中 文 摘 要

量化投资本质上是通过对金融数据建模来预测股票或期货等金融产品的收益。基于机器学习的量化投资依然面临着诸多挑战，市场竞争加剧有降低了某些策略的效率。尽管如此，技术进步和数据获取能力的提升，尤其是近两年人工智能发展迅速，仍然使得该领域还具有巨大潜力。将机器学习应用于量化投资，不仅为投资提供了新的研究思路和建模方法，还有助于投资者更准确地预测股价，从而降低投资风险并提高收益。

本文主要研究通过机器学习模型如何选择预期收益高，风险低的股票。研究过程如下：（1）下载沪深 300 成分股股票日频数据，对股票进行筛选并构建一系列的异象因子，预处理异象因子，使其符合机器学习模型输入要求。（2）构建机器学习模型网络模型，为了更充分利用样本数据，我们通过滚动方式将样本分为 80%的训练集和 20%的测试集用于模型训练。（3）紧接着我们将梯度提升树与支持向量机等六个机器学习模型与 Ols 模型全部进行评估比较，并发现前馈神经网络的评价指标优于其他六个模型。（4）模型优化，然后我们进一步讨论了第一层全连接层、激活函数、第二层全连接层、输出层以及优化器对前馈神经网络预测结果的影响并就行调参对模型进行相应优化。（5）建立投资策略根据前馈神经网络预测结果去构建具体的投资策略，并分析投资策略的收益。表现出色，多头策略和空头策略投资年化收益分别达到了 70.62%、76.67%。（6）为进一步探讨前馈神经网络模型的有效性，构建基于前馈神经网络模型的沪深 300 选股策略，并通过总收益、夏普比率、年化收益率等指标来评估策略的表现与模型的有效性。

研究结果表明，机器学习模型在量化投资中的有效性优于传统分析方法，特别是前馈神经网络。这些模型的适应性和学习能力是它们能够与市场变化同步、持续提供投资洞察的关键。研究结论认为，机器学习模型在量化投资中显示出较大的潜力和优势，能够识别出传统方法无法覆盖的复杂市场趋势和投资机会。投资者能更有效地管理风险、捕捉投资机会，从而可能提高投资回报。运用机器学习技术进行投资分析，是提升投资决策质量和盈利潜力的一种有效手段。

关键词：量化投资策略；机器学习方法；定价因子；个股特征

ABSTRACT

Quantitative investing is essentially modeling financial data to predict the returns of financial products such as stocks or futures. Quantitative investment based on machine learning still faces many challenges, and increased competition in the market has reduced the efficiency of certain strategies. Nonetheless, technological advances and improved access to data, especially the rapid development of artificial intelligence in the last two years, still leave the field with great potential. Applying machine learning to quantitative investment not only provides new research ideas and modeling methods for investment, but also helps investors to predict stock prices more accurately, thus reducing investment risks and increasing returns.

This paper focuses on how to select stocks with high expected returns and low risk through machine learning modeling. The research process is as follows: (1) Download the daily frequency data of CSI 300 constituent stocks, screen the stocks and construct a series of anomalous factors, and preprocess the anomalous factors to make them meet the input requirements of the machine learning model. (2) Construct the machine learning model network model, in order to make fuller use of the sample data, we divide the sample into 80% of the training set and 20% of the test set for model training by rolling. (3) Immediately after that, we evaluate and compare all six machine learning models such as gradient boosting tree and support vector machine with Ols model, and find that the evaluation index of feedforward neural network is better than the other six models. (4) Model optimization, then we further discuss the impact of the first fully-connected layer, activation function, second fully-connected layer, output layer, and optimizer on the prediction results of feed-forward neural networks and optimize the model accordingly with respect to row tuning parameters. (5) Establishing investment strategies According to the feed-forward neural network prediction results to build specific investment strategies, and analyze the return of investment strategies. The performance is excellent, with annualized returns of 70.62% and 76.67% for long and short strategies, respectively. (6) In order

to further explore the effectiveness of the feed-forward neural network model, the CSI 300 stock selection strategy based on the feed-forward neural network model is constructed, and the performance of the strategy and the effectiveness of the model are evaluated through the indicators of the total return, Sharpe ratio, and annualized return.

The results show that the effectiveness of machine learning models in quantitative investment is better than traditional analytical methods, especially feed-forward neural networks. The adaptability and learning capabilities of these models are key to their ability to keep pace with market changes and consistently provide investment insights. The study concludes that machine learning models show greater potential and advantage in quantitative investing, identifying complex market trends and investment opportunities that cannot be covered by traditional methods. Investors are able to manage risks and capture investment opportunities more effectively, thereby potentially improving investment returns. The use of machine learning techniques for investment analysis is an effective means of enhancing the quality of investment decisions and profit potential.

Key Words: Quantitative investment strategy; Machine learning methods; Pricing factors; Individual stock characteristics

目 录

第 1 章 绪论

1.1 研究背景

随着计算机技术的飞速发展和数据采集技术的日益成熟，机器学习算法的应用领域不断拓展，特别是在量化投资领域，机器学习算法已成为提高投资决策准确性的关键工具之一。量化投资，作为一种以数学模型为基础，通过大量历史数据分析来驱动投资决策的方法，近年来越来越受到投资者的青睐。其中，基于机器学习模型的基本面量化投资策略，结合了传统的基本面分析和先进的机器学习技术，旨在挖掘公司基本面数据中隐藏的、对股票价格有预测价值的信息，以实现对资产超额收益率的准确预测。

传统的投资分析方法，无论是基于公司基本面的分析还是技术面的分析，都依赖于分析师的主观判断，这不可避免地带来了投资决策的不确定性。而机器学习算法的应用，则为量化投资提供了一种新的思路。通过训练机器学习模型来识别和处理大量的历史数据，可以更加客观和系统地分析影响股票价格的各种因素，减少人为因素的干扰，从而提高投资决策的准确率和客观性。此外，多因子模型理论提供了一个分析资产超额收益的框架，通过识别影响资产价格的关键因子，并结合机器学习算法强大的数据处理和模式识别能力，可以有效地预测资产的未来表现。

将机器学习算法应用于基本面量化投资策略，不仅能够充分利用现有的大数据资源，而且可以通过机器学习模型的强大预测能力，为投资者提供更加科学、客观和准确的投资决策支持。这种结合了传统基本面分析和现代机器学习技术的新型量化投资策略，无疑具有重要的理论意义和实践价值，值得在未来的研究中进一步探索和深化。

1.2 研究意义

1.2.1 理论意义

本文的研究成果在理论层面促进了金融技术和量化投资的发展，并且重点强调了提高决策过程的透明度和可理解性。通过使用易于解释的机器学习模型，可以让复杂的算法和模型对投资者及监管机构更为透明和可信，这有利于提高投资策略的有效性，并减少误解及投资风险。此研究促进了机器学习与金融理论的整合，为金融市场的稳健发展提供了新的视角。同时，它还为增强机器学习模型的伦理性和责任性奠定了理论基础。综上所述，这项研究在技术创新方面取得进展，对于推动金

