更为进步的是，将三个解释收益率的因子确定下来，分别是市场因子，企业规模因子以及企业价值因子。将这三个因子和资产收益率建立模型，解释资产收益率。Fama-French 三因子模型也成为多因子模型的开山鼻祖，三因子模型的数学表示如下：

$$\mathrm{E}(r_i) = r_f + \beta_{i,m}RP_m + \beta_{i,\mathrm{SMB}}F_{SMB} + \beta_{i,HML}F_{HML} + e_i \#(2.6)$$

其中 $\mathrm{E}(r_i)$是资产 i 未来的超额收益率，r_f为无风险利率，$\beta_{i,m}$是暴露于市场风险的暴露程度，RP_m是市场组合获得的超额收益率，$\beta_{i,\mathrm{SMB}}$是暴露于公司规模因子下的暴露大小，F_{SMB}规模因子收益率，$\beta_{i,HML}$是暴露于公司价值因子的暴露大小，F_{HML}是价值因子收益率，e_i是定价误差。

Fama-French 提出的三因子模型，首次将资产的预期超额收益率归因于市场因子、企业规模因子以及企业价值因子的暴露上。也成为了学者展开多因子模型

研究的基础。

2.2 IC 分析法

IC 来自于 Information Coefficient，由这两个单词的大写首字母组成，中文中称为信息系数，是主动管理领域最喜欢用的指标。IC 衡量预测变量的预测能力，其定义通常为 t+1 时刻的预测收益率与真实收益率在截面上的相关系数。在实际应用中，t+1 时刻的预测收益率往往由 t 时刻的预测变量代替，因此 IC 的定义为 t 时刻预测变量和 t+1 时刻股票收益率在截面上的相关系数[5]：

$$\mathrm{IC} = corr(z_{it}, R_{it+1}) \quad (2.7)$$

式中z_{it}为 t 时刻股票 i 的预测变量取值，R_{it+1}为该股票在 t+1 时刻的收益率。IC 衡量了预测变量所含的未来收益率信息含量，从定义中可以得知，IC 取值范围是-1 到 1，绝对值越高表示预测能力越强。根据经验研究，若按日频收益率来评价，IC 高于 2%就是优秀的预测变量。除了看 IC 绝对值，其他相关的指标还包括信息比率（InformationRatio，IR）、IC 绝对值大于 2%的比例和 IC 的 t 值等，详细见表

表 2.1 IC 分析法指标

指标	计算方法	含义
IC	$corr(z_{it}, R_{it+1})$	变量信息含量
IR	IC 均值/IC 标准差	预测能力的稳定性
|IC|≥2%	IC 序列均值大于 2%的比例	预测能力强弱

2.3 随机森林算法模型

2.3.1 决策树

决策树算法模型是 1966 年 Hunt 等人提出，学者 Quinlan1975 年在进一步研究的时候，也给出了一种分类预测的算法，名称为 ID3，在该算法沿用了信息论中的相关观点及思想，该算法提出特征判别能力的测度用信息增益来描述，建模的算法采用迭代算法。Breiman 等人在 1984 年提出既可以分类又可做回归的二叉树算法，称为 CART 算法，在各类领域都看得到该算法的身影。1993 年 Quinlan 对 ID3 算法进行改进，改进方面包括遵循什么样的规则进行派生、剪去不必要的

枝干、更加有效的进行缺失值的处理。将改进后的算法命名成了 C4.5 算法，不就进行商业改进后进化成了 C5.0 算法，CART 算法和 C5.0 算法也是目前在学界业界应用较为广泛的决策树算法。

归纳算法是决策树算法的核心构建，在建模开始阶段，把所有数据都集中在一个节点上，将这个节点称为根节点，再通过数据特征来选取最优的划分数据的划分条件，进而将原始数据集合划分为多个纯度更大的数据子集，继续进行下去，直到达到满足预先确定的条件才终止，最终呈现的结果形似一个倒着长的树，决策树算法名字的来源也正是如此，决策树算法的建立步骤：

（1）选择数据特征：从初始的数据特征中选取一个数据特征，用于进行向下分裂的判断，找到分类效果最强的数据特征

（2）生成决策树：根据上一步选择的数据特征进行判断向下生成节点，通过计算信息增益大小的方式来决定是否还继续向下分裂，直到最后信息增益很小和不能再对数据集进行分类后，停止决策树的向下分裂

（3）剪枝：由于决策树算法容易出现过拟合的状况，因此需要通过剪去一些纯度不够的节点减小树的结构和规模，通常也称为剪枝缩小树的结构和规模。不同的算法衡量树节点不纯度的方法不同，通常衡量指标有 Gini 指数、错分率以及熵等。

与同为分类算法的逻辑回归相比较，决策树算法每次只划分单一的特征，进行拟合，拟合函数为区间的阶梯函数，决策树算法跟人类的思考方式较为相似，模型可以获得很好的解释性，并且分类规则可视化，因此，决策树算法的构造，决定了它在构建择时模型方面有很强的适用性。

2.3.2 集成方法

集成学习方法，是一种用来增强预测有效性的统计学习数学算法。该方法不是分类模型构建，而是通过采用某种方法组合多个模型，用来训练数据。集成方法时通一系列分类器的构建，构建的分类器往往是小型的决策树，在预测的时候进行加权投票，直至数据点分类的完成。

集成的学习方法分为两类：分别是 bagging 和 boosting 集成方法，bagging 是来自于 bootstrap aggregating，指的是抽样方法采用 Bootstrap，在输入的数据集

中进行随机抽样，抽样得到一组新的训练集数据，并对得到的每一个新的训练集构建预测器，称该预测器为基预测器，并将所有的预测器进行组合得到最终的预测模型，所有的基预测器进行投票产生最终的预测模型。，通过这种集成方法可以极大程度上降低方差，也避免出现过度拟合的情况。Boosting 集成方法是一类目标设定成着重训练上一步模型误分类，该集成放大从一个简单的分类器进行原数据拟合，在第一个模型拟合完毕之后，计算出此模型进行分类的误差，再重新给数据一个权重，给上一个模型被错误分类的样本相对正确分类一个更高的权重，观测样本在每一步的计算中不断被重新分配权重，即分配给上一个模型错误分类的样本集一个更高的权重，迭代过程中生成的每一课树进行投票以及根据树的精确度计算权重，得到最终的预测结果。相比于 Bootstrap，Boosting 方法有较高的预测准确率，但是 Boosting 方法也更可能出现过度拟合的情形。在此之后，随机森林算法被提出，与 bagging 的算法很相似，随机森林算法是 bagging 方法的一个改良算法，随机森林算法在训练数据、调节参数方面简单并且有效，因此也成为了现在被广泛使用的集成算法。

2.3.3 随机森林算法

与 bagging 集成算法相似，随机森林算法的构建基本过程为，先创建若干个初始决策树，每个决策树创建时，不仅使用了 bootstrap 方法进行样本的随机选取，而且还随机选取特征变量数据，利用这种方式，树与树之间的相关系数得到了降低，但最后还是要通过每棵树统计投票的结果，来确认最后的分类信息。随机森林作为一种特定的集成方法，在解决大量特征变量的问题尤为合适。而在其他的集成算法模型中，更显著的特征变量可能会掩盖某些特征变量的效果。随机森林每科决策树上适用的特征变量是随机选取的，这样才能有效的检测每个特征变量的行为以及产生的贡献，因此对比与单颗决策树，随机森林提高预测精确度以及误差降低方面有很大的提升。

图 2.1 随机森林机器学习计算原理图[1]

2.4 LSTM 神经网络模型

2.4.1 RNN

RNN，是 Recurrent Neural Network 的单词的简写，中文名叫做循环神经网络，它与普通的神经网络模型相比不同的是，循环神经网络加入了对时序统计数据的处理。以股票多因子模型为例，普通的神经网络在一给定的时间截面，输入变量为因子数据，输出变量为下一期的超额回报率；循环神经网络则是将股票的长期因子数据作为时间序列，输入变量为过去一段时间的历史数据，由于充斥在资本市场上的数据都比较持久，使得循环神经网络可以把握住“历史重现”的好时机。

如图表所示，循环神经网络读取某个输入变量 x，经计算后输出一个值 o，通过神经网络的循环结构到下一个结构中，从循环神经网络结构的表面看，网络结构复杂且难以理解，所以将它展开成图表右侧的形状，时间序列数据的序列索引为 1 到 T，在 t 时刻附近的网络结构，将在序列索引号 t 时刻训练样本的输入变量表示为x^t，同理，$t-1$ 和$t+1$ 时刻的训练样本输入为x^{t-1}和x^{t+1}；h^t表示为 t 时刻模型的隐藏状态，隐藏状态受到x^t和h^{t-1}两个值的共同影响，o^t表示 t 时刻模型的输出变量，并且o^t只被当前模型的隐藏状态h^t决定，与其他变量无关系；y^t表示为 t 时刻该样本序列的真实值，L^t表示模型 t 时刻的损失函数，又o^t和y^t通过计算得出，三个矩阵 U、V、W 表示为模型的参数，并且在这个模型中是共享的

[1] 林晓明，《人工智能选股之随机森林模型》，华泰证券

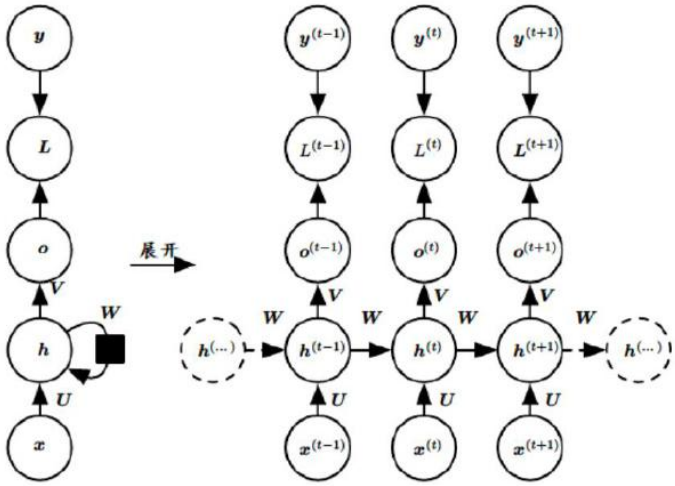


图 2.2 LSTM 神经网络计算原理图[2]

循环神经网路在本质上也是虽权值和阈值进行优化调整，采用的方法是 BP 算法，仅仅增加了时间序列，结合一起就称为 BPTT，公式表示如下：

$$令：\delta^{k-1}=\frac{\partial C^t}{\partial h_*^{k-1}}=\left(W_{hh}^t\delta^k\right)\odot\sigma'\left(h_*^{k-1}\right)\#(2.8)$$

$$则有\frac{\partial C^t}{\partial \omega_{ij}}=\sum_{k=1}^{t}\frac{\partial h_*^k}{\partial \omega_{ij}^k}.\frac{\partial C^t}{\partial h_*^k}=\sum_{k=1}^{t}h_j^{k-1}\,\delta_i^k\#(2.9)$$

其中C^t用来表示在 t 时刻模型输出变量值与真实值之间的交叉熵，在上述公式中，如果 sigma 为 tanh 或 sigmoid 函数，根据 delta 的递推式关系，当时间出现较大跨度时，delta 就会变的很小，进而 BP 梯度的变小带来“梯度消失”的效果，另一方面，对于W_{hh}^t而言，在循环神经网络中，每个时刻的W_{hh}^t都指的是相同的参数，因此在W_{hh}^t的累乘不会出现在 delta 中，而在多次累乘之后，数值的分布较为统一，趋近于无穷大或者 0。趋近无穷大的情况称作“梯度爆炸”，趋近于 0 的情况称作“梯度消失”。

2.4.2 LSTM

上文提及，循环神经网络出现产生梯度消失的问题，不能很好用于处理长序列的数据的情况，学者随后展开对此问题的研究，Hochreater 和 Schmidhuber 在循环神经网络的基础上，提出了一个深化后的神经网络，成为长短期记忆神经网络，简写为 LSTM，该神经网络通过设计隐藏层的神经元很大程度上缓解了循环神经网络的梯度消失问题

[2] 林晓明，《人工智能选股之循环神经网络》，华泰证券

在循环神经网络模型中，t 时刻都有一个隐藏状态h^t如果将每一层神经网络中的o^t、y^t和L^t都省略掉，模型可以简化成如下图表所示，通过图表的显示我们可以看出，h^{t-1}与x^t一起影响h^t，h^t一方面用于计算模型当前层的损失，另一方面也可用于计算模型下一层的隐藏状态h^{t-1}。

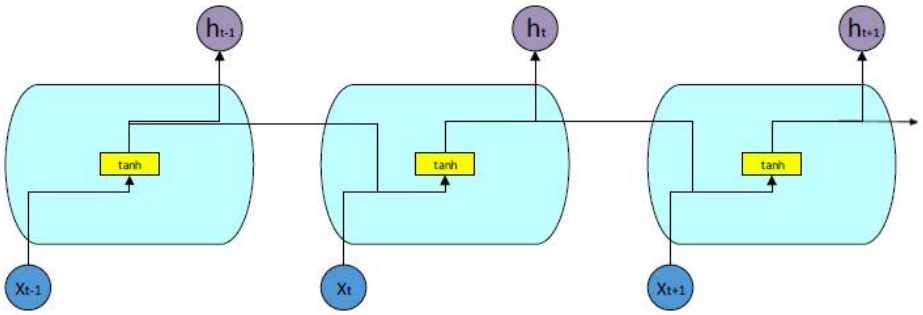


图 2.3 LSTM 神经网络计算原理图[3]

在长短期记忆神经网络模型中，在 t 时刻每个序列索引的位置都会被向前传播，除了和循环神经网络有一样的隐藏状态，还比循环神经网络多了另外一个隐藏状态，如图表 2.4 中的标黑色横线处，多出来的隐藏状态我们叫做C^t细胞状态（Cell State），该细胞也是长短期记忆神经网络模型的一个单元，细胞状态在长短期记忆神经网络模型中的作用实质上相当于隐藏层状态在循环神经网络模型中的作用。

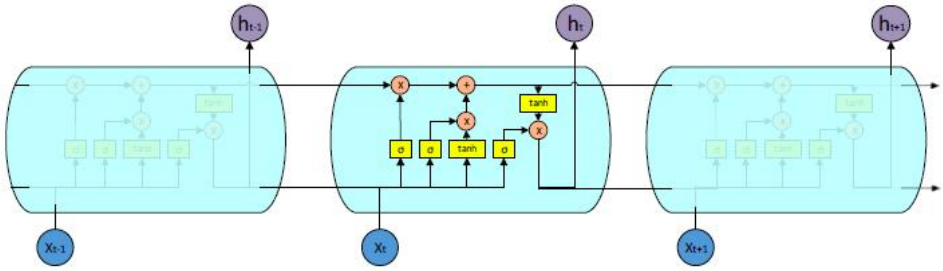


图 2.4 LSTM 神经网络计算原理图[4]

2.4.3 LSTM 的单元结构

[3] 林晓明，《人工智能选股之循环神经网络》，华泰证券

[4] 林晓明，《人工智能选股之循环神经网络》，华泰证券

长短期机器神经网络结构复杂，不仅仅包含细胞状态C^t，还有门控结构，门控结构是除了细胞状态C^t之外的长短期记忆神经网路内部结构单元的统称。在每个序列索引位置 t 长短期记忆神经网络模型的门控结构通常包括输入门、输出门以及忘记门：本质上这三个门就是权重值，也可以类比于电路中控制电流的开关。当该值取 1 时，表示电路处于闭合状态，电流量在无损耗的通过；当该值取 0 时，表示开关处于断开状态，流量被完全堵塞住，需要通过激活函数来实现[0,1]的取值，

