2022 届硕士专业学位研究生学位论文　学校代码：10269

学　　号：51204407092

華東師範大學

基于 Transformer 的选股因子挖掘

院　　系：经济与管理学部

专业学位类别：应用统计硕士

专业学位领域：应用统计

论文指导教师：於州 教授

论 文 作 者：曹昀炀

2022 年 5 月

MASTER DISSERTATION 2022 UNIVERSITY CODE：10269

STUDENT NO：51204407092

EAST CHINA NORMAL UNIVERSITY

Stock Factor Construction Based on Transformer

College: Faculty of Economics and Management

Major: Master of Applied Statistics

Specialty: Applied Statistics

Advisor: Prof. Zhou YU

Candidate: Yunyang CAO

May of 2022

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于 Transformer 的选股因子挖掘》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：曹明炀　　日期：2022 年　5 月　30 日

华东师范大学学位论文著作权使用声明

《基于 Transformer 的选股因子挖掘》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的研究成果归华东师范大学所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和相关机构如国家图书馆、中信所和“知网”送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（　）1.经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，

于　　年　　月　　日解密，解密后适用上述授权。

（√）2.不保密，适用上述授权。

导师签名　　本人签名曹明炀

2022 年　5 月　30 日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

曹昀炀硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
刘玉坤	教授	华东师范大学	主席
丁帮俊	教授	华东师范大学	
郎大为	高级职称	蔚来汽车销售有限公司	

摘 要

股票收益率的预测是量化交易领域最为关键的任务。多因子模型作为最成熟的股票收益预测模型之一，被广泛应用于股票筛选等任务中。多因子模型的预测准确度依赖于选股因子的质量，传统的因子由人工构造而成，忽视了交易数据的部分重要信息。因此，针对传统因子信息量不足的缺陷，本文提出了基于Transformer 的选股因子挖掘方法，并应用于构造新的选股因子。

本文提出了一种选股因子挖掘的 Transformer 模型框架。首先，本文根据股票交易数据计算日频基本面因子和技术因子，并将这些因子作为输入的特征数据。通过引入市值因子和行业因子，本文利用回归分析计算出股票未来 10 个交易日的超额收益率，并将超额收益率作为预测目标。其次，本文对原始数据中的缺失值、离群值进行处理，并按照交易日对自变量和因变量进行规范化。然后，对规范化后的变量进行嵌入和位置编码使其符合 Transformer 输入维数要求。本文通过引入多头注意力机制对自变量之间和股票之间的信息进行提取，构造出新的选股因子。接着，本文利用多因子模型对 Transformer 解码模块进行改进，使得解码器能够实现从因子到超额收益率的预测。最后，从 2017 年至 2020 年中国 A 股市场的实验结果表明，本文提出的选股因子挖掘的 Transformer 模型在测试集上超额收益率预测值和真实值的相关系数为 0.2579，优于线性回归、深度神经网络、XGBoost 模型，从而表明本文模型能够构造新的选股因子，并且能够基于这些新的选股因子建立多因子模型更准确地预测股票的超额收益率，具有可行性和实用性。

关键词：多因子模型 选股因子挖掘 Transformer 量化选股

ABSTRACT

Prediction of stock returns is the most critical task in the field of quantitative trading. As one of the most mature stock return prediction models, multi-factor models are widely used in stock screening and other tasks. The prediction accuracy of the multi-factor model depends on the quality of the stock selection factors. The conventional factors are manually constructed, ignoring some important information of the transaction data. Therefore, in view of the deficiency of insufficient information of traditional factors, this paper proposes a stock factor construction method based on Transformer to construct new stock selection factors.

This paper proposes a novel Transformer framework for stock selection factor construction. Firstly, the daily frequency fundamental factors and technical factors can be calculated based on the stock trading data, and these factors are considered as the input features. By introducing the market value factor and industry factor, this paper uses regression analysis to calculate the excess return of the stock in the next 10 trading days. The excess rate of return is expressed as the prediction target. Secondly, missing values and outliers are processed in the original data, the independent and dependent variables are normalized according to the trading day. Thirdly, the normalized variables are embedded and positional encoded to satisfy the Transformer input dimension requirements. New stock selection factors are constructed by introducing the multi-head attention mechanism to extract the information between independent variables and between stocks. Fourthly, the multi-factor model are applied in the Transformer decoding module, so that the decoder is able to predict from factors to excess returns. Finally, the experimental results of China's A-share market from 2017 to 2020 show that the proposed Transformer model for stock factor construction has a correlation coefficient of 0.2579 between the predicted value and true value of excess returns on the test set. The results of Transformer are better than linear regression, deep neural network and XGBoost model, which shows that the proposed model in this paper is able to construct new stock factors. And the multi-factor model built by these new stock factors can more accurately predict the excess return of stocks, which shows the proposed model is feasible and practical.