融市场的合理性、透明性和效率性，具有重大的理论意义。

传统的计量方法主要有两类，分别是截面回归和时序回归。在因子投资或资产定价领域，前者常用滞后的股票特征对个股的未来收益率进行回归；后者则往往将投资组合的整体收益对部分宏观变量进行回归。然而，这些传统计量方法具有相当的潜在局限性。而机器学习中所应用的更先进的统计工具可以帮助克服这些局限性，包括各种高维预测模型、正则化手段以及模型优化筛选算法。虽然机器学习方法本身并不能识别预测目标与特征变量之间深层的逻辑关系。但机器学习仍然有助于我们了解其中的机制。

1.2.2 现实意义

1.2.2.1 提高投资决策的效率和准确性

随着金融市场的复杂性不断增加，传统的投资方法在处理大规模数据和捕捉市场微妙变化方面显得力不从心。机器学习模型以其在数据处理和模式识别方面的优势，为量化投资策略提供了更高效、更准确的决策支持。在数据处理能力方面，机器学习模型能够分析和处理海量的市场数据，包括价格波动、交易量、财务指标等，帮助投资者在庞杂的信息中迅速发现价值。机器学习模型又能够识别隐藏在复杂市场数据中的模式，并对市场走势进行预测。与传统方法相比，机器学习在预测精度和速度上都有显著优势。

1.2.2.2 促进金融市场稳定性和监管效率

机器学习模型基于客观数据做出决策，减少了人为干预和情绪波动的影响，降低了市场的非理性投机行为。这些模型能快速响应市场变化，提高市场反应速度和效率。此外，通过深入分析和预测市场风险，这些模型帮助投资者更好地管理和分散风险，提高市场稳定性。同时，这也有助于加强市场监管，提高监管效率和有效性。

1.2.2.3 推动金融科技的发展和创新

机器学习模型在量化投资策略中的应用，推动了金融科技领域的发展和创新。这不仅促进了新技术在金融领域的应用，也为其他行业提供了技术创新的范例。不仅促进跨学科融合，将机器学习与金融投资结合，促进了计算机科学、统计学与经济学之间的交叉融合。又能够驱动创新，金融科技的快速发展，尤其是在大数据、云计算和人工智能领域的进步，为量化投资策略提供了新的思路和工具，推动了整个行业的创新和发展。

总而言之，“基于机器学习模型的量化投资策略”不仅在理论上具有深刻意义，其在实际应用中也展现出巨大的价值。这些策略通过提高投资决策的效率和准确性，促进市场稳定性和透明度，以及推动金融科技的创新发展，对现代金融市场产生了深远的影响。随着科技的不断进步和金融市场的日益复杂化，基于机器学习的量化投资策略将继续在金融领域扮演重要角色，成为推动金融创新和稳定的关键因素。

1.3 研究内容与思路

第一部分为绪论。主要阐述了量化投资的背景、发展历程以及研究意义。重点放在了量化投资策略的重要性和应用范围，以及可解释机器学习模型在此领域的应用前景。从而对文章整体内容及结构起到介绍的作用。

第二部分为理论基础与文献综述。先介绍了国内外与量化投资策略研究相关的文献研究，以及机器学习模型在因子投资中的应用的研究。接着介绍了国内外量化投资策略研究方面有关的理论以及量化投资与机器学习模型的相关理论。基于对上述相关研究文献的整理总结，从而对本文的研究内容及状况有一定的了解。

第三部分为数据处理部分。主要介绍了数据来源、数据内容以及基本面因子选取和数据的预处理方法以及具体做法。再对候选因子进行有效性检验，详细说明了如何构建基本面量化投资策略和实现可解释机器学习模型的具体方法，以及用于实现可解释机器学习模型的算法和技术。

第四部分为实证研究，详细说明模型设计和评估指标。本文选取 2010 年至 2022 年的沪深 300 成分股数据为样本，构建收益公告异常交易量、股东权益变化等 11 个基本面异象因子。再进行模型训练、测试并对模型结果进行可视化的分析和解释，对各机器学习模型进行评价从而选择。再对评价最优的模型进行进一步的优化，进行再一次的训练提出投资策略，

第五部分为结论，总结了本文实证研究结果，还分析了本文在研究中的不足。最后展望未来在机器学习模型在基本面量化投资策略中的应用前景。

1.4 主要贡献及创新

本文创新点是：第一，在中国 A 股市场中比较了多个机器学习模型的优劣，包括梯度提升树、支持向量机和前馈神经网络等。通过在实际市场数据上测试这些模型，研究揭示了它们在股票选择和投资策略方面的有效性和性能。这种方法提供了一种新的角度来评估和选择适用于量化投资的机器学习模型，对于寻找适合中国股市的高效投资策略具有重要意义。第二，本文对复杂金融时间序列数据进行的深度

分析。本文不仅关注了常规的股票价格和交易量数据，还结合了宏观经济指标、公司财务报表数据等多维度信息。这有助于捕捉更多潜在的投资机会，并提高投资策略的有效性和精准度。

第 2 章 文献综述

2.1 国内外量化投资概述

量化投资(Quantitative Investment)，也称为系统化投资(Systematic Investment)，是现代金融领域的一种重要投资策略，它通过应用数学模型、统计学方法和计算技术来指导投资决策。与传统基于直觉和经验的投资方法不同，量化投资更加依赖于数据分析和算法处理，目标是在复杂多变的金融市场中寻找投资机会，并努力实现超越市场平均水平的回报。

在量化投资中，数据的重要性不言而喻。投资者会搜集包括股票价格、交易量、公司财务数据以及市场经济指标等大量历史和实时数据。这些数据不仅限于传统的数值型数据，还可能包括来自社交媒体、新闻报道等的非结构化数据。通过对这些数据的综合分析，量化投资旨在识别市场趋势和投资机会，为投资决策提供科学依据。量化投资策略的核心是建立和应用复杂的数学模型和统计算法。这些模型和算法能够处理和分析大量数据，帮助投资者预测市场走势、评估风险和构建投资组合。在实际操作中，量化投资经常采用算法交易，即通过自动化的交易系统按照预设的策略执行买卖指令。这种方法不仅提高了交易效率，还有助于降低交易成本和市场冲击。然而，量化投资并非没有风险。市场的复杂性和不可预测性意味着即使是最先进的模型也无法完全消除投资风险。此外，模型和数据的错误可能导致意外的损失。因此，风险管理成为量化投资的一个重要组成部分，包括利用多种风险控制模型来评估和管理潜在风险。

量化投资策略通过应用数学模型、计算机算法和统计分析来指导投资决策，可分为多种类型，每种类型都依据不同的理论和市场行为进行操作。量化投资策略的主要分类有动量策略、均值回归策略、套利策略、因子投资策略、事件驱动策略、机器学习策略、高频交易（HFT）等。不同的策略各有特点，适用于不同的市场环境和投资目标。量化投资者通常会根据市场条件、风险偏好和投资期限来选择和调整适合的投资策略。随着金融科技的发展，这些策略也在不断演进，以适应市场的变化。