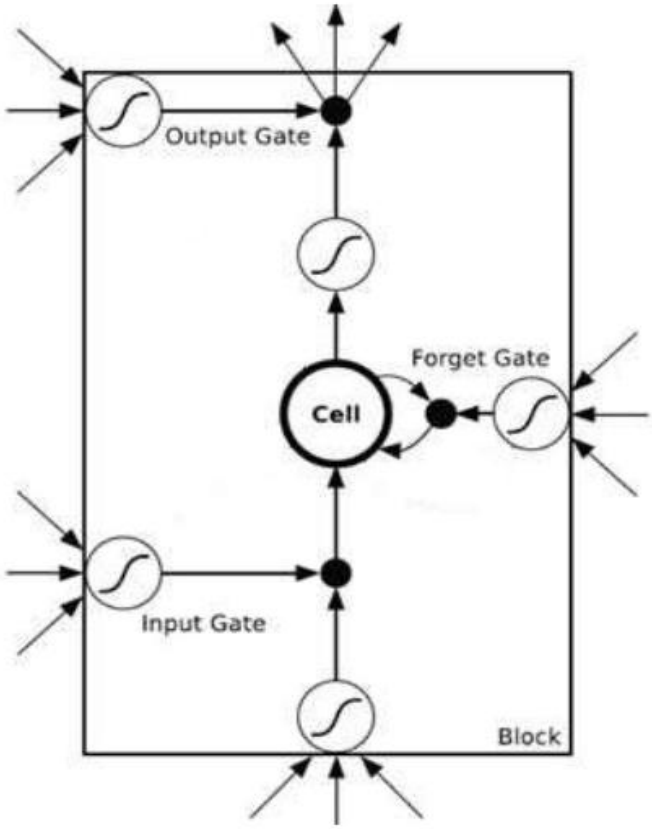


图 2.5 LSTM 神经网络计算原理图[5]

由上文的长短期记忆神经网络模型结构分析中可知，长短期记忆神经网络模型智能只能避免出现像循环神经网络模型一样出现“梯度消失”的情况，“梯度膨胀”并不是一个很严重的问题，但是梯度膨胀会导致将参数修改的与当前的数值相差很大，这就使得，会把之前做的大量的优化工作变成无用功，如何避免出现“梯度膨胀”的情况，可以采用梯度裁剪的方法进行优化。

[5] 黄志文，邹璐，《递归神经网络 RNN——长短期记忆细胞 LSTM 的分行业多因子预测》，国信证券

总结来说，长短期记忆神经网络模型的内部主要可分为以下几个部分：第一个阶段是忘记阶段，在忘记阶段的工作主要为对前一个节点的输入进行选择性的忘记。用通俗的语言解释就是“忘掉不重要，记住重要的”。可以用忘记门进行判断；第二个阶段是选择记忆阶段，相反，在这个阶段会对前一个节点的输入进行选择的记忆，重点记录下来哪些很重要，哪些是不重要的，可以通过输入门进行控制。最后一个阶段是输出阶段，在输出阶段，需要将上面两步得到的结果进行相加，随后就可以将结果传输到下一个隐藏状态h^t，在这个阶段决定出当前状态输出哪些结果，这个阶段主要通过内部机构的输出门控制。

2.5 LightGBM 模型

2.5.1 直方图算法

直方图算法是 LightGBM 模型中的算法，主要是先进行离散化的操作，将连续的特征值离散成 m 个整数，再构造一个直方图或者柱形图，柱形图的宽度为 m.

在进行数据遍历时，离散化后的 m 个整数将会作为索引,依次在直方图上累计统计量，遍历万所有数据后后，直方图已经累计完成了所需的统计量，再根据离散值，进行再一次遍历，找到最佳的分割点。而在 XGBoost 的模型中，则需要将所有离散化的值都进行遍历，在直方图算法中只需要遍历 m 个直方图的离散值。

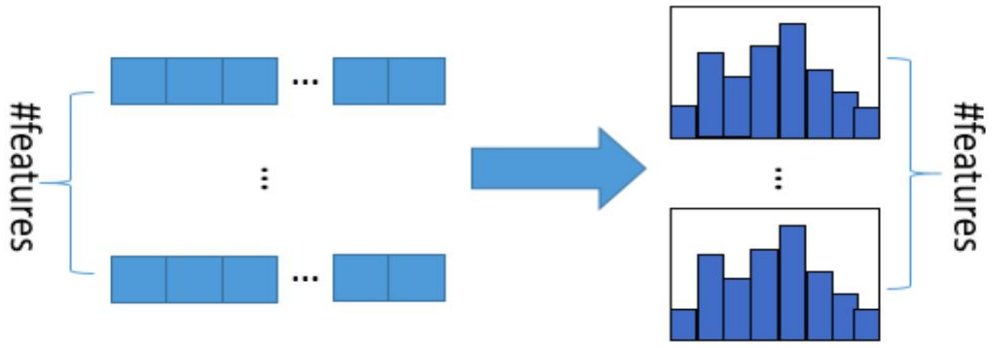


图 2.6 直方图算法计算原理图[6]

直方图算法的使用存在很多的优点，首先内存损耗的明显降低，其次，直方

[6] 微软亚洲研究院，《开源|LightGBM：三天内收获 GitHub1000 星》

图算法不需要再额外去存储排序之后的结果。只需要将特征离散化后的离散值进行保存即可。直方图算法在计算上的代价也相比于 XGboost 大幅减少。XGboost 的预排序算法在进行每一次特征值遍历的时候，都需要进行一次分类的增益的计算，直方图则只需要 k 次的计算。

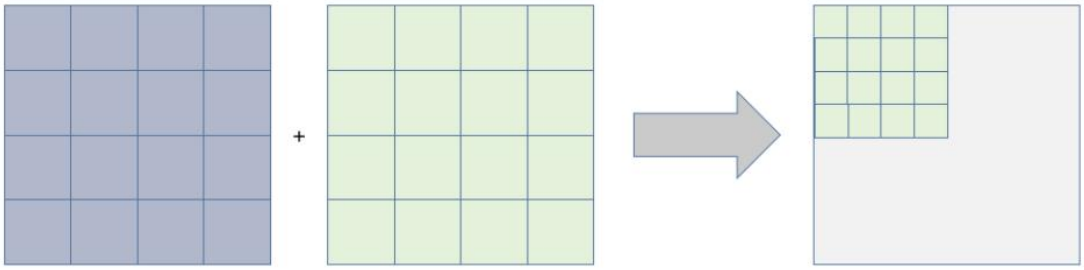


图 2.7 直方图算法计算原理图[7]

然而，直方图的算法也并不是完美的，在特征值被离散化之后，找到的分割点并不是很精确，因此结果会受此影响。但在不同的数据集上使用直方图算法的结果却表明，特征值被离散化之后的分割点对于精度的影响程度并不大，甚至有时候会更好。因为决策树算法本来就属于一类相对较弱的模型，分割点找的精不精确并不会对结果产生太大的影响，相对不精确粗糙的分割点也会有正则化的效果，也非常有效的防止出现过度拟合的状态；在一颗树的情况下，即使训练误差比精确分割的算法要大，但是在梯度提升的框架下，对于最后的结果并没有太大的影响

2.5.2 LightGBM 的直方图做差加速

父节点的直方图与它兄弟的直方图做差可以得到一个叶子的直方图，通过直方图的构造，需要将叶子上的所有数据都进行遍历，但是如果直接进行直方图做差，仅需遍历直方图的 m 个桶。LightGBM 将可以利用这个方法构造单个叶子的直方图，得到兄弟叶子的直方图也仅需要付出微小的代价，并且可以提升一倍的速度。

[7] 微软亚洲研究院，《开源|LightGBM：三天内收获 GitHub1000 星》

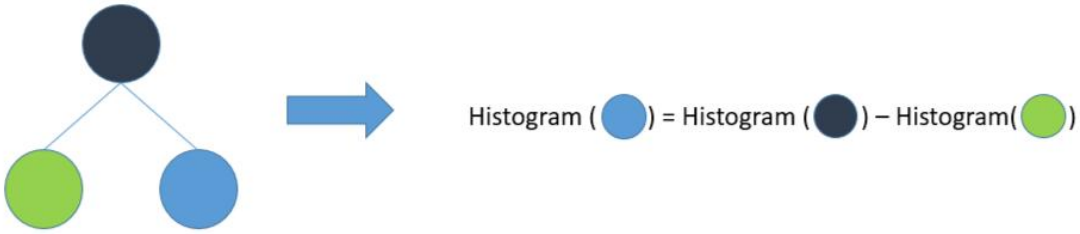


图 2.8 直方图作差计算原理图[8]

2.5.3 带深度限制的 Leaf-wise 的叶子生长策略

在过一次数据之后 Level-wise 可以同时分裂同一层的叶子，在进行多线程优化时更加容易，也可以很好的控制复杂度，不容易出现过拟合的情况。但是实际上 Level-wise 算法相对来讲较为低效，因其在对待同一层叶子的时候不加以区分，存在很多没必要的计算。实际上多叶子的分类增益很低，没有必要进行分裂和搜索。

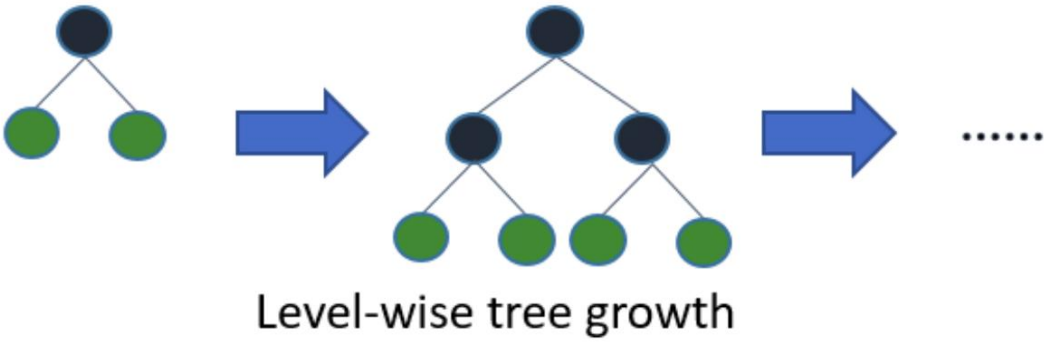


图 2.9 叶子生长策略原理图[9]

相对来讲 Leaf-wise 则是一种更加高效的策略，从当前的叶子中每次找到分裂增益最大的一个叶子，然后进行分离，如此持续循环下去。因而相比于 Level-wise，分裂次数在相同的情况下，Level-wise 将降低很多的误差，精度更好。Leaf-wise 可能会长出比较深的决策树，而出现过度拟合的情况。因此 LightGBM 在 Level-wise 上加入了一个最大深度的限制，不仅保证高效率的运行，还防止出现过拟合的情况。

[8] 微软亚洲研究院，《开源|LightGBM：三天内收获 GitHub1000 星》

[9] 微软亚洲研究院，《开源|LightGBM：三天内收获 GitHub1000 星》

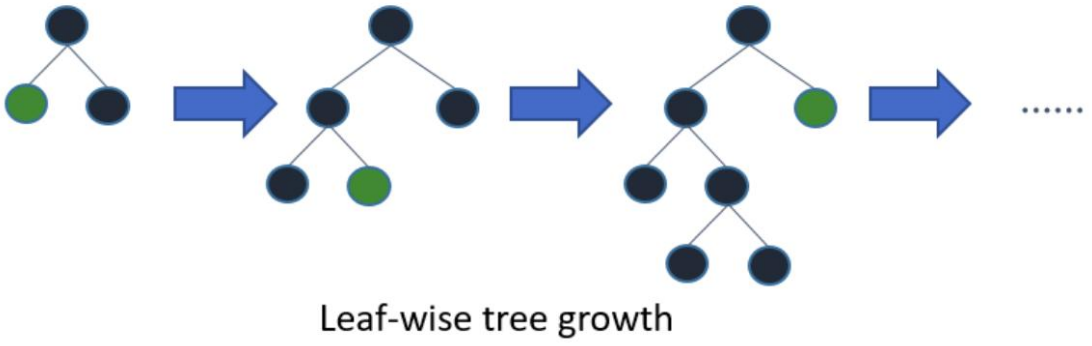


图 2.10 叶子生长策略原理图[10]

2.5.4 直接支持类别特征

目前，大部分机器学习算法都无法直接支持类别特征，通常要将类别特征进行转化，转化成多维的 one-hot 编码特征，这样就降低了时间和空间上的效率。而在实践中使用类别特征是很常见的，因此也正是基于这个考虑，LightGBM 对类别特征的支持进行了优化，类别特征可以直接输入，并不需要再进行额外的 one-hot 编码展开。增加了类别特征的决策规则在决策树算法中。相关实验证明，训练速度可以达到 8 倍的增长，并且有一致的精度。

2.5.5 直接支持高效并行

LightGBM 还能支持高效运行，原生支持并行学习，目前 LightGBM 支持两种并行方式，特征并行和数据并行。

特征并行是分别在不同机器不同特征集合上找到最精确的分割点，然后在不同机器之间同步找到的最精确的分割点。数据并行则并不是让不同的机器在本地进行直方图的构造，然后合并，最后在合并后的直方图上找到最精确的分割点。

LightGBM 算法都优化了这两种并行方法，在数据并行中使用的是分散的规则，分摊直方图合并的任务到不同的机器上，降低计算量和通信量。再利用直方图做差，更进一步的减少一半的通信量。

2.6 本章小结

本章主要对本文涉及到的金融学相关理论知识进行了梳理，多因子模型的基础理论包括资产组合理论、资本资产定价模型、套利定价理论以及三因子模型。涉

[10] 微软亚洲研究院，《开源|LightGBM：三天内收获 GitHub1000 星》

及到的因子研究方法有 IC 分析法，以及本文主要运用的随机森林、LSTM 神经网络、Lightgbm 等机器学习算法。

3.因子数据处理

本文是建立在多因子模型的基础上，经过 IC 分析或随机森林机器学习算法对因子进行筛选，再使用机器学习算法对未来股价的收益率进行预测，从而构建一个量化投资策略，因此需要获得完整的因子数据以及股票收益率数据，并对获得的数据进行预处理工作，从而得到科学的有效的计算结果。

3.1 数据来源及数据内容

本文的数据获取及处理工作都基于 BigQuant 人工智能量化投资平台，该平台拥有 A 股、美股、港股、期货、期权等多市场海量数据，全面支持主流 AI 框架。可进行数据获取数据清洗处理等工作。

本文选取 2015-01-01 至 2020-12-31 时间段的中国 A 股市场的因子数据以及股票收益率数据作为训练集，2021-01-01 至 2021-11-30 时间段的数据作为测试集，构建量化投资策略进行回测分析。

3.2 候选因子选取

根据多因子模型理论，股票的超额收益率是来自于不同的因子，因此在构建多因子量化投资策略之前，对候选因子库的选取是关键一步。为了能够对股票收益率进行充分解释，学术界也为此展开了许多年的研究，从三因子模型开始，学术界及业界不断提出新的多因子模型，同时挖掘出众多可以用来解释股票收益率的因子，本文候选因子建立在学术界及业界对因子研究的基础上，选取了财务类因子、行情类因子以及预期类因子中共 64 个指标因子作为候选因子。其中财务类因子包括估值类因子、规模类因子、成长类因子、质量类因子；行情类因子包括风险因子、流动性因子、技术因子、动量因子以及资金流因子。估值因子反映当前股票高低，即该公司是否具有投资价值，规模类因子反映目前公司的市值规模大小，体现了市值大小对收益率的影响，成长类因子用来反映未来公司的成长以及发展潜力，质量类因子主要反应了公司财务质量好坏程度。风险因子主要反映的是过去一段时间内资产价格的波动性大小；流动性因子主要反映的是过去一段时间内资产的流通性强弱；技术因子是各类技术指标的集合；动量因子反映了过去一段时间股票的价格动量，即能够根据动量效应的大小来对股票价格走势进