总的来说，量化投资通过减少投资决策的主观性和随意性，使投资过程更加科学和系统化。随着技术的不断进步和数据量的日益增长，量化投资的方法和策略也在不断发展和完善，为投资者带来了新的机遇和挑战。

2.1.1 国内量化行业的发展历程

量化投资的发展在中国可以划分为若干阶段。在 2004 年以前，该领域还处于起步阶段，缺少历史悠久的公募量化基金。自 2004 年起，随着首个公募量化基金的推出，量化投资开始逐步进入发展期。然而，由于早期缺少有效的对冲工具，策略主要集中在套利上，公募基金主要关注指数和类似指数的产品，而专注于量化的基金则较为少见。到了 2008 年，随着全球金融危机的发生，大批海外量化专家返回国内，为该行业注入了新的动力。2010 年，随着沪深 300 股指期货的引入，为量化投资提供了关键的对冲工具，从而促进了私募行业的规范发展和量化投资的加速成长。然而，由于策略偏好小市值风格，2015 年的股市危机对量化策略造成较大回撤，促使基金管理者开始重视差异化竞争。此后，随着新技术的引入，量化策略变得日益丰富和多样化。到了近年来，随着人工智能和机器学习技术的发展和应用，量化投资领域不断创新，市场环境和监管政策的变化也在影响着策略和发展方向，同时投资者对量化投资的认知和接受度逐渐提高，有助于该领域的进一步成长和成熟。中国的量化投资行业经历了从起步到快速发展，再到策略多元化和技术革新的过程，目前正处于不断成熟和创新的阶段。

截至 2023 年第三季度，中国公募量化基金的总规模达到 3198 亿元。虽然行业整体规模略有收缩，这一增长在当前市场环境下仍然突出。在权益型公募基金中，量化基金的市场份额由 4.4%微增至 4.8%。在业绩方面，量化基金在超额收益上表现出色，沪深 300、中证 500、中证 1000 指数增基金的超额收益分别为 1.3%、1.0%和 2.4%，与上一季度持平。尽管量化对冲基金在过去两个季度连续获得正收益，但季度收益中枢降至-0.1%，全年来看仍显示出可观的防跌能力。关于市场展望和风险，当前股指市场估值可能并不处于高位，为指数增强型产品提供 Beta 支撑，同时交易结构倾向于中小盘股，这有助于中小市值股票的 Alpha 收益。长期来看，随着海外加息政策结束和国内经济基本面的改善，CTA 产品预计将实现反弹。从行业发展趋势来看，尽管存在挑战，量化基金行业显示出韧性和成长潜力，其适应市场变化和策略创新的能力预计将在未来发挥关键作用。

2.1.2 国外量化行业的发展历程

量化金融行业的发展历程始于 20 世纪初，当时路易斯·巴舍利耶引入了期权定价的数学模型。在 1950 年代，哈里·马科维茨的投资组合选择理论推动了现代投资组合理论的发展。20 世纪后半叶，随机微积分和连续时间过程的引入，特别是在

1960 年代至 1980 年代期间，进一步推进了量化金融的发展。黑-休尔斯模型在 1970 年代对衍生品定价产生了重要影响。近几十年来，机器学习和人工智能在量化金融中的应用不断增加。2007-2008 年金融危机后，该行业开始更加重视风险管理和监管合规。这个历程涵盖了从早期的基础理论到 21 世纪的数学、统计和计算技术的发展。

2.2 国内外量化投资策略相关研究

2.2.1 国外关于量化投资策略的研究

量化投资最初起源于海外，早在 20 世纪初，法国学者 Bachelier 开始采用量化手段来描述布朗运动，并在《投机理论》一书中提出了随机游走假说，即股票价格遵循一定的随机路径，特定时间内达到某价格的可能性呈现正态分布[1]。他也发展了随时间变化的股价模型，奠定了随机游走理论的基础。此理论模型在量化投资研究中被广泛认为是一个重要的起点。但这一理论当时与主流观点相悖，因而被长时间忽视。直到 1955 年，经济学家 Samuelson 重新发现了这篇论文，并对其包含的多个前沿学术成果表示惊讶。1960 年，Cootner 发表了《股票市场的随机性》，将 Bachelier 的成果纳入学术讨论并进行了深化[2]。在此之前，在 1952 年，Markowitz 提出了均值-方差组合模型，用以量化收益和风险，将数理统计应用于投资组合管理，对量化投资的发展产生了深远影响[3]。学术界随后对金融和投资学领域进行了更深入的数理统计方法研究。在随机游走理论提出同年，Sharpe 等人（1964）扩展了 Markowitz 的方法，提出 CAPM 模型，这一模型分析了无风险利率和与市场相关的超额收益对证券价格的影响[4]。Fama（1964）基于这些研究，提出了有效市场假说，强调了市场信息的即时反应性[5]。1976 年，Ross 基于资本资产定价模型提出了套利定价理论，该理论指出无风险套利的出现暗示着市场价格不均衡[6]。Kahneman 和 Tversky（2000）观察到投资者对风险的态度依其盈亏状态变化[7]。Shefrin 和 Statman（1985）探讨了投资者在亏损时的行为倾向[8]。Fama 和 French（1992）提出了三因子模型，为业绩归因提供了理论框架[9]。Haugen 和 Baker（1996）通过实证分析揭示了多种因素对股票期望收益的影响[10]。Piotroski（2004）通过分析财务报表来构建多因子选股模型[11]。Arlen 等（2006）讨论了构建量化投资策略的关键元素。这些研究成果为量化投资领域的发展奠定了坚实基础[12]。

随着量化投资领域的发展 Wesley R. Gray 和 Jack R. Vogel（2016）提供了一个基于动量的股票选择系统的实用指南[13]。Marcos López de Prado（2012）展示了机器学习技术在预测模型和算法交易策略中的应用[14]。Irene Aldridge（2013）详细讨论了

高频交易的各个方面[15]。Emmanuel Jurczenko（2017）深入探讨了因子投资的不同方面[16]。这些理论成果及其变形仍广泛运用于投资实务界中，为后续的量化投资领域奠定了坚实的基础。

2.2.2 国内关于量化投资策略的研究

量化投资策略在国外已较为成熟，而中国则较晚开始发展。尽管这样，得益于信息技术的迅速进步，量化投资在中国资本市场逐渐得到了认可。2005 年，范龙振发现了中国 A 股市场的市值效应、市盈率效应和价格效应，并提出了解释这些市场指数差异的三因子模型[17]。到了 2017 年，李斌等人设计了基于机器学习和技术指标的量化投资算法（ML-TEA），验证了其在预测投资组合走势上的有效性[18]。接下来，黄卿等人在 2018 年利用 XGBoost、支持向量机和神经网络来预测沪深 300 指数，进一步证明了这些方法在中国市场的适用性[19]。同年，王云凯等人结合基本面多因子数据和梯度提升机、随机森林等方法构建了投资组合[20]。2020 年，张虎等人通过筛选因子和采用自注意力机制的神经网络，开发了新的量化多因子策略。进一步展示了机器学习技术在量化投资领域的发展和有效性[21]。