行预测；资金流因子反映市场的对股票的追捧程度。部分候选因子如下图所示，完整版详见附录。

表 3.1 候选因子库

财务类因子	
估值因子	
EP_TTM	净利润 TTM/总市值
EP_LYR	净利润（最新年报）/总市值
BP_LF	净资产 TTM/总市值
OCF_TTM	经营性现金流 TTM/总市值
SP_TTM	营业收入 TTM/总市值
SP_LYR	营业收入（最新年报）/总市值
FCFP_LYR	自由现金流 TTM/总市值
PEG	市盈率/净利润同比增长率 * 100
规模类因子	
LN_MV	对数总市值
LN_FLOAT_MV	对数流通市值
成长类因子	
SALES_GR_TTM	营业收入增长率_TTM 同比
NET_PROFIT_GR_TTM	净利润增长率_TTM 同比

3.3 数据预处理

在获取到 A 股市场的因子数据以及股票收益率数据之后，数据通常是不完整缺失一些数据、不一致的数量级以及存在一些极端值，则在使用这些数据之前，需要进行一些预处理工作，从而得到高质量的数据。数据预处理包括缺失值处理、去极值、标准化处理、中性化处理等，因本文将用到不同的因子筛选方法，不同的方法会存在些许数据预处理上的差异。

3.3.1 缺失值处理

缺失值的处理方法大致分为两类，一类是直接删除，一类是插补法。直接山出发即将含有缺失值数据的样本行（列）或者特征数据行（列），以得到一个完整的不含任何缺失值的列表，直接删除法一般在缺失值较少的时候使用。插补法则是用数据填充缺失值部分，包括用特殊值进行填充或者用平均值进行填充，插补法一般在数据缺失较多是使用。本文在用 IC 分析法及随机森林机器学习算法

对因子进行筛选中，因数据量庞大，缺失值数据相对较少，因此采用的是对缺失值所在行（列）进行直接删除处理。

3.3.2 极值处理

为避免一些可能存在的极端值对整个数据集合的影响，在使用数据之前还需要进行去极值化的处理，常见的去极值方法有 MAD 法、3 倍标准差法、百分位法。

（1）MAD 法

MAD 是 Median Absolute Deviation 的首字母缩写，叫做绝对值差中位数法。绝对值中位数法的基本思想是将因子和平均值之间的距离计算出来然后进行离群值的检测。第一步是找到所有因子的中位数，然后将每个因子减去中位数，计算得到绝对偏差值的中位数 MAD，最后确定 n，得到$[X_{median} - nMAD, X_{median} + nMAD]$的合理范围，调整超出该合理范围的因子值。本文使用的 MAD 方法 n 值取值为 5。

（2）3 倍标准差法

3 倍标准差法又称为标准差法。标准差继基于因子的平均值X_{mean}体现因子的离散程度，可以通过用$X_{mean} \pm n\sigma$在离群值的处理过程中衡量因子和平均值之间的距离，该方法与 MAD 方法处理的逻辑相似，第一步是计算因子的平均值和因子的标准差，再确定参数 n，从而类似于 MAD 方法将因子值的合理范围确定为$[X_{mean} - n\sigma, X_{mean} + n\sigma]$，对因子值做出调整，在该方法下，n 的取值为 3

（3）百分位法

本文利用百分位法去极值计算的逻辑是将因子值进行升序的排序，对排位百分位高于 99%或排位百分位低于 1%的因子值，进行类似于 MAD、3σ的方法进行调整。

3.3.2 标准化

不同的数据拥有不同的量级，数量级的差异将导致量级较大的属性占据主导地位，因此需要消除样本不同属性具有不同量级时的影响，对数据进行标准化处理，学术界及业界一般采用标准化及 z-score 标准化方法。

（1）$min - max$标准化

$min-max$标准化的基本思想是将最大值标准化为 1，最小值标准化为 0 或 1，其他值在 0 到 1 之间分布，对于每个属性 A，设置属性的最小和最大值为$minA$和$maxA$，通过该方法每一个原始值都讲被标准化到[0,1]之间，计算公式为：

$$x=\frac{x-x_{min}}{x_{max}-x_{min}}\#(3.1)$$

其中x为原始值，x_{min}为最小值，x_{max}为的最大值。

（2）z-score 标准化

在 z-score 标准化的方法下，通常将均值归一化为数值 1，方差也同样归一为 1，在属性 A 的最大值和最小值未知的情况或者有超过范围数据的情况下，该种方法特别适用，计算公式为：

$$x=\frac{x-\overline{x}}{\sigma}\#(3.2)$$

其中x为原始数据的均值，$\overline{x}$为原始数据的标准差，目前 z-score 标准化方法时应用最多的一种方法，本文在进行数据标准化时使用的是 z-score 标准化方法。

3.3.3 行业及市值中性化

行业和市值是两个十分显著对因子有影响力的因素，大部分的因子中都包含了行业和市值因子的影响，因此在利用 IC 分析法进行因子有效性分析前，还需要对行业及市值进行中性化处理，避免单纯按因子排序造成的行业或公司大小的偏向性。本文将利用回归法对因子进行行业哑变量和对数市值多元线性回归，将残差作为中性化后的因子值

3.4 基于 IC 分析法因子有效性检验

本文基于 2015-01-01 至 2020-12-31 的数据作为样本数据，对样本数据进行 IC 法分析因子有效性，即按照 2.2 节 IC 分析法的内容，先对所有的因子值进行缺失值处理、去极值处理、标准化处理以及行业市值中性化处理等预处理，计算预处理之后的因子与滞后 22 个交易日后的股票收益率的秩相关系数作为 IC 值，并且得到了这些因子在不同截面期的 IC 序列，再进行计算得到 IC、IR、IC 标准差等各类检验指标值。

本文的实证研究都基于 bigqunant 量化投资平台，部分实证结果如下，完整实证结果详见附录：

表 3.2 候选因子库 IC 分析法计算结果

财务类因子	IC 均值	IC 标准差	IR 值	|IC|>0.02 的比率
估值因子				
EP_TTM	-0.03	0.06	-0.55	80.67%
EP_LYR	-0.03	0.06	-0.57	78.15%
BP_LF	-0.05	0.11	-0.45	88.24%
OCF_TTM	0.04	0.06	0.58	87.39%
SP_TTM	-0.04	0.1	-0.36	90.76%
SP_LYR	0.03	0.1	0.36	89.08%
FCFP_LYR	0	0.04	0.12	56.30%
PEG	0.01	0.04	0.18	68.91%
规模类因子				
LN_MV	-0.01	0.17	-0.04	90.76%
LN_FLOAT_MV	0.01	0.12	0.05	85.71%
成长类因子				
SALES_GR_TTM	0.03	0.06	0.4	80.67%
NET_PROFIT_GR_TTM	0.03	0.06	0.49	79.83%

根据 2.2 节的 IC 分析法理论，IC 值大于 2%的的因子为有效因子，IC 大于 2%的比例以及 IR 值则反映了因子的稳定性，即可对所得数据进行因子有效性分析及筛选出量化投资模型所需的因子。本文将对初始因子做如下筛选，筛选出 IC 均值大于 2%以及 IC 大于 2%比例大于 80%比例的因子，将其作为绝对规则，IR 值作为相对规则。并对同类型不同期现的因子，仅保留一个相对有效且稳定的因子，因规模类因子未达到因子筛选规则，但为能包含所有类型因子，对一个规模类因子做保留处理。其中估值因子、规模类因子、成长类因子因子计算结果如下，其它因子计算结果详见附录：

表 3.3 基于 IC 分析法的因子计算结果

财务类因子	IC 均值	IC 标准差	IR 值	|IC|>0.02 的比率
估值因子				
EP_TTM	-0.03	0.06	-0.55	80.67%
BP_LF	-0.05	0.11	-0.45	88.24%
OCF_TTM	0.04	0.06	0.58	87.39%
SP_TTM	-0.04	0.1	-0.36	90.76%
SP_LYR	0.03	0.1	0.36	89.08%
规模类因子				
LN_MV	-0.01	0.17	-0.04	90.76%
成长类因子				

SALES_GR_TTM	0.03	0.06	0.4	80.67%

3.5 基于随机森林算法因子重要性排序

本文基于 2015-01-01 至 2020 年 12-31 的数据作为样本数据，对样本数据进行随机森林机器学习算法分析因子重要性，根据 2.3 节的随机森林机器学习算法理论内容，仅对数据进行缺失值处理，随后将因子数据作为特征输入变量，将股票收益率分为两类并作为输出变量，利用随森林机器学习算法进行训练并输出因子重要性。随机森林机器学习算法参数如下：

表 3.4 随机森林机器学习算法输入参数

树的个数	最多考虑特征个数	数据的最大深度	每个叶子节点最少样本数	并行度	随机种子	算法类型
10	auto	30	200	1	0	分类

本文的实证研究都基于 bigqunant 量化投资平台，部分实证结果如下，详情见附录：

表 3.5 基于随机森林算法的因子重要性排序

财务类因子	重要性得分
估值因子	
FCFP_LYR	0.148612
OCF_TTM	0.131956
BP_LF	0.126081
PEG	0.125612
SP_LYR	0.119159
EP_TTM	0.118628
EP_LYR	0.115401
SP_TTM	0.114552
规模类因子	
LN_MV	0.124356
LN_FLOAT_MV	0.123331
成长类因子	
SALES_GR_TTM	0.124572
NET_PROFIT_GR_TTM	0.121704

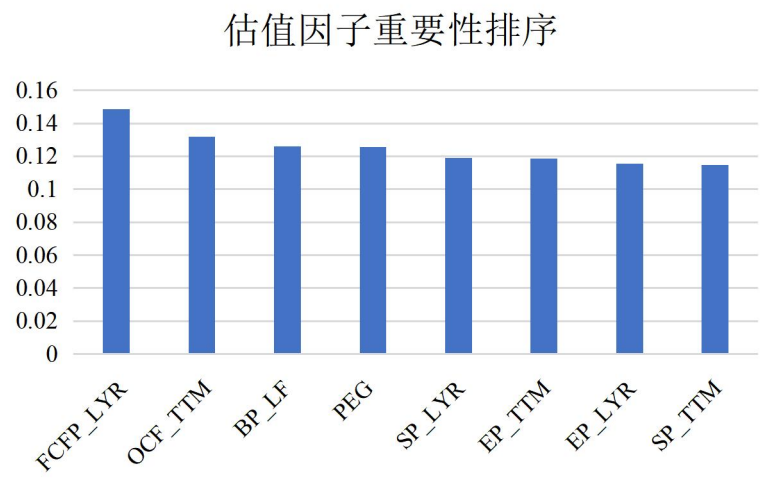


图 3.1 估值因子重要性排序

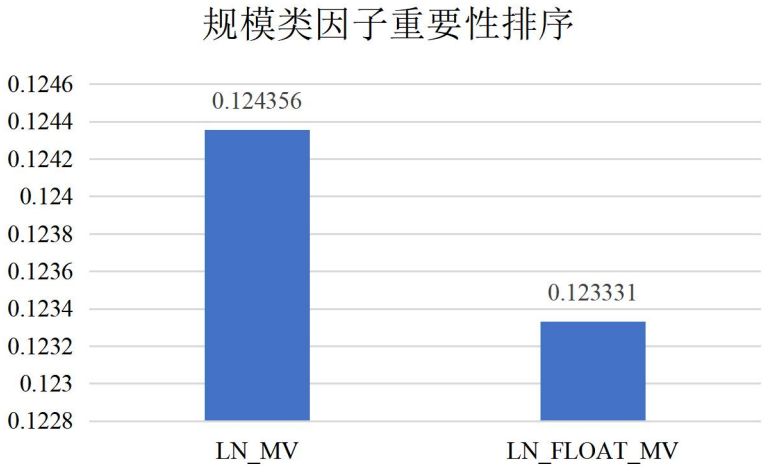


图 3.2 规模类因子重要性排序

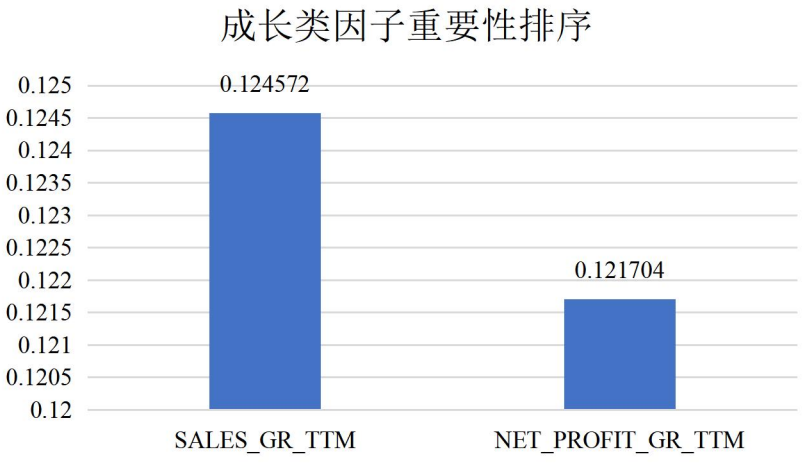


图 3.3 成长类因子重要性排序

因子的重要性得分越高表示该因子对于股票收益率进行分类越重要，即对股

票收益率的解释力度更好，在得到因子重要性得分之后，本文将选择相对得分更高的因子，为与 IC 分析法筛选后的因子形成对比，在数量以及种类上与经 IC 分析法保持一致，因此经过随机森林机器学习算法筛选后的因子部分结果如下，详情见附录：

表 3.6 基于随机森林算法的因子筛选结果

财务类因子	重要性得分
估值因子	
FCFP_LYR	0.148612
OCF_TTM	0.131956
BP_LF	0.126081
PEG	0.125612
SP_LYR	0.119159
规模类因子	
LN_MV	0.124356
成长类因子	
SALES_GR_TTM	0.124572

将经过随机森林机器学习算法以及 IC 分析筛选后的结果放在一起比较，从比较结果上可以看出，两种方法筛选后的因子存在较大差异，如财务类因子中的估值因子，在 IC 分析法中 FCFP_LYR、OCF_TTM、PEG 因为相对非有效而被排除在有效因子之外，而在随机森林机器学习算法重要性排序的方法中 FCFP_LYR、OCF_TTM 属于估值类因子中相对重要性较高的因子。两种方法在质量因子、流动性、资金流因子的筛选结果则完全不同。虽然两种方法对因子筛选后的结果存在较大差异，但对于本文的量化投资策略来讲，筛选出有效的因子仅是构建量化投资策略模型的第一步，在下一章将会对两种方法筛选后的因子作为量化投资策略模型的输入变量，对股票收益率进行预测，再进行回测，比较不同模型投资策略的效果，从而比较出两种方法筛选出来的因子对于构建量化投资策略模型更优。

表 3.7 IC 分析法与随机森林算法因子筛选结果对比

随机森林算法因子筛选结果	IC 分析法因子筛选结果
财务类因子	财务类因子
估值因子	估值因子
FCFP_LYR	EP_TTM
OCF_TTM	BP_LF

BP_LF	OCF_TTM
PEG	SP_TTM
SP_LYR	SP_LYR
规模类因子	规模类因子
LN_MV	LN_MV
成长类因子	成长类因子
SALES_GR_TTM	SALES_GR_TTM

3.6 本章小结

本章主要阐述了数据的来源以及本文如何对数据进行预处理工作，而后基于学术界以及业界的研究成果构建了候选因子库，用 IC 分析法及随机随机森林机器学习算法两种方法对因子进行了筛选，并将两组结果放在一起进行比较，为下一章量化投资策略模型的构建提供因子数据基础。

4.量化投资策略构建、回测及评价

4.1 量化投资策略评价指标

对于本文即将构建的量化投资策略，对多构建的量化投资策略进行实际环境下的模拟回测是必需的，本文将用以下指标衡量投资组合策略表现：

（1）策略总收益

策略总收益，即该策略在回测期间的绝对收益，计算公式为：

$$Total\ Returns = \frac{PV_{end} - PV_{start}}{PV_{start}} * 100\%\#(4.1)$$

其中，PV_{end}=策略最终股票和现金总价值，PV_{start}=策略开始股票和现金总价值。

（2）策略年化收益率

策略年化收益率即将策略总收益通过转化成一年的平均收益，是常用的衡量收益率的指标，计算公式为：

$$Total\ Annualized\ Returns = R_p = \left((1+P)^{\frac{252}{n}} - 1\right) * 100\%\#(4.2)$$

其中，P=策略总收益，n=策略执行天数。

（3）基准总收益

基准总收益即某个基准指数在回测期间的绝对收益率，本文用沪深 300 指数在回测期间的绝对收益率作为基准总收益，计算公式为：

$$Benchmark\ Returns = \frac{M_{end} - M_{start}}{M_{start}} * 100\%\#(4.3)$$

其中，M_{end}=基准最终价值，M_{start}=基准开始价值。

（4）基准年化收益

基准年化收益率即将基准总收益通过转化成一年的平均收益，是常用的衡量收益率的指标，计算公式为：

$$Benchmark\ Annualized\ Returns = R_p = \left((1+P)^{\frac{252}{n}} - 1\right) * 100\%\#(4.4)$$

其中，P=基准总收益，n=策略执行天数。

（5）Alpha 阿尔法

投资者构建的投资组合将面临着两种风险，系统性风险 Beta 和非系统性风

险 Alpha，我们把投资组合获得的与市场波动无关的回报叫做 Alpha，也叫做超额收益率，当 Alpha 取值大于 0 是，投资组合相对于市场收益率获得了超额收益，当 Alpha 取值为 0 时，投资组合相对于市场收益率获得了适当的收益，Alpha 取值小于 0 时，投资组合相对于市场收益率，获得了较少的收益。Alpha 的计算公式如下：

$$Alpha = \alpha = R_p - \left(R_f + \beta_p\left(R_m - R_f\right)\right) \#(4.5)$$

其中R_p=策略年化收益率，R_m=基准年化收益率，R_f=无风险利率，β_p=策略 Beta 值。

（5）Beta 贝塔

Beta 表示的是投资组合面临的系统性风险，是不能暴露的风险，它反应了投资组合对大盘变化的敏感性，也表示暴露的市场风险大小。当 Beta 取值小于 0 时，表面投资组合和基准收益走势反方向，当 Beta 的取值为 1 时，表明投资组合收益和基准收益走势相同，当 Beta 取值大于 1 时，投资组合收益和基准收益走势相同，但是比基准收益走势移动的幅度更大。Beta 的计算公式如下：

$$Beta = \beta_p \frac{Cov\left(D_p, D_m\right)}{Var(D_m)} \#(4.6)$$

其中，D_p=策略每日收益，D_m=基准每日收益，$Cov(D_p, D_m)$=策略每日收益与基准每日收益的协方差，$Var(D_m)$=基准每日收益的方差。

（6）夏普比率

夏普比率表示投资组合每承担一单位的风险，会给投资组合带来多少的超额收益，该指标同时将风险和收益都进行综合考量。

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p} \#(4.7)$$

其中，R_p=策略年化收益率，R_f=无风险利率，σ_p=策略年化波动率。

（7）胜率

胜率的计算公式为盈利的总资产除以总交易场次并乘以百分数，即：

$$\text{胜率} = \frac{\textit{盈利总次数}}{\textit{总交易次数}} * 100\% \#(4.8)$$

（8）盈亏比

盈亏比是在投资市场里每次交易的盈利和亏损的比例，计算公式为：

$$盈亏比 = \frac{盈亏的平均金额}{亏损的平均金额} * 100\% \#(4.9)$$

（9）信息比率

信息比率用来衡量每承担一单位超额风险，给投资组合带来的超额收益，经过计算，信息比率值越大，说明该投资组合进行跟踪误差时，获得更高的超额收益，也可以认为投资组合追求较大的信息比率，因此在设定承担适度风险的目标下，尽可能获得较高的信息比率

$$information\ Ratio = \frac{R_p - R_m}{\sigma_t} \#(4.10)$$

其中，R_p=策略年化收益率，R_m=基准年化收益率，σ_t策略与基准每日收益差值的年化标准差。

（10）策略日收益率标准差

$$Daily\ Volatility = \sigma = \sqrt{\frac{1}{n}\sum_i^n \left(R - \overline{R}\right)^2} \#(4.11)$$

其中，R=策略每日收益率，$\overline{R}$=策略每日收益率的平均值，n=策略执行天数。

（11）策略波动率

策略每日收益率的标准差的年化值，即年度波动率。用来测量策略的风险性，波动越大代表策略风险越高。

$$Algorithm\ Volatility = \sigma_p = \sqrt{\frac{252}{n}\sum_i^n \left(R_p - \overline{R}_p\right)^2} \#(4.12)$$

其中，R_p=策略每日收益率，$\overline{R}_p$=策略每日收益率的平均值，n=策略执行天数。

（12）基准波动率

基准每日收益率的标准差的年化值。用来测量基准的风险性，波动越大代表基准风险越高。

$$Benchmark\ Volatility = \sigma_m = \sqrt{\frac{252}{n}\sum_i^n \left(R_m - \overline{R}_m\right)^2} \#(4.13)$$

其中R_m=基准每日收益率，$\overline{R}_m$=基准每日收益率的平均值，n=策略执行天数。

（13）最大回撤

描述策略可能出现的最糟糕的情况，最极端可能的亏损情况

$$Max\ Drawdown = \frac{Max\left(P_x - P_y\right)}{P_x} \#(4.14)$$

其中P_x, P_y为策略某日股票和现金的总价值，并有$y > x$。

其中策略总收益、策略年化收益率指标为策略模型盈利性的衡量指标，也是最受关注的衡量指标，而 Alpha 阿尔法、Beta 贝塔、夏普比率、信息比率、策略波动率、最大回撤是衡量策略模型稳定性的指标，在本章之后对模型效果进行评价将主要集中在盈利性及稳定性。

4.2 基于线性回归模型的量化策略构建

基于线性回归的模型基本思想是：将前章 IC 分析法和随机森林算法筛选出来的因子作为线性回归的模型输入变量（解释变量），将未来 30 日的收益率作为输入变量（被解释变量）进行回归计算，建立线性回归模型，找到两者之间的关系，随后将这种线性关系用于对股票未来收益率的预测，从而构建一个股票投资组合。具体做法是：选取 2015-01-01 至 2020-12-31 的因子数据和经过计算的股票收益率数据作为样本内数据，训练出线性回归模型，将 2021-01-01 至 2021-11-30 的因子数据和经过计算的股票收益率数据作为样本外数据，将因子数据作为解释变量，预测未来 30 交易日的收益率，在将对于股票收益率排名靠前的股票进行买入持有操作构建投资组合，通过模拟回测得到相应的投资组合收益相关数据。基于线性回归模型构架量化投资策略模型的基本流程如下：

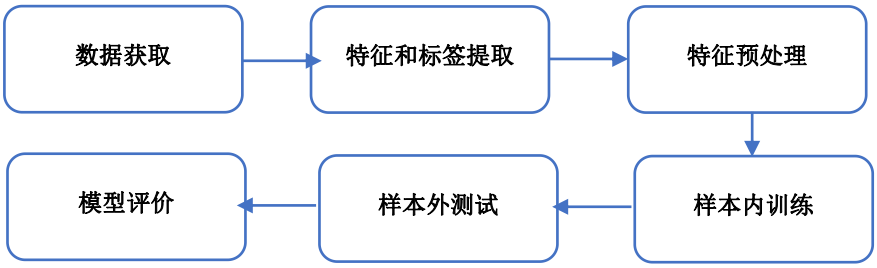


图 4.1 基于线性回归模型量化投资策略构建流程图

获取数据：获取中国 A 股市场中所有的股票收益率数据和因子值数据

提取特征和标签：计算 23 个预定的因子值作为特征变量或解释变量，计算股票未来 30 日的收益率作为标签或被解释变量

特征变量预处理：进行去极值、缺失值处理

策略回测步骤：利用 2015-01-01 至 2020-12-31 的数据作为样本内数据进行模型训练，2021-01-01 至 2021-11-30 的数据作为样本外测试数据用于预测股票收益率，每日买入预测值排名前 5 的股票，持有日至少为 30 日，具体而言，预测排名越靠前，分配到的资金越多且最大资金占用比例不超过 3%；初始 30 日平均分配资金，之后，尽量使用剩余资金。

基于 Bigquant 量化平台的回测结果如下：本次回测有两组结果，一组结果是将 IC 分析法筛选后的因子作为输入因子，进行回测，第二组是将随机森林机器学习算法筛选后的因子作为输入因子，进行回测。其中第一组回测结果如下：

图 4.2 回测期间策略收益率走势图

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	27.1%	31.61%	-7.28%	33%	31%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
1.58	0.52	1.41	16.34%	11%	9.67%

表 4.1 回测返回指标值

第一组结果为基于 IC 分析法筛选因子及线性回归模型预测的量化投资策略，

从策略收益率和基准收益率的走势图来看，在大部分时间跑赢了基准收益率，并且在回测期间获得27.1%的绝对收益，而基准收益率为-7.28%，超额收益率为33%，策略收益率换算年化收益率超过 30%，盈利性极佳。从其他指标上看，稳定性还不错，收益波动率仅为 16.34%，并且每承担一单位系统性风险获得 1.58 个单位的回报，风险回报控制极其佳。另外该策略的最大回撤仅为 9.67%，控制在 10%以内，对于大多数投资股票的投资者来说，10%以内的回撤是一个非常难达到的目标。但是该策略的胜率仅为 0.52，意味着进行交易只有一半的时间交易之后马上盈利。总的来说该策略投资组合是有效并且表现极佳。

最后一个交易日的持仓情况：

表 4.2 投资策略最后一个交易日持仓情况

日期	股票代码	股票名称	持仓均价	收盘价	股数	持仓价值	收益
2021/11/30	603040.SHA	新坐标	21.99	23.38	659	15407.421	916.143
2021/11/30	603031.SHA	安德利	23.83	46.93	100	4693	2310
2021/11/30	601086.SHA	国芳集团	3.96	3.57	1396	4983.72	-543.742
2021/11/30	601068.SHA	中铝国际	3.49	3.89	1300	5057	520
2021/11/30	600884.SHA	杉杉股份	16.686	37.43	903	33799.29	18732.013
2021/11/30	601011.SHA	宝泰隆	4.007	5.12	2529	12948.48	2815.788
2021/11/30	603157.SHA	*ST 拉夏	1.66	2.43	7600	18467.999	5849.999
2021/11/30	600976.SHA	健民集团	35.943	51.92	201	10435.92	3211.417
2021/11/30	600966.SHA	博汇纸业	17.598	10.15	201	2040.15	-1497.028

最后一个交易日的交易情况：

表 4.3 投资策略最后一个交易日交易情况

日期	时间	股票代码	股票名称	买/卖	数量	成交价	总成本	交易佣金
2021/11/30	9:30	002486.SZA	嘉麟杰	买入	2700	2.71	7317.001	5
	9:30	601890.SHA	亚星锚链	买入	800	9.16	7328	5
	9:30	300308.SZA	中际旭创	买入	200	37.89	7578	5
	9:30	002479.SZA	富春环保	买入	2200	5.52	12144	5
	9:30	300399.SZA	天利科技	买入	1600	11.44	18304.001	5.49
	15:00	601225.SHA	陕西煤业	卖出	900	11.88	-10692	13.9
	15:00	601228.SHA	广州港	卖出	6300	3.17	-19970.997	25.96
	15:00	002512.SZA	达华智能	卖出	2100	3.36	-7056	9.17
	15:00	002915.SZA	中欣氟材	卖出	200	33.68	-6736	8.76

	15:00	002901.SZA	大博医疗	卖出	100	47.24	-4724	6.14

按以上步骤，只将因子进行更改后得到第二组的回测结果如下：

图 4.3 投资策略回测期间收益率走势图

表 4.4 投资策略回测指标情况

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	25.83%	30.11%	-7.28%	31%	31%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
1.72	0.51	1.7	14.14%	12%	6.57%

第二组结果则是基于随机森林算法筛选因子及线性回归模型预测的量化投资策略，单独看该组策略，绝对收益率为 25.83%，超额收益率为 31%，远远超过基准收益率，并且有一个不错的稳定性，收益波动率为 14.14%，风险回报为 1.72。另外最大回撤仅为 6.57%，该策略组合拥有较强的风险控制能力。但该组合在胜率上也仅为 51%，不过也并不影响策略的盈利性。

最后一个交易日的持仓情况如下：

表 4.5 投资策略最后一个交易日的持仓情况

日期	股票代码	股票名称	持仓均价	收盘价	股数	持仓价值	收益
2021/11/30	600879.SHA	航天电子	8.173	8.37	1900	15903	375
2021/11/30	601038.SHA	一拖股份	10.889	12.3	1705	20971.499	2406.339
2021/11/30	600962.SHA	国投中鲁	7.54	8.97	500	4485	715
2021/11/30	601058.SHA	赛轮轮胎	9.382	13.91	3630	50493.299	16437.242

2021/11/30	603106.SHA	恒银科技	5.152	5.64	2356	13287.84	1149.492
2021/11/30	600979.SHA	广安爱众	2.883	3.28	1423	4667.44	565.5
2021/11/30	601231.SHA	环旭电子	17.831	15.38	206	3168.28	-504.844
2021/11/30	600520.SHA	文一科技	6.3	9.02	600	5412	1632

最后一个交易日的交易情况如下：

表 4.6 投资策略最后一个交易日的交易情况

日期	时间	股票代码	股票名称	买/卖	数量	成交价	总成本	交易佣金
2021/11/30	9:30	002831.SZA	裕同科技	买入	200	31.6	6320	5
	9:30	002832.SZA	比音勒芬	买入	300	23.96	7188	5
	9:30	002837.SZA	英维克	买入	200	47.42	9484	5
	9:30	002897.SZA	意华股份	买入	200	46.48	9296	5
	15:00	002340.SZA	格林美	卖出	1800	11.3	-20340.002	26.44
	15:00	002561.SZA	徐家汇	卖出	1900	6.91	-13128.994	17.07
	15:00	002512.SZA	达华智能	卖出	2300	3.36	-7728	10.05
	15:00	002918.SZA	蒙娜丽莎	卖出	300	23.28	-6984	9.08

总结来讲，两组策略组合的结果有一定相似性，并且都实现了较好的策略收益，同时在波动性以及风险控制上都表现很强，只是胜率有所欠缺，只有一半时间的交易在盈利，但是作为一个中长期持有的策略组合，从结果上也可以看出并不影响策略组合的盈利性，因此，通过 IC 分析法和随机森林算法筛选出的因子，通过线性回归模型构建量化投资策略，是有效且值得向实践中推广的。

4.3 基于 LSTM 模型的量化策略构建

基于 LSTM 的模型基本思想是：将前章 IC 分析法和随机森林算法筛选出来的因子作为 LSTM 的模型输入特征，将未来 30 日的收益率作为标签进行计算，建立 LSTM 模型，找到两者之间的关系，随后将这种关系用于对股票未来收益率的预测，从而构建一个股票投资组合。具体做法是：选取 2015-01-01 至 2020-12-31 的因子数据和经过计算的股票收益率数据作为样本内数据，基于 LSTM 模型进行训练，将 2021-01-01 至 2021-11-30 的因子数据和经过计算的股票收益率数据作为样本外数据，将因子数据作为输入特征，预测未来 30 交易日的收益率，在将对于股票收益率排名靠前的股票进行买入持有操作构建投资组合，通过模拟回测得到相应的投资组合收益相关数据。基于 LSTM 模型构建量化投资策略模型的基本流程如下：