量化投资策略的研究和实践在中国市场已经取得了显著的进步，尤其是在信息技术和机器学习的驱动下。国内外的研究都强调了特征因子选取的重要性，指出其直接影响量化选股模型的预测能力和策略收益。总体而言，量化投资领域的快速发展预示着这一领域将继续涌现出新的策略和技术，尤其是人工智能技术在金融领域的应用前景令人憧憬。尽管国内金融市场还在探索和规范中，但量化投资的研究和实践正在不断进步，展示出广阔的发展潜力。

2.3 机器学习在因子投资中的应用

近期，学者们开始利用机器学习解决传统研究方法难以应对的大量因子问题。研究主要分为三个方向：一是，使用变量选择技术评估因子对资产定价的影响。例如，Feng 等（2017）应用 Lasso 技术评价因子在资产定价中的作用，发现与众多因子相比，盈利和投资因子具有更明显的统计意义[22]。二是，通过机器学习提取因子共性以解释截面收益率。例如，Light 等（2017）利用偏最小二乘法（PLS）检验企业特征对预期收益的预测效果[23]；Kozak 等（2019）[24]和 Kelly 等（2019）[25]分别使用 PCA 和 IPCA 技术挖掘因子共性，结果表明基于少数主成分的模型能独立预测截面收益。三是，开发新的集成技术以提升预测性能。例如，Lewellen（2015）通过 FM 回归整合 15 个因子，有效预测了股票的超额收益[26]；DeMiguel 等（2017）

从投资者效用角度探索公司特征对截面收益的预测能力，找到了六个独立预测平均收益的公司特征[27]；Gu 等（2018）评估了美国市场上常见的机器学习算法性能，证实机器学习模型超越了传统线性回归模型[28]。

相对于美国市场，中国股市结合机器学习和因子策略的研究还较少。胡熠和顾明（2018）从多个角度选取了八个因子构建综合指标，应用于中国 A 股市场，验证了价值投资策略的有效性[29]。Hsu 等（2018）比较了美国市场常见因子在中国市场的效果，指出两国市场因子有效性有显著差异[30]。Jiang 等（2019）使用 FM 回归、PCA、PLS 和 FC 方法整合 A 股市场的多个因子，有效提取了预测信息[31]。

此外，大量文献应用机器学习预测股价或收益。李斌等（2019）使用支持向量机、神经网络、Adaboost 等算法根据多个技术指标预测股价变动，证明了这些算法的高预测准确性和投资绩效[32]。Krauss 等（2017）整合了深度神经网络、梯度提升树和随机森林策略，预测了标普 500 指数的涨跌情况[33]。Fischer 和 Krauss（2018）利用长短期记忆模型（LSTM），基于日频收益率数据预测股票相对表现，其构建的投资组合表现优于传统模型[34]。

2.4 有关多因子模型的研究

中国金融学家们已经对三因子模型进行了大量的实证研究，以深入了解中国股市的资本资产定价机制。然而，他们发现，CAPM 模型在中国的各个时期都存在着一定的局限性。杨朝军在 1998 年的实证研究表明，系统性风险只是影响投资者收益的一个重要因素，但股票收益率的变化可能会对投资者的决策产生更大的影响，这一点在沪市 1993 年至 1995 年的 A 股股票中得到了明显的体现[35]。孙爱军和陈小悦 2000 年深入研究了从 1994 年到 1998 年沪深 A 股和 B 股的股票走势，并以此为基础，发现β值对于理解股票价格变化的规律性影响甚微，因此，他们得出的结论是，CAPM 模型的有效性可能无法得到充分的验证。 陈冬华、张田余和陈信元从 1996 年到 1999 年，采用多种方法，但均未能找到β值的解释[36]。刘霖、靳云汇则从 1997 年到 2000 年，分别挑选出 A 股和 B 股，经过实证分析，发现β值之外的其他因素与股票的平均收益率也存在一定的联系，而这种联系也非常复杂。近年来，大量研究结果表明，资本资产定价模型并不能有效应用于国内股市[37]。

随着三因子模型的推广，国内金融学者纷纷展开了深入的探索，通过实证分析，他们发现三因子模型在中国股市的各个时期都起到了至关重要的作用。经过余世典和范龙振在 2002 年的实证研究，他们从 1995 年至 2000 年的所有 A 股中收集了大量

数据，发现账面市值比和规模效应对股市的发展有着重要的影响，这一发现为股市的发展提供了重要的理论支撑[38]。马永凯和邓长荣在 2005 年通过对深圳 A 股 1996 年至 2003 年的实证研究，发现三因子模型的运行效果良好，表明国内市场存在价值效应和规模效应，这大大提高了超额收益的解释度。不同行业，也建立了三因子模型[39]。黄建山和刘辉在 2013 年的实证研究表明，从 1995 年到 2011 年，国内 A 股市场三因子模型解释能力远超 CAPM 模型，并呈现出明显的价值效应与规模效应[40]。

四因子和五因子模型推出之后，国内金融研究者也做了大量的实证分析，然而观点和研究没有一致结果，对中国市场的适用性也有很多的讨论。李飞、欧阳志刚于 2016 年在三期中加入了滞后六个月的动量因子。该因子模型发现四个新因子中国股市具有较好的解释力[41]。周晓华、高春亭 2016 年的研究发现，投资因子 CMA、盈利因子 RMW、价值因子 HML、规模因子 SMB 和的显著性使得规模依次递增，这是因为五因子模型的解释力明显高于三因子模型，从而使得投资效应得到了更好的体现，所以五因子模型是基本适用于我国证券市场的[42]。

综上研究，三因子模型在 CAPM 模型框架下引入了 SMB 规模因子和 HML 价值因子，四因子模型在此基础上加入 WML 动量因子，五因子模型又加入投资因子 CMA。

2.5 文献述评

文献涵盖了量化投资策略的历史发展、重要理论模型、以及机器学习技术在该领域的应用。从 Bachelier 关于随机游走模型的早期工作，到 Markowitz 的均值-方差组合模型，以及后续的资本资产定价模型（CAPM）和有效市场假说，文献展示了量化投资如何从基本的数学模型发展成为一个复杂且高度统计驱动的领域。强调了多个关键的理论进展，如 Fama 和 French 的三因子模型，这些理论模型为投资组合的构建和风险管理提供了基础。此外，它们也提到了使用机器学习技术进行股票市场分析的较新研究，包括但不限于利用多因子分析、高频交易和算法交易策略。

国内关于量化投资的研究相对较晚开始，但随着信息技术的发展，量化投资在中国市场也取得了显著的进展。这包括对基本面数据的分析，如市值、市盈率和价格效应等。国内研究者也发现，在 A 股市场中，这些因素可以有效地解释股价波动和市场表现的差异。