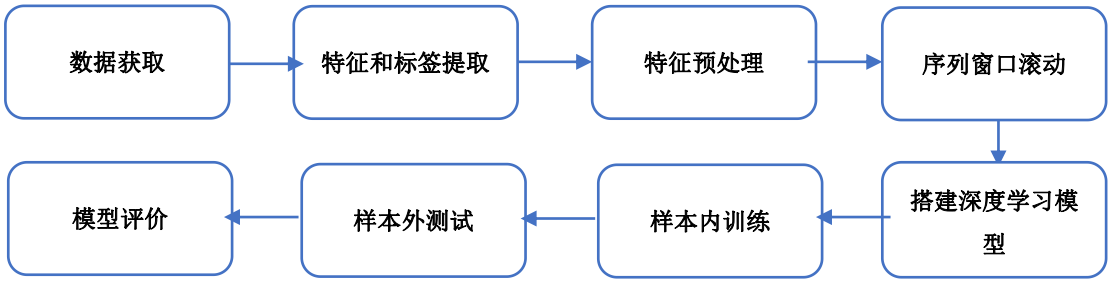


图 4.4 基于 LSTM 神经网络模型量化投资策略构建流程图

获取数据：获取中国 A 股市场中所有的股票收益率数据和因子值数据

提取特征和标签：计算 23 个预定的因子值作为特征变量或解释变量，计算股票未来 30 日的收益率作为标签或被解释变量

特征预处理：先进行缺失值处理，去除存在特征异常值的股票，再进行标准化处理，去除存在较大数量级特征值的影响

序列窗口滚动：窗口大小设置为 5，滚动切割。使用过去 5 天的因子数据作为输入。

搭建 LSTM 模型：构建两个隐含层的 LSTM 长短期记忆神经网络预测股票收益率（回归模型）。在可视化策略中表现为 1 个输入层；一个 LSTM 和一个全连接层作为隐藏层，每构建一层进行 dropout 断开一些神经元防止过拟合；最后一个全连接层作为输出层（输出维度调整为 1）。

模型训练与预测：使用 LSTM 模型进行训练和预测；策略使用 tanh 激活函数

策略回测：利用 2015-01-01 至 2020-12-31 的数据作为样本内数据进行模型训练，2021-01-01 至 2021-11-30 的数据作为样本外测试数据用于预测股票收益率，每日买入预测值排名前 5 的股票，持有日至少为 30 日，具体而言，预测排名越靠前，分配到的资金越多且最大资金占用比例不超过 3%；初始 30 日平均分配资金，之后，尽量使用剩余资金。

模型评价 ：查看模型回测结果。

第一组结果如下

图 4.5 投资策略回测期间收益率走势图

表 4.7 投资策略回测指标情况

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	55.32%	65.59%	-7.28%	67%	19%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
2.87	0.64	2.87	18.55%	16%	10.16%

第一组结果是基于 IC 分析法筛选因子及 LSTM 预测模型构建量化投资策略的回测结果，收益指标上看，收益率为 55.32%，超额收益率也达到了 67%，该策略在回测期间实现了一个非常不错的收益。从稳定性及风险控制上讲，收益波动率为 18.55%，风险回报比为 2.87，最大回撤为 10.16%，收益较为稳定，风险控制能力水平也极佳。

最后一个交易日持仓情况

表 4.8 投资策略最后一个交易日持仓情况

日期	股票代码	股票名称	持仓均价	收盘价	股数	持仓价值	收益
2021/11/30	300312.SZA	邦讯技术	13.775	8.853	560.902	4965.701	-2760.455
2021/11/30	002766.SZA	ST 索菱	8.429	11.857	89.862	1065.47	308.044
2021/11/30	300022.SZA	吉峰科技	17.211	17.673	394.823	6977.839	182.593
2021/11/30	300325.SZA	ST 德威	14.3	25.894	2175.121	56322.859	25217.825
2021/11/30	300278.SZA	华昌达	11.927	18.825	1597.742	30077.5	11021.639
2021/11/30	000709.SZA	河钢股份	33.048	24.598	996.639	24515.25	-8421.548
2021/11/30	002427.SZA	ST 尤夫	12.476	13.238	550.204	7283.558	419.05
2021/11/30	300612.SZA	宣亚国际	38.905	31.627	44.632	1411.589	-324.807

2021/11/30	000856.SZA	冀东装备	13.559	12.883	87.474	1126.894	-59.137
2021/11/30	002494.SZA	华斯股份	9.696	10.725	239.135	2564.641	246.026

表 4.9 投资策略最后一个交易日交易情况

日期	时间	股票代码	股票名称	买/卖	数量	成交价	总成本	交易佣金
2021/11/30	9:30	002473.SZA	*ST 圣莱	买入	72.623	23.386	1698.324	5
	9:30	600105.SHA	永鼎股份	买入	26.294	67.26	1768.527	5
	9:30	600336.SHA	澳柯玛	买入	62.292	29.491	1837.06	5
	9:30	600522.SHA	中天科技	买入	13.182	158.214	2085.565	5
	9:30	300006.SZA	莱美药业	买入	61.872	45.023	2785.629	5
	15:00	002785.SZA	万里石	卖出	1396.431	30.253	-42245.633	54.92
	15:00	300746.SZA	汉嘉设计	卖出	327.866	16.988	-5569.92	7.24
	15:00	600841.SHA	上柴股份	卖出	52.375	40.058	-2098.009	5
	15:00	300338.SZA	开元教育	卖出	238.895	30.245	-7225.372	9.39
	15:00	300109.SZA	新开源	卖出	28.518	96.038	-2738.811	5

第二组情况：

图 4.6 投资策略回测期间收益率走势图

表 4.10 投资策略回测指标情况

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	31.8%	37.2%	-7.28%	36%	14%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
2.03	0.59	1.92	14.63%	11%	10.12%

第二组结果是基于随机森林机器学习算法筛选因子及 LSTM 预测模型构建量化投资策略的回测结果，从盈利指标上看获得了 31.8%的收益率，超额收益率

达到了 36%，稳定性及风险控制方面，收益波动率为 14.63%，风险回报比为 2.03，最大回撤也刚刚超过 10%，从指标上看是一个不错的策略。

但是相较于上一组结果无论是盈利水平还是稳定性都有所下降，对于 LSTM 神经网络模型来讲，不同的输入因子策略效果呈现出较大差异，从这两组回测结果上看，基于 IC 分析法筛选后的因子能更好的适用于 LSTM 神经网络模型的量化投资策略构建，综合来讲基于 LSTM 神经网络模型的量化投资策略模型构建取得了量化的结果，是有效且值得使用的一个模型。

表 4.11 投资策略最后一个交易日持仓情况

日期	股票代码	股票名称	持仓均价	收盘价	股数	持仓价值	收益
2021/11/30	002356.SZA	*ST 赫美	7.22	25.029	3756.633	94025.832	66902.69
2021/11/30	002071.SZA	*ST 长城	1.072	0.476	16552.045	7884.121	-9867.127
2021/11/30	000673.SZA	*ST 当代	17.305	20.138	517.133	10414.19	1464.965
2021/11/30	002316.SZA	亚联发展	11.252	13.779	984.039	13559.294	2486.696
2021/11/30	600311.SHA	*ST 荣华	6.42	8.683	510.533	4433.137	1155.443
2021/11/30	600146.SHA	*ST 环球	2.511	3.036	15075.658	45769.698	7907.339
2021/11/30	002633.SZA	申科股份	9.699	11.926	168.227	2006.27	374.571
2021/11/30	000502.SZA	绿景控股	19.071	19.272	1758.052	33882.052	355.105
2021/11/30	002072.SZA	*ST 凯瑞	4.967	7.736	3489.269	26994.403	9664.929
2021/11/30	002188.SZA	*ST 巴士	5.432	6.349	2095.335	13303.95	1922.839

表 4.12 投资策略最后一个交易日交易情况

日期	时间	股票代码	股票名称	买/卖	数量	成交价	总成本	交易佣金
2021/6/30	09:30	600297.SHA	广汇汽车	买入	81.605	18.879	1540.642	5
	09:30	600973.SHA	宝胜股份	买入	42.949	36.394	1563.083	5
	09:30	300483.SZA	首华燃气	买入	59.782	26.209	1566.822	5
	09:30	002092.SZA	中泰化学	买入	46.687	35.154	1641.207	5
	09:30	601717.SHA	郑煤机	买入	65.635	25.065	1645.156	5
	15:00	000889.SZA	中嘉博创	卖出	529.597	20.707	-10966.336	14.26
	15:00	600738.SHA	丽尚国潮	卖出	33.739	45.601	-1538.545	5
	15:00	000937.SZA	冀中能源	卖出	50.107	30.379	-1522.203	5
	15:00	600939.SHA	重庆建工	卖出	1319.302	3.588	-4733.563	6.15
	15:00	603113.SHA	金能科技	卖出	438.864	19.224	-8436.702	10.97

4.4 基于 LightGBM 模型的量化策略构建

基于 LightGBM 的模型基本思想是：将前章 IC 分析法和随机森林算法筛选出来的因子作为 LightGBM 的模型输入特征，将未来 30 日的收益率作为标签进

行计算，建立 LightGBM 模型，找到两者之间的关系，随后将这种关系用于对股票未来收益率的预测，从而构建一个股票投资组合。具体做法是：选取 2015-01-01 至 2020-12-31 的因子数据和经过计算的股票收益率数据作为样本内数据，基于 LightGBM 模型进行训练，将 2021-01-01 至 2021-11-30 的因子数据和经过计算的股票收益率数据作为样本外数据，将因子数据作为输入特征，预测未来 30 交易日的收益率，在将对于股票收益率排名靠前的股票进行买入持有操作构建投资组合，通过模拟回测得到相应的投资组合收益相关数据。基于 LightGBM 模型构架量化投资策略模型的基本流程如下：

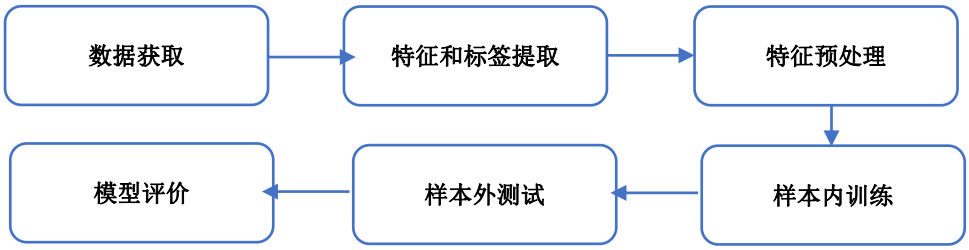


图 4.7 基于 LightGBM 模型量化投资策略构建流程图

获取数据：获取中国 A 股市场中所有的股票收益率数据和因子值数据

提取特征和标签：计算 23 个预定的因子值作为特征变量或解释变量，计算股票未来 30 日的收益率作为标签或被解释变量

特征变量预处理：缺失值处理

模型训练与预测：使用 LightGBM 模型进行训练和预测。相关参数情况如下：

表 4.13 LightGBM 模型输入参数

迭代次数	损失函数类型	树最大叶子树	学习率	每个叶子最少样本量	特征值离散化数量
30	排序	30	0.1	200	255

策略回测：利用 2015-01-01 至 2020-12-31 的数据进行训练，预测 2021-01-01 至 2021-11-30 期间的股票表现。每日买入预测排名最靠前的 5 只股票，至少持有 30 日，同时淘汰排名靠后的股票。具体而言，预测排名越靠前，分配到的资金越多且最大资金占用比例不超过 3%；初始 30 日平均分配资金，之后，尽量使用剩余资金。

模型评价 ：查看模型回测结果，

基于 Bigquant 量化平台的回测结果如下：本次回测有两组结果，一组结果

是将 IC 分析法筛选后的因子作为输入因子，进行回测，第二组是将随机森林机器学习算法筛选后的因子作为输入因子，进行回测。其中第一组回测结果如下：

图 4.8 投资策略回测期间收益率走势图

表 4.14 投资策略回测期间指标情况

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	23.25%	27.06%	-7.28%	27%	20%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
1.69	0.71	2.0	12.97%	10%	6.33%

第一组结果是基于 IC 分析法筛选因子及 LightGBM 预测模型构建的量化投资策略，从盈利性上看，表现较佳，回测期间内获得了 23.25%的收益率，超额收益率也达到 27%，相对于负的基准收益率，该策略的盈利性表现也是很强的。稳定性上讲，收益波动率为 12.97，风险回报比为 1.69，也仅为 6.33%，稳定性及风险控制能力处于非常不错的水平，胜率也到达了 71%，该策略总的来讲各项指标较为均衡，并未有明显的短板，也并无超乎寻常的部分，随后在进行个策略对比是在详细讨论相对效果。

最后一个交易日的持仓情况：

表 4.15 投资策略最后一个交易日持仓情况

日期	股票代码	股票名称	持仓均价	收盘价	股数	持仓价值	收益
2021/11/30	002476.SZA	宝莫股份	18.292	25.5	81.255	2072.034	585.721
2021/11/30	300038.SZA	ST 数知	10.491	8.414	614.547	5170.907	-1276.38
2021/11/30	000005.SZA	世纪星源	17.222	20.296	94.685	1921.732	291.072

2021/11/30	600192.SHA	长城电工	6.92	8.423	1163.724	9801.909	1749.228
2021/11/30	300325.SZA	ST 德威	15.696	25.894	152.779	3956.079	1558.053
2021/11/30	002289.SZA	ST 宇顺	9.199	11.677	3076.36	35922.902	7624.709
2021/11/30	300306.SZA	远方信息	34.961	44.158	200.481	8852.778	1843.754
2021/11/30	600721.SHA	ST 百花	4.386	13.055	262.022	3420.697	2271.584
2021/11/30	002381.SZA	双箭股份	54.487	44.733	45.012	2013.51	-439.056

最后一个交易日的交易情况如下：

表 4.16 投资策略最后一个交易日交易情况

日期	时间	股票代码	股票名称	买/卖	数量	成交价	总成本	交易佣金
2021/11/26	9:30	300059.SZA	东方财富	买入	2.195	1944.008	4267.295	5
	15:00	603178.SHA	圣龙股份	卖出	579.851	17.534	-10167.229	13.22

第二组的回测结果如下：

图 4.9 投资策略回测期间收益率走势图

表 4.17 投资策略回测期间指标情况

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	38.42%	45.13%	-7.28%	45%	23%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
2.64	0.7	5.11	13.32%	15%	7.58%

第二组回测结果是基于随机森林筛选因子及 LightGBM 预测模型构建的量化投资策略，收益率 38.42%，超额收益率 45%，远远超过基准，同时收益波动率为 13.32%，风险回报比为 2.64，最大回撤也仅为 7.58%，该策略无论从盈利

性还是稳定性及风险控制水平都较为出众，是一个优秀的策略组合。

相比第一组结果，第二组结果收益率、收益波动率、夏普比率等指标上面有所提升，最大回撤相对于第一组结果也有所增大，但幅度不大，总体上对比第二组结果从盈利性及稳定性有一个不错的提升，若究其原因，因为随机森林机器学习算法和 LightGBM 都属于机器学习树模型的一种，原理相似，筛选出来的因子更适用于用 LightGBM 构建量化投资策略模型。LightGBM 模型同样可以输出因子重要性，因此我们将利用第一组的训练结果输出因子重要性排序，将排名前50%的因子进行再筛选，随后用这些因子作为输入特征因子，利用 LightGBM 模型构建量化投资策略模型，在相同的条件下，观察策略效果是否有所提升。

最后一个交易日的持仓情况：

表 4.18 投资策略最后一个交易日持仓情况

日期	股票代码	股票名称	持仓均价	收盘价	股数	持仓价值	收益
2021/11/30	603316.SHA	诚邦股份	6.675	6.946	1398.487	9713.194	378.623
2021/11/30	000609.SZA	中迪投资	22.684	31.411	401.74	12619.2	3506.211
2021/11/30	600615.SHA	丰华股份	309.701	421.938	28.518	12032.826	3200.801
2021/11/30	002027.SZA	分众传媒	132.231	95.425	333.137	31789.597	-12261.415
2021/11/30	002199.SZA	东晶电子	15.734	23.19	439.787	10198.626	3279.154
2021/11/30	002209.SZA	达意隆	9.453	9.81	396.361	3888.204	141.21
2021/11/30	300056.SZA	中创环保	29.481	33.259	401.485	13352.902	1516.798
2021/11/30	300345.SZA	华民股份	14.961	22.317	232.048	5178.544	1706.868
2021/11/30	603089.SHA	正裕工业	16.432	20.73	122.042	2529.878	524.526

最后一个交易日的交易情况如下：

日期	时间	股票代码	股票名称	买/卖	数量	成交价	总成本	交易佣金
2021/11/30	9:30	300023.SZA	*ST 宝德	买入	68.712	23.137	1589.78	5
	9:30	000534.SZA	万泽股份	买入	27.173	70.504	1915.804	5
	9:30	000688.SZA	国城矿业	买入	31.435	66.219	2081.614	5
	15:00	300165.SZA	天瑞仪器	卖出	50.818	44.295	-2250.972	5

表 4.19 投资策略最后一个交易日交易情况

因 LightGBM 也属于一类树模型，因此我们从第一组回测结果可以得到输入因子的重要性，随后我们将因子重要性排序结果取排名前 50%的因子，作为输入因子，在相同环境下进行投资策略构建及回测，将第一组的两次回测结果进行比

较，观察是否有较好的改进。

部分因子排序结果如下，详情见附录：

表 4.20 基于 LightGBM 模型的因子重要性排序

log(market_cap_0)	144
fs_operating_revenue_yoy_0	70
swing_volatility_30_0	70
fs_net_cash_flow_ttm_0/market_cap_0	61
pb_lf_0	58
avg_turn_5	56
ta_rsi_28_0	55
volatility_30_0	53
ta_macd_macd_12_26_9_0	48
ta_trix_14_0	42
pe_ttm_0	40

我们选取排名前 50%的因子重新作为策略模型的输入因子特征，在相同的环境下进行策略构建及回测，得到以下结果：

图 4.10 投资策略回测期间收益率走势图

表 4.21 投资策略回测期间指标情况

指标	收益率	年化收益率	基准收益率	阿尔法	贝塔
	45.99%	54.25%	-7.28%	54%	19%
夏普比率	**胜率**	**盈亏比**	**收益波动率**	**信息比率**	**最大回撤**
2.77	0.59	3.21	14.97%	15%	9.04%

跟第一组第一次的回测结果对比发现，收益性方面有了很大的提升，回测期间收益接近第一次回测收益的两倍，风险回报比也有所提升。另外，虽然收益波动率以及最大回撤也在增加，但我们能清除的看到，相比增长快一倍的收益率，增加的幅度很小，因子对第一组结果进行以上改进，能使投资策略模型获得较好的改进。

4.5 量化策略效果评价及比较

上一节，以 IC 分析法、随机森林机器学习算法筛选因子，分别利用线性回归模型、LSTM 神经网络模型、LightGBM 机器学习模型进行量化投资策略构建，并且基于相同的条件进行回测，得到了六组结果和一组改进后的策略回测结果，现将七组回测结果相关指标整理所得如下：

表 4.22 投资策略回测指标情况汇总

	IC 线性模型	随机森林线性模型	ICLSTM 模型	随机森林 LSTM 模型
指标				
收益率	27.10%	25.83%	55.32%	31.80%
年化收益率	31.61%	30.11%	65.59%	37.20%
基准收益率	-7.28%	-7.28%	-7.28%	-7.28%
阿尔法	33%	31%	67%	36%
贝塔	31%	31%	19%	14%
夏普比率	1.58	1.72	2.87	2.03
胜率	0.52	0.51	0.64	0.59
盈亏比	1.41	1.7	2.87	1.92
收益波动率	16.34%	14.14%	18.55%	14.63%
信息比率	11%	12%	16%	11%
最大回撤	9.67%	6.57%	10.16%	10.12%

续表 4.22

	ICLightGBM 模型	随机森林 LightGBM	改进 ICLightGBM 模型
指标			
收益率	23.25%	38.42%	45.99%
年化收益率	27.06%	45.13%	54.25%
基准收益率	-7.28%	-7.28%	-7.28%
阿尔法	27%	45%	54%
贝塔	20%	23%	19%

夏普比率	1.69	2.64	2.77
胜率	0.71	0.7	0.59
盈亏比	2	5.11	3.21
收益波动率	12.97	13.32%	14.97%
信息比率	10%	15%	15%
最大回撤	6.33%	7.58%	9.04%

从表上对比发现，七组量化投资策略都实现了超过 20%的收益率，而回测期间的基准收益率仅为-7.28%。从结果对比上讲，将 IC 分析法作为因子筛选的方法，筛选后的因子作为量化投资策略模型的输入因子，基于 LSTM 的量化投资策略回测收益超过了基于线性回归模型的量化投资策略回测收益，而基于 LightGBM 模型的量化投资策略回测收益并未超过基于线性模型的量化投资策略回测收益，但对该 LightGBM 模型构建的量化投资策略模型进行改进后，获得了超过基于线性回归模型的量化投资策略回测收益；将随机森林方法作为因子筛选的方法，筛选后的因子作为量化投资策略模型的输入因子，基于 LSTM 和 LightGBM 两种机器学习模型的量化投资策略回测收益均超过了基于线性回归模型构建的量化投资策略回测收益。并且在盈利水平上，利用 IC 分析法筛选的因子作为输入特征因子，利用 LSTM 神经网络模型构建的量化投资策略表现出来的收益水平在七组结果中表现最佳。

从其他指标上看，风险回报比、盈亏比等指标基于机器学习的量化投资策略模型较线性回归模型构建的量化投资策略都有所提升，贝塔值也有所下降，表明暴露在市场组合的风险水平有所下降。收益波动率并未出现较大变化，基于 LightGBM 中模型的量化投资策略，最大回撤有所下降，但基于 LSTM 模型的量化投资策略虽然实现了七组结果中较强的收益水平，最大回撤水平也是三组模型中较大的一组策略模型。

4.6 本章小结

本章首先简单介绍了模型结果的回测指标，随后分别基于线性回归模型、LSTM 神经网络模型、LightGBM 机器学习模型构建量化投资策略并进行回测，随后先分别对各组结果进行评价，随后将各组回测指标进行对比分析并进行讨论。

5.结论与展望

5.1 本文主要结论

随着股票市场的增大，对投资者来说选择未来收益率不错的股票越来越困难，并且常常存在“7 亏 2 平 1 负”的魔咒，所以本文综合考虑多个维度的因子，建立因子库，通过 IC 分析法和随机森林机器学习算法对因子库进行筛选，随后分别基于线性回归模型、LSTM 神经网络模型、LightGBM 机器学习模型进行量化投资策略构建，本文通过实证研究得出的主要结论如下：

（1）分别基于随机森林机器学习算法以及基于 IC 分析法对因子库进行筛选，IC 分析法和随机森林机器学习算法都显现出因子筛选的效能。最后筛选出包括行情类因子及技术类因子等的 23 个因子数据，作为量化投资策略模型构建的输入因子特征。将两种方法筛选后的因子对比发现：存在较大的差异，有几小类因子发生了完全的变化。

（2）将 IC 分析法和随机森林机器学习算法筛选后的因子作为输入因子特征，并且基于三种模型进行构建量化投资策略，总共输出七组回测结果，分别来看：回测期间内，七组结果都有较好的盈利水平，并且在稳定性和风险控制上面都呈现出不错的水平，总体来讲呈现的是不错的量化投资策略,可以为投资者提供参考

（3）将基于线性回归模型、LSTM 神经网络模型、LightGBM 机器学习模型构建的量化投资策略回测结果对比发现：回测期间内，基于线性回归、LSTM 模型构建的量化投资策略，将 IC 分析法作为因子筛选方法，能够获得获得更好的收益；基于 LightGBM 模型构建的量化投资策略，将随机森林作为因子筛选方法，能够获得更好的收益。

（4）回测期间内，将 IC 分析法作为因子筛选方法，仅 LSTM 模型构建的量化投资策略收益与改进后的 LightGBM 模型构建的量化投资策略回测收益优于线性回归模型构建的量化投资策略回测收益；将随机森林方法作为因子筛选方法，LSTM 和 LightGBM 机器学习模型构建的量化投资策略回测收益均优于线性回归模型构建的量化投资策略回测收益；同时，与线性回归模型构建的量化策略

回测结果对比发现，LSTM 和 LightGBM 机器学习算法构建的量化投资策略回测结果稳定性、风险控制等指标也都有所提升。

（5）回测期间内，将 IC 分析法作为因子筛选方法，基于 LSTM 模型的量化投资策略收益水平最佳。但同时最大回撤水平也最大，但是绝对幅度并不大，略超过 10%，在几组策略结果里面是较好的一组策略。在承担一定的最大回测水平的情况下，推荐选择 LSTM 神经网络模型构建量化投资策略，为投资决策提供参考。

5.2 建议与启示

5.2.1 对我国完善资本市场的建议

1、壮大机构投资者力量，推动资本市场发展

目前我国资本市场还不是强有效资本市场的原因之一是国内的资本市场投资者个人投资者数量庞大，市场上充斥的非理性的行为，致使资产的价格不能很好的反应出该资产所有可以获得的信息。市场有效的一个重要意义是杜绝内部交易，有效的保护投资者的利益。

非自然投资者以机构投资者为主，是由一群拥有专业知识专业投资人组成，背后有成熟的调研团队，因此在进行投资决策是都比个人投资者更加理性专业，我国近年也正在引导去散户化，走向机构化，推动形成一个机构投资者主导的有效的资本市场，因此如若借助机器学习算法在量化投资上的应用，进一步提高专业投资机构的投资水平，引导散户投资者转向聘请专业的机构进行投资，将能够有效的壮大机构投资者的力量，推动资本市场朝更加有效的方向发展。

2、完善信息披露制度，保护中小投资者

一方面，资本市场作为一个信息市场，如何保护好和服务好中小投资者，首先要做到的是能够及时、准确的让中小投资者获得真实信息，并且能够基于这些真实的信息做出自主判断；另一方面，资本市场上公布的上市公司财务数据信息，是量化研究的基础，提高信息披露的质量，为量化研究提供良好的基础。我国前些年频发财务造假事件，给许多投资者带来损失，需要进一步完善信息披露制度，提升透明度，规划信息披露规则体系，督促上市公司及股东等信息披露义务人能

够及时、准确、真实、完整的进行信息披露，同时，对信息披露编报等规则进一步优化，提升上市公司披露的财务信息的质量。为开展量化研究提供高质量的数据来源，帮助中小投资者获得超额收益，保障中小投资者利益，进一步推动资本市场健康发展。

3、加大金融科技人才培养，鼓励金融科技创新

在产业革命和科技革命的背景下，金融科技发展迅速，金融业务与人工智能、大数据、物联网等信息技术深度融合，不断为金融发展提供源源不断的创新活力。目前，我国在金融科技发展方面，走在世界的前列，但国内的发展主要集中在金融科技应用方面，一些底层技术核心还被国外科技巨头所掌握，如何突破底层技术壁垒，亟需加强金融科技人才的培养。国内具备金融和科技素养的人才来源有限，专业人才体系也面临着不完善的问题，加之其他领域对金融科技人才的争夺，金融科技人才的短缺是中国正在面临的困境，人才市场上金融科技人才的缺口进一步扩大。对于金融科技企业和传统的金融机构，不仅仅要求人才具备专业的技术，还要有金融思维。具备金融和科技素养的复合型人才是推动金融科技发展的主力军。

高校在我国教育体系中是不可或缺的一部分，是为社会培养人才的重要渠道，高校在培养金融科技复合型人才方面作用不可或缺，高校应该积极开展金融科技人才的培养实践，设计科学的课程培养体系，鼓励金融科技创新，为金融科技发展提供源源不断的人才。

5.2.2 对投资者的启示

1、科学构建投资组合，合理分散风险

构建投资组合是现代金融学理论部分的重要内容，组合投资旨在收益和风险之间找到一个平衡点，即实现在收益一定的条件下尽可能的降低风险，或是在风险一定是寻求收益的最大化。资产组合理论已经证明，随着组合中证券数量的增加，证券组合的风险会降低，相关性较低的资产减组成的多元化资产组合可以有效的降低风险。虽然根据资产组合的理论，理性的投资者能够在收益和风险之间找到最佳的平衡点，这当中涉及到对大量金融数据的处理以及计算，实际上普通投资者很难做到构建收益和风险平衡的投资组合。

基于机器学习算法的人工智能技术可以进行复杂的技术面、基本面分析，以及文本分析对金融投资组合中的资产配置进行优化。在对传统投资组合优化方法实际操作中，会遇到很多挑战，相比于传统的方法人工智能机器学习算法通常能够提供更好的收益和协方差估计。而这些估计可以直接应用到资产配置的决策过程中，从而构建出较传统方法构建的投资组合更接近绩效目标的投资组合。大量基于人工智能机器学习的量化研究，将会很有效的帮助投资者科学构建投资组合，合理风险投资风险。

2、为投资者提供科学决策工具，提高投资者投资决策能力

在我国股票市场，有着“七亏二平一赚”的魔咒，大意是投资者中百分之70的投资者是亏损的，百分之20的投资者处于盈亏平衡的状态，仅仅还有百分之10的投资者投资股票是盈利的，因此我们可以看出，在我国以散户投资者为主的市场，市场上充斥的许多不理性的投资行为，追涨杀跌，无法进行系统科学的决策。

一方面如若机器学习算法在量化投资中的应用成熟，开发出辅助投资者进行投资决策的应用场景，使个人投资者也像机构投资一样进行专业的投资决策，对个人投资者打破“七亏二平一赚”的魔咒会有很大的帮助；另外一方面，投资者了解到机器学习算法在投资中的应用，投资者应该意识到可能在和机器进行交易，在算法越来越接近于人的大脑的时代，普通人战胜算法，也几乎不可能的，认识到自己在投资中的不足，可以转而投向寻求专业投资机构的帮助也将是一个不错的选择。

5.3 研究展望

本文通过实证分析证明了基于IC分析法及随机森林机器算法对候选因子进行筛选，随后再基于线性回归模型、LSTM神经网络模型及LightGBM机器学习模型构建的量化投资策略可以获得较高的超额收益，但仍然有很多工作待进一步开展。

（1）因子筛选是本文量化投资策略构建的基础，本文在候选因子库建立时还有许多因子未包含进来，因此可以通过增加许多相关因子，从而达到优化因子库的目的。

（2）本文采用固定长度的静态训练期进行训练和测试，有一定的局限性和复杂性，接下来的研究可以尝试进行超参数的动态寻优以及滚动预测。

参考文献

[1]Chen T , Guestrin C. XGBoost: A Scalable Tree Boosting System[C]//Proceedings of the 221d1 ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016: 785-94 100%.S

[2]Eugene F. Fama,Kenneth R. French. A five-factor asset pricing model[J]. Journal of Financial Economics,2015,116(1).

[3]Fama Eugene F.,French Kenneth R.. Common risk factors in the returns on stocks and bonds[J]. North-Holland,1993,33(1).

[4]Fama E F. Efficient market hypothesis: A Review of Theory and Empirical Work[J]. Journal of Finance, 1970, 25(2).

[5]G.H. Lawson,Richard Pike. Capital Asset Prices: Risk and Return[J]. Managerial Finance,1979,5(1).

[6]Gupta Deepak,Pratama Mahardhika,Ma Zhenyuan,Li Jun,Prasad Mukesh. Financial time series forecasting using twin support vector regression.[J]. PloS one,2019,14(3).

[7]Kim K. Financial time series forecasting using support vector machines[J].

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

[9]Manoj1ovic T, Stajduhar I. Predicting stock market trends using random forests; sample of the Zagreb stock exchange[C] 113841 Internationa1 Convention on Information & Communication Technology, Electronics & Micro electronics. IEEE, 2015.100%.

[10]Manolis Maragoudakis, Dimitrios Serpanos . Artificial Intelligence Applicating and Innovations[M]. Springer Berlin Heidelberg 2010.

[11]Piotr Ladyzynski, Kamil Žbikowski ,Przemyslaw Grzegorzewski. Stock Trading with Random Forests,Trend Detection Tests and Force Index Vo1umeIndicators[C]// 12t1 International c 100%ce on Artificial Intelligence and Soft Computing,ICAISC 2013.

[12]Ross Stephen A. The arbitrage theory of capital asset pricing[J]. Academic Press,1976,13(3).

[13]Sharpe, W.F. Capital Asset Prices: A theory of market of market equilibrium under conditions of risk[J]. Journal of Finance, 1964(19): 425-442.