特别值得注意的是，机器学习在量化投资中的应用日益增长。这包括使用机器学习算法进行股价预测、投资组合管理和因子投资策略。这些方法超越了传统的统

计模型，提供了更高的预测准确率和投资绩效。

总体而言，这些文献展示了量化投资策略的演变，从基本的数学模型到复杂的机器学习应用。随着技术的不断发展和金融市场的不断变化，预计未来这个领域还会继续进步和发展。

最后通过文献阅读，整合了文献中所使用最频繁的 38 个异象因子用于此论文的因子构建。并且可以通过这些论文可以知道此 38 个因子对预测股价由很大贡献。

2.6 相关理论和模型

2.6.1 量化投资概念及其理论基础

在传统的投资分析领域，基本面分析和技术面分析是两种主要的方法。量化交易，另一方面，涉及使用现代信息技术来创建与金融数据相关的模型，以指导投资决策。这种方法结合了计算机和数学技术，用于构建投资组合，以更精确和高效的方式处理投资问题。区别于传统投资，量化投资结合了理论分析、市场风险感知和个人经验，形成了一种基于数据分析的投资模型。它利用信息技术分析数据，探索股市规律，制订合理的投资策略，以实现投资组合的超额收益。

量化投资被视为一种主动的投资策略，旨在识别并利用市场提供的高收益机会。通过构建优化的投资组合，投资者努力分析目标投资，考察市场和行业动态，以超越市场平均水平。追求超额收益的投资者需深入了解市场动态，探究影响股票未来收益的多种因素，如公司估值和行业趋势，从而构建一个能够实现超额收益的最佳投资组合。主流量化投资方法分为量化选股、量化择时、期货套利、统计套利、算法交易及资产配置和现代数据方法的应用等（丁鹏，2012）。

2.6.1.1 现代投资组合理论

哈里·马科维茨（Harry Markowitz）首次提出了有效边界、风险报酬以及投资组合有效前沿的概念，用结构化、数据化的方式，解决了投资过程中风险与收益度量的问题。马科维茨通过用投资组合的历史收益率代替投资组合的期望收益，用历史收益率的方差来表示投资组合的风险。投资者通过计算资产的收益与风险，构建所有投资组合的有效边界，并结合投资者的无差异曲线，选择最优的投资组合。该理论的前提是投资者都是理性人，并且还包含以下假设：

（1）投资者在面对各种可选择的投资组合时，在相同的风险水平下，都会选择预期收益最大的投资组合。或者在相同的预期收益水平下，投资者都会选择风险最小的投资组合。

（2）理性投资者的效用曲线是一个复杂的函数，它受到预期收益和方差的双重影响，从而决定了最终的结果。

（3）所有的投资者都是理性投资人，会寻求更高的收益，但是边际效用是递减的。

该理论的目标函数公式如下：

$$min\delta^2(r_p) = \sum\sum w_i w_j\, cov(r_i, r_i) \qquad \text{公式（2-1）}$$

$$E(r_p) = \sum w_i r_i \qquad \text{公式（2-2）}$$

其中，r_p为组合收益；r_i, r_i为第i种、第j种资产的收益；w_i, w_j为资产 i 和资产 j 在组合中的权重；$\delta^2(r_p)$为组合收益的方差即组合的总体风险；$cov(r_i, r_i)$为两种资产之间的协方差。通过上述公式，可以得到所有可能的投资组合有效边界，位于有效边界上的投资组合具有不同的收益率与风险，但总能取得相同风险水平下的最高收益。投资者作为理性人会在有效边界上选择合适的投资组合，即个人效用函数与有效边界的交点。

2.6.1.2 套利定价理论

1976 年，斯蒂芬·罗斯（Stephen Ross）提出了著名的套利定价理论（Arbitrage Pricing Theory，简称 APT），并且认为风险资产的收益与若干共同因子（代表系统风险）相关。与资本资产定价模型（CAPM）相比，APT 更少依赖于理论假设，并且更加适用于现实的证券市场。APT 认为，资产价格受到多种经济因素的影响，这些因素可以通过一系列的因子来表示。通过对这些因子进行加权和求和，可以计算出证券的合理价格。相对于 CAPM，APT 的一个优点是它不要求所有投资者都对证券收益的未来概率分布达成一致，这使得投资者可以根据自己的预期进行投资决策。

2.6.1.3 有效市场假说

20 世纪 70 年代，Eugene Fama 提出了市场有效性水平的评判标准。他认为市场的有效性程度与市场价格反映信息的程度直接相关，因此提出了三种有效市场：弱有效、半强有效和强有效。在弱有效市场下，证券价格已经充分反映了所有的历史数据信息，包括股价和成交量等，因此技术分析将无效。在半强有效市场下，价格已经充分反映了所有已公开的信息，包括股价、成交量、盈利数据、盈利预测、公司管理情况和其他公开的财务信息等，因此基本面分析将失去作用。在强有效市场下，价格已经充分反映了所有公开和未公开的信息，即使投资者掌握内幕消息也无法获得超额收益。实际上，市场存在很多异常现象，表明市场至多是弱有效的，有

时可能是半强有效的。市场无法完全有效的原因很多，比如获取信息的成本高昂、信息传递速度慢、投资者无法处理大量信息等。正因为存在这些限制，许多量化投资者认为市场并非完全有效，他们的努力可以带来回报。

2.6.2 线性回归模型

线性回归是一种基本的统计分析方法，用于建立自变量（解释变量）和因变量（响应变量）之间的线性关系。它包括简单线性回归（一个自变量）和多元线性回归（多个自变量），线性回归以其强大的解释性、简便的计算方法和广泛的应用而闻名，适用于从经济学到社会科学的各个领域。虽然线性回归假设变量间存在线性关系，且对异常值敏感，可能在某些情况下限制了它的适用性，但它仍然是分析变量关系和进行预测的有效工具。线性回归在多重共线性情况下可能遇到问题，但总体而言，它是理解变量间影响和提供决策支持的基础分析模型。

普通最小二乘法（Ordinary Least Squares，OLS）。这是一种在统计学中广泛应用的技术，用于在线性回归模型中估计未知参数。普通最小二乘法的目标是最小化误差项（即实际观察值与模型预测值之间差异）的平方和。在一个线性回归模型中，我们假设一个因变量（或被解释变量）可以通过一个或多个自变量（或解释变量）的线性组合来预测。普通最小二乘法通过找到这样一组参数（通常表示为线性方程中的系数），使得模型预测值和实际观测值之间的总误差（即所有误差的平方和）最小，从而确定这些参数的最佳估计值。普通最小二乘法的主要优点是其计算简单、直观，并且在满足某些基本假设（如误差项的独立性、同分布性、正态性等）时，可以提供最佳的无偏估计。然而，当这些假设不满足时，OLS 估计可能不再是最佳的或者可信的。此外，OLS 对异常值非常敏感，可能会由于异常值的存在而产生误导性的估计。

Lasso 回归（Least Absolute Shrinkage and Selection Operator）是一种线性回归的扩展，它通过引入 L1 正则化来进行特征选择和模型的简化。L1 正则化的关键特点是它倾向于产生稀疏模型，即模型中的许多系数会变成零。这种性质使得 Lasso 回归不仅能够处理过拟合问题，还能自动进行特征选择，因此在变量众多且相互关联的数据集（例如在金融量化分析中常见）上特别有用。在量化领域，Lasso 可以帮助筛选出对预测目标变量影响最大的少数特征，从而简化模型并提高预测的可解释性。然而，当存在一组高度相关的变量时，Lasso 可能只选择其中一个变量，忽略其他变量，这可能会导致模型解释性的降低。