[14]Stambaugh Robert F.,Yuan Yu. Mispricing Factors[J]. Narnia,2017,30(4).

[15]Thomas Fischer,Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research,2018,270(2).

[16]Wenbin Chen,Kun Fu,Jiawei Zuo,Xinwei Zheng,Tinglei Huang,Wenjuan Ren. Radar emitter classification for large data set based on weighted-XGBoost[J]. IET Radar, Sonar & Navigation,2017,11(8).

[17]Xiong R , Nichols E P , Shen Y . Deep Learning Stock Volatility with Google Domestic Trends[J]. Papers, 2016.

[18]Yihua Zhong , Lan Luo , Xinyi Wang , Jinlian Yang. Multi-factor Stock Selection Model Based on Machine Learning[J]. Engineering Letters,2021,29(1).

Neurocomputing, 2003, 55 (1): 307-319.

[19]Yutong S, Zhao H. Stock selection model based on advanced AdaBoost algorithm[C]//International Conference on Modelling. IEEE, 2015.

[20]陈浪南,屈文洲.资本资产定价模型的实证研究[J].经济研究,2000(04):26-34.

[21]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(02):21-27.

[22]邓凤欣,王洪良.LSTM 神经网络在股票价格趋势预测中的应用——基于美港股票市场个股数据的研究[J].金融经济,2018(14):96-98.

[23]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(02):61-65.

[24]李泽远.可超越评分卡模型么？基于 LightGBM 与卷积神经网络在贷款违约风险预测的研究[J].特区经济,2021(05):67-69.

[25]范龙振,余世典.中国股票市场的三因子模型[J].系统工程学报,2002(06):537-546.

[26]冯宇旭,李裕梅.基于 LSTM 神经网络的沪深 300 指数预测模型研究[J].数学的实践与认识,2019,49(07):308-315.

[27]葛橹漠,周显.基于 XGBoost 的多因子选股模型[J].信息技术与标准化,2020(05):36-41.

[28]黄卿,谢合亮.机器学习方法在股指期货预测中的应用研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析[J].数学的实践与认识,2018,48(08):297-307.

[29]靳云汇,刘霖.中国股票市场 CAPM 的实证研究[J].金融研究,2001(07):106-115.

[30]贾秀娟.基于随机森林的支持向量机量化选股[J].区域金融研究,2019(01):27-30.

[31]李志冰,杨光艺,冯永昌,景亮.Fama-French 五因子模型在中国股票市场的实证检验[J].金融研究,2017(06):191-206.

[32]马晓君,沙靖岚,牛雪琪.基于 LightGBM 算法的 P2P 项目信用评级模型的设计及应用[J].数量经济技术经济研究,2018,35(05):144-160.

[33]欧阳志刚,李飞.四因子资产定价模型在中国股市的适用性研究[J].金融经济学研究,2016,31(02):84-96.

[34]欧阳红兵,黄亢,闫洪举.基于 LSTM 神经网络的金融时间序列预测[J].中国管理科学,2020,28(04):27-35.

[35]裴大卫,朱明.基于多因子与多变量长短期记忆网络的股票价格预测[J].计算机系统应用,2019,28(08):30-38.

[36]彭燕,刘宇红,张荣芬.基于 LSTM 的股票价格预测建模与分析[J].计算机工程与应用,2019,55(11):209-212.

[37]王珺,杨晓红,杨凤霞.三因子模型在中国 A 股市场的有效性探讨[J].湖北经济学院学报(人文社会科学版),2013,10(09):27-28+72.

[38]王源昌,汪来喜,罗小明.F-F 三因子资产定价模型的扩展及其实证研究[J].

金融理论与实践,2010(06):45-50.

[39] 王茵田,朱英姿.中国股票市场风险溢价研究[J].金融研究,2011(07):152-166.

[40]王淑燕,曹正凤,陈铭芷.随机森林在量化选股中的应用研究[J].运筹与管理,2016,25(03):163-168+177.

[41]王燕,郭元凯.改进的XGBoost模型在股票预测中的应用[J].计算机工程与应用,2019,55(20):202-207.

[42]杨炘,陈展辉.中国股市三因子资产定价模型实证研究[J].数量经济技术经济研究,2003(12):137-141.

[43]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/OL].计算机工程与应用:1-12[2021-06-17].

附录

附录 1：候选因子库

财务类因子	
估值因子	
EP_TTM	净利润 TTM/总市值
EP_LYR	净利润（最新年报）/总市值
BP_LF	净资产 TTM/总市值
OCF_TTM	经营性现金流 TTM/总市值
SP_TTM	营业收入 TTM/总市值
SP_LYR	营业收入（最新年报）/总市值
FCFP_LYR	自由现金流 TTM/总市值
PEG	市盈率/净利润同比增长率 * 100
规模类因子	
LN_MV	对数总市值
LN_FLOAT_MV	对数流通市值
成长类因子	
SALES_GR_TTM	营业收入增长率_TTM 同比
NET_PROFIT_GR_TTM	净利润增长率_TTM 同比
质量因子	
ROE_TTM	净资产收益率
ROA_TTM	资产回报率
GORSS_PROFIT_MARGIN_TTM	销售毛利率
NET_PROFIT_MARGIN_TTM	销售净利率
行情类因子	
风险因子	
SWING_VOLATILITY_5_0	5 日振幅波动率
SWING_VOLATILITY_10_0	10 日振幅波动率
SWING_VOLATILITY_30_0	30 日振幅波动率
SWING_VOLATILITY_60_0	60 日振幅波动率
SWING_VOLATILITY_120_0	120 日振幅波动率
VOLATILITY_5_0	5 日波动率
VOLATILITY_10_0	10 日波动率
VOLATILITY_30_0	30 日波动率

VOLATILITY_60_0	60 日波动率
VOLATILITY_120_0	120 日波动率
流动性因子	
AVG_TURN_5	过去 5 个交易日的平均换手率
AVG_TURN_10	过去 10 个交易日的平均换手率
AVG_TURN_20	过去 20 个交易日的平均换手率
AVG_TURN_60	过去 60 个交易日的平均换手率
AVG_TURN_120	过去 120 个交易日的平均换手率
技术性因子	
TA_MACD_MACD_12_26_9_0	MACD
TA_MACD_MACDHIST_12_26_9_0	MACD
TA_MACD_MACDSIGNAL_12_26_9_0	MACD
TA_CCI_14_0	CCI 指标，timeperiod=*14*
TA_CCI_28_0	CCI 指标，timeperiod=28
TA_AD_0	收集派发指标
TA_ADX_14_0	ADX 指标，timeperiod=14
TA_ADX_28_0	ADX 指标，timeperiod=28
TA_AROONOSC_14_0	AROONOSC 指标，timeperiod=14
TA_AROONOSC_28_0	AROONOSC 指标，timeperiod=*28*
TA_ATR_14_0	ATR 指标，timeperiod=14
TA_ATR_28_0	ATR 指标，timeperiod=28
TA_MFI_14_0	MFI 指标，timeperiod=14
TA_MFI_28_0	MFI 指标，timeperiod=28
TA_OBV_0	OBV 指标
TA_RSI_14_0	RSI 指标，timeperiod=14
TA_RSI_28_0	RSI 指标，timeperiod=28
TA_SAR_0	SAR 指标
TA_STOCH_SLOWD_5_3_0_3_0_0	STOCH (KDJ) 指标 D 值
TA_STOCH_SLOWK_5_3_0_3_0_0	STOCH (KDJ) 指标 K 值
TA_TRIX_14_0	TRIX 指标，timeperiod=14
TA_TRIX_28_0	TRIX 指标，timeperiod=28
TA_WILLR_14_0	WILLR 指标，timeperiod=14
TA_WILLR_28_0	WILLR 指标，timeperiod=28
动量因子	
TA_MOM_10_0	过去 10 个交易日的价格动量
TA_MOM_20_0	过去 20 个交易日的价格动量
TA_MOM_30_0	过去 30 个交易日的价格动量
TA_MOM_60_0	过去 60 个交易日的价格动量
资金流因子	

MF_NET_AMOUNT_L_0	大单净流入净额
MF_NET_AMOUNT_M_0	中单净流入净额
MF_NET_AMOUNT_MAIN_0	主力净流入净额
MF_NET_AMOUNT_S_0	小单净流入净额
MF_NET_AMOUNT_XL_0	超大单净流入净额

附录 2：候选因子库 IC 分析法计算结果

财务类因子	IC 均值	IC 标准差	IR 值	|IC|>0.02 的比率
估值因子				
EP_TTM	-0.03	0.06	-0.55	80.67%
EP_LYR	-0.03	0.06	-0.57	78.15%
BP_LF	-0.05	0.11	-0.45	88.24%
OCF_TTM	0.04	0.06	0.58	87.39%
SP_TTM	-0.04	0.1	-0.36	90.76%
SP_LYR	0.03	0.1	0.36	89.08%
FCFP_LYR	0	0.04	0.12	56.30%
PEG	0.01	0.04	0.18	68.91%
规模类因子				
LN_MV	-0.01	0.17	-0.04	90.76%
LN_FLOAT_MV	0.01	0.12	0.05	85.71%
成长类因子				
SALES_GR_TTM	0.03	0.06	0.4	80.67%
NET_PROFIT_GR_TTM	0.03	0.06	0.49	79.83%
质量因子				
ROE_TTM	0.03	0.08	0.36	80.67%
ROA_TTM	0.02	0.09	0.27	83.19%
GORSS_PROFIT_MARGIN_TTM	0.01	0.08	0.18	83.19%
NET_PROFIT_MARGIN_TTM	0.02	0.09	0.21	79.83%
行情类因子				
风险因子				
SWING_VOLATILITY_5_0	-0.06	0.07	-0.75	87.39%
SWING_VOLATILITY_10_0	-0.07	0.09	-0.74	90.76%
SWING_VOLATILITY_30_0	-0.08	0.1	-0.77	87.39%
SWING_VOLATILITY_60_0	-0.07	0.11	-0.67	94.12%
SWING_VOLATILITY_120_0	-0.07	0.11	-0.61	84.87%
VOLATILITY_5_0	-0.05	0.09	-0.56	87.39%
VOLATILITY_10_0	-0.07	0.11	-0.68	87.39%
VOLATILITY_30_0	-0.08	0.13	-0.59	88.24%
VOLATILITY_60_0	-0.07	0.14	-0.53	88.24%
VOLATILITY_120_0	-0.07	0.14	-0.48	90.76%
流动性因子				

AVG_TURN_5	-0.11	0.11	-0.97	91.60%
AVG_TURN_10	-0.1	0.11	-0.94	92.44%
AVG_TURN_20	-0.1	0.11	-0.86	89.92%
AVG_TURN_60	-0.08	0.11	-0.69	86.55%
AVG_TURN_120	-0.07	0.11	-0.65	87.39%
技术性因子				
TA_MACD_MACD_12_26_9_0	-0.06	0.1	-0.61	87.39%
TA_MACD_MACDHIST_12_26_9_0	-0.04	0.07	-0.53	81.51%
TA_MACD_MACDSIGNAL_12_26_9_0	-0.05	0.1	-0.56	87.39%
TA_CCI_14_0	-0.04	0.09	-0.44	82.35%
TA_CCI_28_0	-0.05	0.09	-0.53	88.24%
TA_AD_0	0	0.13	-0.04	84.87%
TA_ADX_14_0	0.01	0.09	0.08	78.15%
TA_ADX_28_0	0	0.08	0.04	84.03%
TA_AROONOSC_14_0	-0.04	0.08	-0.56	83.19%
TA_AROONOSC_28_0	-0.05	0.08	-0.63	84.03%
TA_ATR_14_0	-0.03	0.08	-0.35	83.19%
TA_ATR_28_0	-0.02	0.08	-0.27	84.87%
TA_MFI_14_0	-0.04	0.08	-0.52	81.51%
TA_MFI_28_0	-0.05	0.08	-0.63	73.95%
TA_OBV_0	0	0.13	-0.03	87.39%
TA_RSI_14_0	-0.06	0.11	-0.6	89.92%
TA_RSI_28_0	-0.07	0.11	-0.64	94.12%
TA_SAR_0	0	0.06	-0.07	76.47%
TA_STOCH_SLOWD_5_3_0_3_0_0	-0.02	0.09	-0.24	87.39%
TA_STOCH_SLOWK_5_3_0_3_0_0	-0.02	0.09	-0.22	78.15%
TA_TRIX_14_0	-0.07	0.11	-0.62	87.39%
TA_TRIX_28_0	-0.04	0.1	-0.42	92.44%
TA_WILLR_14_0	-0.03	0.1	-0.28	80.67%
TA_WILLR_28_0	-0.04	0.1	-0.37	86.55%
动量因子				
TA_MOM_10_0	-0.05	0.08	-0.6	84.87%
TA_MOM_20_0	-0.06	0.09	-0.68	84.03%
TA_MOM_30_0	-0.06	0.1	-0.59	91.60%
TA_MOM_60_0	-0.05	0.09	-0.53	82.35%
资金流因子				
MF_NET_AMOUNT_L_0	0.04	0.05	0.86	78.90%
MF_NET_AMOUNT_M_0	0.02	0.04	0.41	58.72%
MF_NET_AMOUNT_MAIN_0	0.03	0.06	0.47	72.48%
MF_NET_AMOUNT_S_0	-0.04	0.07	-0.67	84.40%

MF_NET_AMOUNT_XL_0	0	0.05	0.06	43.12%

附录 3：基于 IC 分析法的因子计算结果

财务类因子	IC 均值	IC 标准差	IR 值	|IC|>0.02 的比率
估值因子				
EP_TTM	-0.03	0.06	-0.55	80.67%
BP_LF	-0.05	0.11	-0.45	88.24%
OCF_TTM	0.04	0.06	0.58	87.39%
SP_TTM	-0.04	0.1	-0.36	90.76%
SP_LYR	0.03	0.1	0.36	89.08%
规模类因子				
LN_MV	-0.01	0.17	-0.04	90.76%
成长类因子				
SALES_GR_TTM	0.03	0.06	0.4	80.67%
质量因子				
ROE_TTM	0.03	0.08	0.36	80.67%
ROA_TTM	0.02	0.09	0.27	83.19%
行情类因子				
风险因子				
SWING_VOLATILITY_30_0	-0.08	0.1	-0.77	87.39%
VOLATILITY_30_0	-0.08	0.13	-0.59	88.24%
流动性因子				
AVG_TURN_5	-0.11	0.11	-0.97	91.60%
技术性因子				
TA_MACD_MACD_12_26_9_0	-0.06	0.1	-0.61	87.39%
TA_CCI_28_0	-0.05	0.09	-0.53	88.24%
TA_AROONOSC_28_0	-0.05	0.08	-0.63	84.03%
TA_ATR_14_0	-0.03	0.08	-0.35	83.19%
TA_MFI_14_0	-0.04	0.08	-0.52	81.51%
TA_RSI_28_0	-0.07	0.11	-0.64	94.12%
TA_STOCH_SLOWD_5_3_0_3_0_0	-0.02	0.09	-0.24	87.39%
TA_TRIX_14_0	-0.07	0.11	-0.62	87.39%
TA_WILLR_28_0	-0.04	0.1	-0.37	86.55%
动量因子				
TA_MOM_30_0	-0.06	0.1	-0.59	91.60%
资金流因子				
MF_NET_AMOUNT_S_0	-0.04	0.07	-0.67	84.40%

附录 4：基于随机森林算法的因子重要性排序

财务类因子	重要性得分
估值因子	

FCFP_LYR	0.148612
OCF_TTM	0.131956
BP_LF	0.126081
PEG	0.125612
SP_LYR	0.119159
EP_TTM	0.118628
EP_LYR	0.115401
SP_TTM	0.114552
规模类因子	
LN_MV	0.124356
LN_FLOAT_MV	0.123331
成长类因子	
SALES_GR_TTM	0.124572
NET_PROFIT_GR_TTM	0.121704
质量因子	
NET_PROFIT_MARGIN_TTM	0.254501
GORSS_PROFIT_MARGIN_TTM	0.129638
ROA_TTM	0.124878
ROE_TTM	0.122502
行情类因子	
风险因子	
VOLATILITY_30_0	0.193327
VOLATILITY_10_0	0.174569
VOLATILITY_120_0	0.109561
VOLATILITY_60_0	0.107543
VOLATILITY_5_0	0.101825
SWING_VOLATILITY_5_0	0.070114
SWING_VOLATILITY_30_0	0.065315
SWING_VOLATILITY_120_0	0.062585
SWING_VOLATILITY_10_0	0.057703
SWING_VOLATILITY_60_0	0.057457
流动性因子	
AVG_TURN_10	0.246048
AVG_TURN_5	0.227244
AVG_TURN_20	0.187807
AVG_TURN_120	0.179117
AVG_TURN_60	0.159785
技术性因子	
TA_TRIX_14_0	0.120166
TA_TRIX_28_0	0.096078

TA_SAR_0	0.073547
TA_ATR_28_0	0.049417
TA_ATR_14_0	0.045382
TA_RSI_28_0	0.039897
TA_ADX_28_0	0.039855
TA_MACD_MACDHIST_12_26_9_0	0.039281
TA_CCI_14_0	0.037899
TA_OBV_0	0.037538
TA_AD_0	0.035248
TA_MACD_MACD_12_26_9_0	0.035198
TA_CCI_28_0	0.034283
TA_MACD_MACDSIGNAL_12_26_9_0	0.033866
TA_AROONOSC_28_0	0.033573
TA_WILLR_28_0	0.031713
TA_STOCH_SLOWD_5_3_0_3_0_0	0.031061
TA_ADX_14_0	0.029783
TA_RSI_14_0	0.028588
TA_MFI_28_0	0.027942
TA_AROONOSC_14_0	0.027113
TA_STOCH_SLOWK_5_3_0_3_0_0	0.026996
TA_WILLR_14_0	0.025496
TA_MFI_14_0	0.02008
动量因子	
TA_MOM_30_0	0.29746
TA_MOM_60_0	0.236801
TA_MOM_20_0	0.234069
TA_MOM_10_0	0.23167
资金流因子	
MF_NET_AMOUNT_S_0	0.228478
MF_NET_AMOUNT_MAIN_0	0.207128
MF_NET_AMOUNT_M_0	0.204822
MF_NET_AMOUNT_L_0	0.198326
MF_NET_AMOUNT_XL_0	0.161246

附录 5：基于随机森林算法的因子筛选结果

财务类因子	**重要性得分**
估值因子	
FCFP_LYR	0.148612
OCF_TTM	0.131956
BP_LF	0.126081
PEG	0.125612

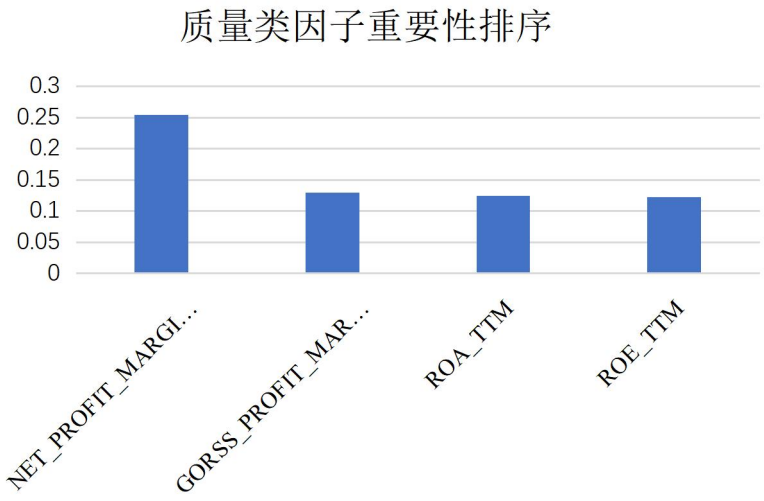
SP_LYR	0.119159
规模类因子	
LN_MV	0.124356
成长类因子	
SALES_GR_TTM	0.124572
质量因子	
NET_PROFIT_MARGIN_TTM	0.254501
GORSS_PROFIT_MARGIN_TTM	0.129638
行情类因子	
风险因子	
VOLATILITY_30_0	0.193327
SWING_VOLATILITY_5_0	0.070114
流动性因子	
AVG_TURN_10	0.246048
技术性因子	
TA_TRIX_14_0	0.120166
TA_ATR_28_0	0.049417
TA_RSI_28_0	0.039897
TA_MACD_MACDHIST_12_26_9_0	0.039281
TA_CCI_14_0	0.037899
TA_AROONOSC_28_0	0.033573
TA_WILLR_28_0	0.031713
TA_STOCH_SLOWD_5_3_0_3_0_0	0.031061
TA_MFI_28_0	0.027942
动量因子	
TA_MOM_30_0	0.29746
资金流因子	
MF_NET_AMOUNT_S_0	0.228478

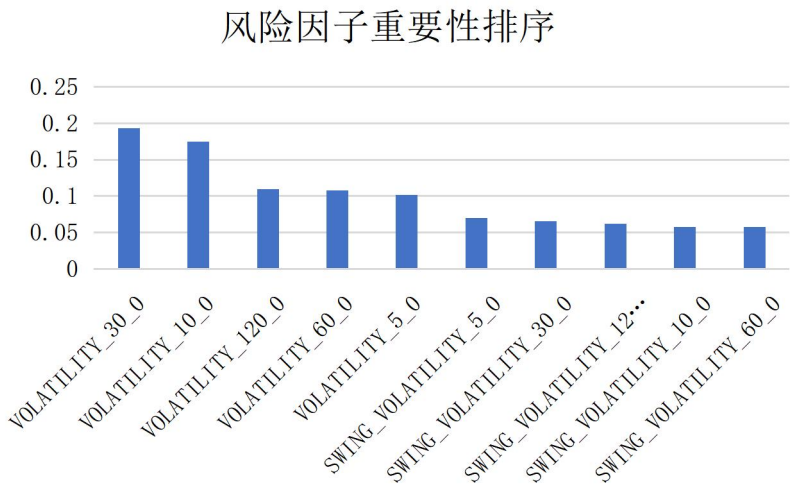
附录 6：IC 分析法与随机森林算法因子筛选结果对比

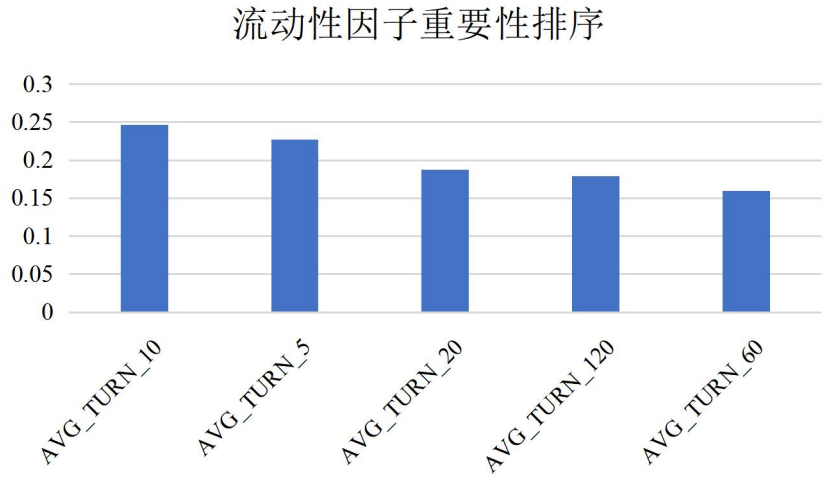
随机森林算法因子筛选结果	**IC 分析法因子筛选结果**
财务类因子	财务类因子
估值因子	估值因子
FCFP_LYR	EP_TTM
OCF_TTM	BP_LF
BP_LF	OCF_TTM
PEG	SP_TTM
SP_LYR	SP_LYR
规模类因子	规模类因子
LN_MV	LN_MV
成长类因子	成长类因子

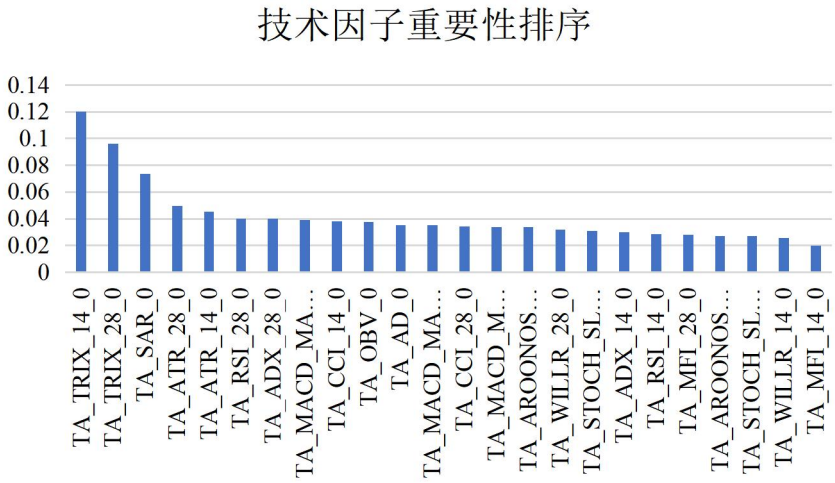
SALES_GR_TTM	SALES_GR_TTM
质量因子	质量因子
NET_PROFIT_MARGIN_TTM	ROE_TTM
GORSS_PROFIT_MARGIN_TTM	ROA_TTM
行情类因子	行情类因子
风险因子	风险因子
VOLATILITY_30_0	SWING_VOLATILITY_30_0
SWING_VOLATILITY_5_0	VOLATILITY_30_0
流动性因子	流动性因子
AVG_TURN_10	AVG_TURN_5
技术性因子	技术性因子
TA_TRIX_14_0	TA_MACD_MACD_12_26_9_0
TA_ATR_28_0	TA_CCI_28_0
TA_RSI_28_0	TA_AROONOSC_28_0
TA_MACD_MACDHIST_12_26_9_0	TA_ATR_14_0
TA_CCI_14_0	TA_MFI_14_0
TA_AROONOSC_28_0	TA_RSI_28_0
TA_WILLR_28_0	TA_STOCH_SLOWD_5_3_0_3_0_0
TA_STOCH_SLOWD_5_3_0_3_0_0	TA_TRIX_14_0
TA_MFI_28_0	TA_WILLR_28_0
动量因子	动量因子
TA_MOM_30_0	TA_MOM_30_0
资金流因子	资金流因子
MF_NET_AMOUNT_S_0	MF_NET_AMOUNT_S_0

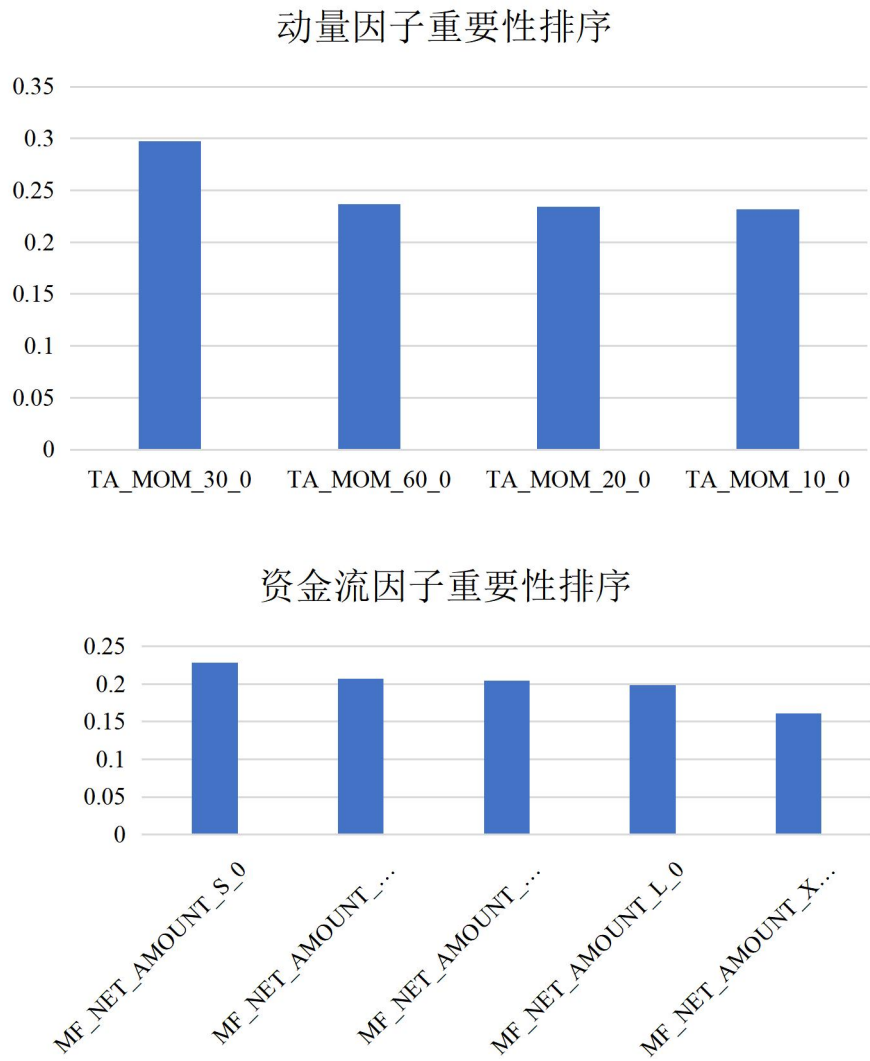
附录 7：因子重要性排序




风险因子重要性排序
0.25
0.2
0.15
0.1
0.05
0
VOLATILITY_30_0
VOLATILITY_10_0
VOLATILITY_120_0
VOLATILITY_60_0
VOLATILITY_5_0
SWING_VOLATILITY_5_0
SWING_VOLATILITY_30_0
SWING_VOLATILITY_12…
SWING_VOLATILITY_10_0
SWING_VOLATILITY_60_0


流动性因子重要性排序
0.3
0.25
0.2
0.15
0.1
0.05
0
AVG_TURN_10
AVG_TURN_5
AVG_TURN_20
AVG_TURN_120
AVG_TURN_60


技术因子重要性排序
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
TA_TRIX_14_0
TA_TRIX_28_0
TA_SAR_0
TA_ATR_28_0
TA_ATR_14_0
TA_RSI_28_0
TA_ADX_28_0
TA_MACD_MA…
TA_CCI_14_0
TA_OBV_0
TA_AD_0
TA_MACD_MA…
TA_CCI_28_0
TA_MACD_M…
TA_AROONOS…
TA_WILLR_28_0
TA_STOCH_SL…
TA_ADX_14_0
TA_RSI_14_0
TA_MFI_28_0
TA_AROONOS…
TA_STOCH_SL…
TA_WILLR_14_0
TA_MFI_14_0



附录 8：基于 LightGBM 模型的因子重要性排序

log(market_cap_0)	144
fs_operating_revenue_yoy_0	70
swing_volatility_30_0	70
fs_net_cash_flow_ttm_0/market_cap_0	61
pb_lf_0	58
avg_turn_5	56
ta_rsi_28_0	55
volatility_30_0	53
ta_macd_macd_12_26_9_0	48
ta_trix_14_0	42
pe_ttm_0	40
fs_roa_ttm_0	36

fs_roe_ttm_0	31
mf_net_amount_s_0	21
ps_ttm_0	19
fs_operating_revenue_0/market_cap_0	16
ta_willr_28_0	14
ta_atr_14_0	13
ta_cci_28_0	11
ta_stoch_slowd_5_3_0_3_0_0	7
ta_aroonosc_28_0	2
ta_mom_30_0	2
ta_mfi_14_0	1

致谢

时光荏苒，3 年研究生学习生活一晃而过，在即将完成毕业论文之际，首先对我的师父陈芳平教授以及授课老师表示感谢，无论平日学习指导，还是此次帮助我完成毕业论文，都离不开师父的指导和授课老师平日的悉心教导，再次由衷的感谢我的师父以及授课老师。