岭回归（Ridge regression），也称为 Tikhonov 正则化，是一种用于分析多重共线性数据的技术，特别适用于当预测变量数量超过观测数量的情况。它通过引入 L2 正则化来解决标准线性回归的一些问题。Ridge 回归的 L2 正则化会惩罚系数的大小，使得模型中的所有系数都被缩小，但不会完全到达零。这有助于降低模型的复杂度，减轻过拟合问题，同时保持所有变量在模型中。在量化领域，Ridge 回归有助于构建更稳健的模型，尤其是在处理涉及大量预测变量和参数的金融数据时。然而，Ridge 回归并不进行特征选择，可能会保留一些不重要的特征。

弹性网络（ Elastic Net）是一种结合了 Lasso 回归和 Ridge 回归优点的线性回归方法。它通过同时加入 L1 和 L2 正则化来整合两种方法的优势。Elastic Net 特别适用于处理具有多重共线性或预测变量数量超过样本数量的情况。在这些情况下，Lasso 可能无法有效工作，而 Ridge 又不能进行特征选择。Elastic Net 通过调整两种正则化之间的平衡，可以保持模型的稀疏性并提高模型的稳定性。在量化领域，这种方法有助于构建既稳健又具有解释性的模型，特别是在面对复杂的金融数据集时。但是，调整 Elastic Net 的正则化参数可能比调整单一正则化更为复杂。

2.6.3 机器学习模型

机器学习是人工智能的一个核心分支，它使计算机能够通过分析和处理大量数据来“学习”，从而能够识别数据中的模式和关系，并在没有特定编程指令的情况下做出决策或预测。这个学习过程主要分为三种类型：监督学习，其中算法通过分析带有标签的数据集来学习输入和输出之间的映射关系；非监督学习，这涉及到处理未标记的数据，以探索数据的潜在结构；以及强化学习，它涉及到算法（代理）通过与环境的互动来优化其决策过程，通常是基于其行为的结果（如奖励）。机器学习的应用非常广泛，涵盖了从图像和语音识别到自然语言处理、医疗诊断和股市交易等多个领域。它包括一系列算法，如决策树、支持向量机（SVM）、神经网络和随机森林等。有效的数据处理，包括数据归一化、去噪和特征提取，对于成功的机器学习至关重要。尽管机器学习有着巨大的潜力，但它也面临诸多挑战，包括对数据质量和数量的依赖、模型的泛化能力、以及避免过拟合和欠拟合的问题。随着技术的不断进步，机器学习在各个行业的应用正在迅速扩展，对未来社会和经济产生深远的影响。

2.6.3.1 支持向量机（SVM）

支持向量机（SVM）是一种强大的监督学习算法，适用于分类和回归任务，特

别擅长处理高维数据。其核心优势在于寻找数据类别间的最大间隔边界，增强模型的泛化能力。利用核技巧，SVM 可以高效地处理非线性可分的数据集，映射到高维空间实现分类。它生成的稀疏模型，仅由关键数据点（支持向量）决定，有助于减少模型复杂度，提升计算效率。SVM 模型对参数选择敏感，核函数的挑选和正则化参数的调整对性能有显著影响，且模型的可解释性较弱。尽管在大规模数据集上训练耗时，SVM 因其出色的分类准确率和鲁棒性而广泛应用于各个领域，如生物信息学、计算机视觉和文本处理等。在深度学习流行之前，SVM 是处理复杂模式识别问题的首选算法之一。

2.6.3.2 集成学习模型

集成学习本身并不是一个单独的机器学习算法模型，它是指结合多个机器学习模型来共同完成学习任务的集成模型。本文采用的梯度提升树模型（GBDT）与极限梯度提升树模型（XG Boost）都属于集成学习模型。

梯度提升树（Gradient Boosting Tree）是一个有效的机器学习算法，它能够有效地处理复杂的决策树问题，并且能够有效地解决实际的数据处理任务。

梯度提升树的核心思想是：将许多弱分类器（即决策树）组合成一个强分类器。通过对前面决策树的残差进行训练，就可以构建出更加准确的决策树。在每一轮迭代中，可以使用更加精确的模型来模拟前面决策树所模拟的数据，从而降低残差。这样一层层地拟合，最终得到的就是一个强分类器。

梯度提升树的训练过程可以概括为以下步骤：首先，将数据集分成训练集和验证集。初始化一个决策树，将所有数据都分配到根节点。计算当前模型的预测值和真实值之间的差异，即残差。训练一棵决策树来预测这些残差。在训练过程中，会基于不同的损失函数进行拟合，如平方损失函数、对数损失函数等。将新的决策树与之前的决策树进行结合，从而得到一个更加准确的模型。这里的结合方式可以是加法模型或乘法模型。重复以上步骤，直到得到一个满意的模型为止。

梯度提升树的优点在于它可以处理高维度数据和非线性数据，并且在特征选择方面具有很好的鲁棒性。同时，它也可以通过正则化来防止过度拟合。不过，梯度提升树的训练时间相对较长，而且需要调节的参数相对较多。

极限梯度提升（Extreme Gradient Boosting，XGBoost）是一种梯度提升算法，它通过对传统阶梯提升算法的优化和改进，实现了更高效的提升效果。XGBoost 在原始梯度提升算法的基础上引入了正则化方法，以避免过度拟合。它还使用了一些

技术来加快算法的运行速度，包括特征子抽样、并行处理和缓存优化等。这些改进使得 XGBoost 在许多数据科学竞赛和实际应用中表现优异。具有处理大数据、自动处理缺失值、内置交叉验证、正则化以防过拟合、后剪枝技术和特征重要性评分等特点。XG Boost 通过迭代地训练决策树来逐步提高模型的预测准确率。。每一轮迭代，它都会比较当前模式的预期值与现实值，以此来权衡两者之间的差异，也就是残差。然后，它再训练一棵决策树来预测这些残差。最后，将新的决策树与之前的决策树进行结合，从而得到一个更加准确的模型。总之，XGBoost 是一种具有高性能和灵活性的机器学习算法，XGBoost 已经被广泛应用于各种领域，取得了令人瞩目的成果。

2.6.3.3 深度学习模型

这种技术的发展基于对人工神经网络和人工智能多年的研究，它是机器学习领域的一个重要分支。直到 2006 年，深度学习才被学界正式命名，并开始受到广泛关注，其出现为机器学习和人工智能领域开辟了新的可能性。随着时间的推进，深度学习技术的发展速度显著加快，产生的新型深度学习模型在性能上已大幅超越传统机器学习算法。

深度学习的核心在于人工神经网络，这是一种模仿人类思维过程来分析和学习数据的技术。这一过程称为深度学习。在进行特征提取时，深度学习展现了卓越的能力，能从原始数据中提取更深层次的本质特征，这是传统机器学习方法难以实现的。在传统机器学习中，特征需手动设计，这一过程既复杂又耗时。而深度学习技术可以自动从数据中识别特征，并需要通过设定适当的网络层次结构和神经元数量来优化学习过程。深度学习的多层隐藏层结构使得模型能学习到更为复杂的抽象特征，通过多次线性变换处理，最终目的是提升模型的预测精度。

（1）卷积神经网络（Convolutional Neural Networks，简称 CNN）是深度学习中的一种强大的前馈神经网络，它特别适用于分析视觉图像数据。CNN 通过模仿人类的视觉感知机制来自动和有效地识别图像中的模式，比如边缘、颜色和纹理等。这种网络的设计包含多个卷积层，这些层通过滤波器（或称为卷积核）扫描输入数据并创建卷积特征图，以提取并学习数据的局部特征。随后的池化层（pooling layers）则用于降低特征图的维度，从而减少计算量并提取更加抽象的特征。

CNN 的这些特性使得它在诸多应用中都表现出色，尤其是那些需要处理高维数据并需要模型具备高度自动化特征提取能力的场合。卷积层的参数共享和池化层的

降维策略大大减少了模型的参数数量，使得 CNN 在大规模图像数据集上的运算速度相比其他深度学习算法更快，同时也降低了过拟合的风险。此外，CNN 通过深层结构能够学习到复杂的函数映射，具有很强的函数表达能力，这使得模型在未知数据上也具有良好的泛化能力。

在量化领域，CNN 可以用于自动检测并识别金融时间序列数据中的复杂模式，这是一项对速度和准确度要求极高的任务。给定大量数据和对实时处理的需求，CNN 的快速处理能力和高准确度输出使其成为理想的选择。例如，CNN 可以从市场数据中提取关键特征，识别出潜在的投资机会和风险，从而有效地辅助量化分析师进行决策。因此，CNN 在量化分析研究中的应用，不仅能够提升数据处理速度，还能够提高策略的预测精度，支持更加精准和有效的投资策略。

（2）循环神经网络（Recurrent Neural Network，简称 RNN）是深度学习模型中的一种，特别适用于处理序列数据和时间序列分析。RNN 的核心特性是网络中存在循环，这意味着网络的输出可以反馈到输入端，因此它可以维持一个内部的状态，用于存储和处理序列中的信息。这使得 RNN 在处理序列相关问题时表现出色，如语言建模、文本生成、语音识别、机器翻译等。

RNN 的设计允许它捕捉时间序列数据中的动态信息。在每一个时间步，网络都会根据当前输入和前一时间步的隐藏状态进行预测，这样就能够在序列的不同部分传递信息。而且无论输入序列的长度如何，模型的参数（权重和偏差）在所有时间步上是共享的。这种参数共享机制使得模型能够处理不同长度的序列，并且大大减少了模型的参数量。RNN 还可以处理不同长度的输入序列，也可以产生不同长度的输出序列。这使得 RNN 能够应用于各种各样的任务，如生成固定大小的输出（例如情感分类）或生成与输入长度相匹配的输出（例如机器翻译）。有着对序列数据的处理的天生的能力、模型的灵活性、时间序列数据的深入理解和持久化记忆等特点。

尽管 RNN 有许多优点，但在实践中也面临着一些挑战：在训练过程中，由于序列太长，误差梯度会随着时间的推移逐渐消失（或在某些情况下变得非常大），这使得模型难以学习和保持长期依赖性。并因此导致 RNN 的训练通常需要更长的时间，因为它们必须按顺序处理序列数据，这限制了并行计算的能力。而且尽管理论上 RNN 能够处理长期依赖性，但在实际应用中，标准的 RNN 结构很难真正学习到这些依赖关系。

（3）长短期记忆网络（Long Short-Term Memory，简称 LSTM）是一种特殊类

型的循环神经网络（RNN），专门设计用于解决标准 RNN 在处理长序列数据时遇到的长期依赖问题。在传统的 RNN 中，由于梯度消失或爆炸问题，模型难以维持和学习时间序列中的长期依赖关系。LSTM 通过引入特殊的结构单元-LSTM 单元来解决这个问题。

每个 LSTM 单元包含三个门结构：输入门、遗忘门和输出门。这些门结构允许 LSTM 单元有选择地存储、更新或删除信息，从而有效地维护和传递长期状态信息。输入门控制新信息的进入，遗忘门决定哪些信息应该被丢弃，输出门则控制从单元状态到输出的信息流。

在量化领域，LSTM 模型的这些特性使其成为理想的工具，用于分析和预测金融时间序列数据。量化分析中涉及的数据通常包含复杂的时间动态和非线性模式，这些都是 LSTM 擅长处理的。LSTM 可以捕获股票市场、货币交易等金融时间序列中的长期趋势和周期性模式，帮助量化分析师在制定交易策略和风险管理方面做出更加准确的决策。此外，LSTM 在处理大量历史数据时的高效性，使其在快速变化的金融市场中尤为有用，能够为量化投资提供强大的预测支持。

（4）密集前馈神经网络（Dense Feedforward Neural Network），也被称为多层感知器（Multilayer Perceptron, MLP），是深度学习领域中一种基础且广泛使用的网络结构。这种神经网络的核心特点是其层次结构，包括输入层、一个或多个隐藏层，以及输出层。每一层由多个神经元组成，这些神经元通过加权连接将信息从一层传递到下一层，而没有任何反向或循环的路径，即信息仅在一个方向上流动。

密集前馈神经网络通过非线性激活函数，能够有效地捕捉和模拟数据中的非线性关系。这些激活函数使网络不仅能够处理简单的线性问题，还能学习复杂的非线性模式。多层隐藏层的设计使得这种网络能够学习数据的层次化特征，从而在更深层次上理解和解释数据。随着层级的增加，网络能够从原始输入中提取更加抽象和复杂的特征。

尽管密集前馈神经网络结构简单，但却拥有强大的功能，这主要归功于其多层结构和大量的可训练参数。这些参数（包括权重和偏差）使得网络具有高度的灵活性和适应性，能够适用于各种不同类型的数据和任务。在实际应用中，根据特定问题的需求，可以调整网络的深度（层数）和宽度（每层的神经元数量）。

然而，密集前馈神经网络也存在一些缺点。最主要的是它们容易发生过拟合，特别是在参数数量众多而训练数据相对有限的情况下。此外，这种网络的训练过程

中涉及大量的超参数调整，如学习率、正则化系数、层数和神经元数量等，这使得训练过程可能变得复杂和耗时。还有，训练这样的网络通常需要大量的数据和计算资源。

密集前馈神经网络在许多领域都有广泛应用，例如图像识别、语音识别、自然语言处理、金融预测等。在这些应用中，网络能够从原始数据中提取重要特征，并进行有效的预测或分类。尽管在某些特定任务上，如图像处理或序列数据处理，密集前馈神经网络可能不如其他更专门化的深度学习模型（如卷积神经网络或循环神经网络）有效，但在处理一般性问题时，它仍然是一种非常有价值的工具。总的来说，密集前馈神经网络因其结构上的简洁性、应用的广泛性以及在处理复杂模式方面的能力，成为了深度学习入门和实际应用的重要基石。

第 3 章 因子处理

3.1 数据来源及数据内容

3.1.1 数据获取

初始的股票选取基于沪深 300 指数的全部成分股。鉴于 ST 类股票因业绩亏损可能面临退市的风险，这类股票已从研究中排除以消除潜在偏差；同理，由于金融行业股票的一些指标计算方法与其他行业不同，故从分析中予以排除；另外，考虑到新上市公司股价在上市初年可能会因 IPO 抑价效应而出现不稳定波动，首年数据亦被剔除。

此研究所用的所有原始数据均源自同花顺 iFinD 金融数据服务。通过该服务，可获取股票的历史交易数据，如日交易量、成交金额、平均日价、收盘价、换手率以及公司的股东权益、毛利率、营业收入等信息。此外，同花顺数据库还提供上市公司发布的各类财务报告，包括季报和年报。

3.1.2 确认股票池

本文选取 2011-01-01 至 2022-12-31 时间段的沪深 300 市场的因子数据以及股票收益率数据。2020-09-01 至 2022-12-31 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。股票池以沪深 300 指数为样本空间，依据中证行业分类，分别选取了属于能源、原材料、主要消费、信息技术、医药卫生、通信服务等分类的共计 133 只股票。

截至 2023 年 6 月 21 日，沪深 300 指数共包含 300 只股票，其中股本前十大权重股如图 3-1 所示：

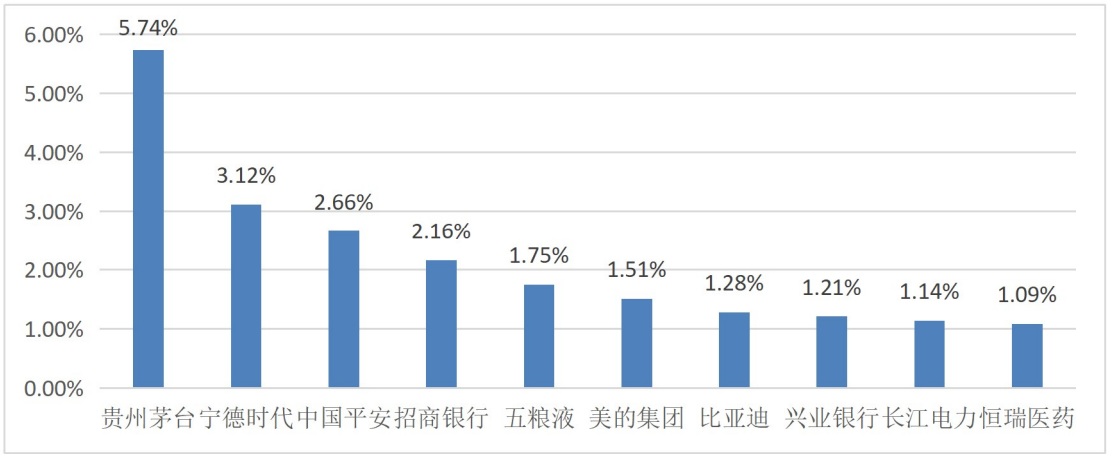

图 3-1 沪深 300 前十大权重股票

数据来源：新浪财经

3.2 候选因子选取

基本面因子的选取和模型建立与上市公司的财务数据息息相关，上市公司的财务数据来源于定期公布的财务报表。财务报表是一家上市公司用来提供信息给投资者和监管者，定期披露财务数据，可以有效解决因为信息不对称，从而使得投资者可以做出更正确的投资决策。

财务报表的组成部分为：现金流量表、资产负债表和利润表。这三张报表体现出一家上市公司整体运营情况和财务状况。其中，资产负债表是这三张报表的重中之重。财务报表的各个组成部分均有自己不同的作用但又有紧密相连的关系，它们的结合反映出了一家上市公司的经营和财务状况。

本文根据各类文献中研究者的观点，对股价可能产生影响的因子分为很多种类，并借鉴 Green et al.（2017），且通过阅读文献选出了研究中国股票市场时使用频率最多的 38 个公司特征变量代理异象因子，本文是对以下六大类指标进行分析：交易摩擦因子、动量因子、价值因子、成长因子、盈利因子、财务流动因子共六大类。在量化投资中这六类因子通过从不同角度评估股票，为投资者提供了全面的决策依据。它们帮助识别交易成本、市场趋势、被低估股票、高成长潜力公司、盈利稳健性以及财务健康状况，从而构建多样化、降低风险并追求超额收益的投资组合。这些因子的综合运用对于发掘投资机会和优化投资策略具有重要意义。其中交易摩擦因子：反映股票交易时的成本和难易程度，对于考虑交易成本和流动性的策略至关重要；动量因子：基于股票价格的历史表现预测未来表现，利用股票的趋势性；价值因子：通过股票的市场价值与其基本面价值的比较，寻找被低估的股票；成长因子：关注公司的成长潜力，如收入增长率，选择具有高成长性的股票；盈利因子：侧重于公司的盈利能力，选择盈利稳健的股票。财务流动性因子：反映公司的财务健康状况和流动性，对于避免选择财务状况较差的公司至关重要。候选因子中的交易摩擦类因子如图 3-1 所示， 完整版详见附录。

表 3-1 部分候选因子

交易摩擦因子	
缩写	因子名称
idvol	特定波动率
vol	总波动率
skew	总偏态
std_turn	交易换手率的波动率
volumed	交易额
std_vol	交易额波动率
illq	非流动性风险
LM	标准化换手率
retnmat	最大日收益率
sharechg	股本增长率
aeavol	收益公告异常交易量

财报数据大部分为季度公布，本文采用季度数据进行了月度填充。由于上市公司财报披露时间存在延迟，填充数据的基本原则是仅在规定的报表全部可用后再进行填充。确保因子的多样性、代表性和实际的预测能力的必要性，选择了 38 个因子进行量化，也为投资决策提供了全面的视角。进行选择时，应考虑因子之间的相关性，避免冗余，确保每个因子都能为模型带来独特的信息价值。同时，也要考虑因子的可行性和数据获取的便利性，确保所选因子能够在实际操作中有效实施。

3.3 数据预处理

在收集因子数据后，发现存在数据不完整、不同数量级和极端值。所以，在实际使用这些数据前，必须进行预处理以确保数据的高质量。数据预处理的步骤包括处理缺失值、剔除极端值、进行标准化和中性化处理等。

3.3.1 缺失值处理

处理数据中的缺失值是数据分析和数据清洗的重要部分。常见的处理缺失值的方法主要有两个：一是删除含有缺失值的行或列，如果数据集很大且缺失值不多，可以考虑直接删除含有缺失值的行或列。如果一个特定的列中有大量缺失值，也可以考虑删除整个列；二是填充缺失值:可以使用 0、空字符串或特定标记等常数来填充。也可以用均值、中位数或众数等统计数据填充，这在数值型数据缺失值处理中比较常见。在时间序列数据中，还可以用前一条或后一条记录的数据来填充缺失值。

本文中使用均值填充初始数据中的缺失值。简单有效而且容易实现，且能有效地处理缺失值问题，还能上保持数据的总体分布和统计特性。

3.3.2 极值处理

在数据处理中，极值（异常值或离群值）可能会对统计分析和机器学习模型产