KEY WORDS: Multi-factor model, Stock factor construction, Transformer, Quantitative stock selection

目 录

第一章 导论

第一节 选题背景和研究意义

一、选题背景

股票收益的预测是指投资者根据当前市场和企业披露的信息，试图确定未来某个交易日的股票价值的行为。成功预测股票的未来走势或价格能够为投资者带来丰厚的利润，因此股票收益预测是量化交易领域里最为关键的任务。有效市场假说认为在一个成熟的市场中，股票的价格能够反映当前所有已知的信息，新发布的信息几乎能立即反映在股票价格中，因此任何缺少市场信息的价格预测都存在偏差。尽管这一假说的前提较为严苛，即需要市场的透明度高、功能齐全、公司竞争充分、投资者足够理性，但是它仍然肯定了充足的市场和企业信息对股票价格的变化会产生影响。在此理论基础上，量化交易领域通过归纳有效的市场信息概括出选股因子。这些因子是判断股票收益走势的重要指标，进而被用于股票筛选和投资之中，根据多种因子建立的预测模型被称作多因子模型。

多因子模型作为最成熟的选股模型之一，被广泛应用于收益率预测、股票筛选等场景。该模型建立在资本投资组合、资产定价、套利定价理论等现代投资理论的基础上，利用价格、成交量、公司会计数据、技术分析数据等各类因子对收益率时间序列进行预测。因此，构造出包含有效信息的因子是准确预测收益率的前提和基础，选股因子挖掘是量化交易的必经过程。

二、提出问题

根据上一节所述的背景可知，多因子模型中因子的有效性对股票未来价格的预测准确性有着至关重要的影响。因此如何挖掘出有效的选股因子是本文重点研究的问题。

一个合理可行的多因子模型，需要满足以下两个条件：第一，每一个单因子能够尽可能多地包含有效信息并减少冗余信息，即单因子和超额收益的相关性越高越好；第二，各个因子之间尽可能相互独立，这一条件要求因子之间不存在

冗余的信息，使得每个单因子都能很好反映股票价格的一部分信息。现有研究主要针对日频数据对日频因子进行挖掘，即通过当前时刻的所有信息对未来某一交易日的股票价格进行预测。日频因子尽可能多地代表股票的历史信息，因此本文选择用日频数据构造因子。相比于更高频因子，日频因子更加实用，它既能保证投资者可以进行日频调仓，也保证了模型的预测速度较快。此外，本文选择未来 10 个交易日的超额收益作为预测目标。股票价格时间序列的自相关系数是随时间增加逐渐递减的。如果选择预测未来过短时间的股票价格，会不便于投资者进行较为及时的调仓。如果选择预测未来过长时间的股票价格，会导致预测准确度大幅降低，甚至出现价格走势预测完全相反的情况。因此，本文选择的未来 10 个交易日的超额收益既满足了投资者对调仓频率的需求，也尽可能保证了解决问题方案的可行性。综上所述，如何构建能够准确预测股票未来 10 个交易日超额收益的选股因子是本文的研究重点。

三、研究意义

本文的研究意义主要可以分为理论意义和社会意义两部分。理论意义主要为模型对现有因子的进一步改进和补充，社会意义体现在准确的价格预测能够促进交易充分程度，使得价格股票回归真实价值，从而加速股票市场变得成熟。

在理论意义方面，现阶段的因子挖掘研究仍处在发展阶段。不同的多因子模型层出不穷，特征构造的算法也在不断发展，挖掘出的因子也在不断扩展。本文选择将 Transformer 模型结合到金融时间序列分析领域，弥补经典的神经网络算法在难以预测微小波动的问题，降低选股因子在信息量上的冗余。此外，合理有效的选股因子能够促进多因子模型的改进。股票收益的预测可以分为两个步骤，第一步时根据市场和企业披露的历史数据构建选股因子，第二步时根据选股因子预测未来的股票收益。因此，因子是从历史数据到未来收益的过渡和桥梁，针对不同的因子可以构造出不同的选股模型。本文利用 Transformer 模型在编码模块和解码模块实现根据历史数据构建选股因子的步骤，在模型输出模块实现选股因子对股票未来收益的预测步骤。其中，经过编码的数据可以作为新的选股因子，能够实现对现有因子库的扩充，从而为后续的选股、预测研究提供理论依据。

在社会意义方面，实用可行的因子模型有助于使得市场更加成熟。一方面，准确的股票收益预测能够鼓励投资者进行交易。中国的证券市场非常庞大，为投资者提供了广大的交易平台和众多的投资选择。更多更频繁的交易从宏观上能够促进这个市场加速发展，从微观上也提供更多的投资组合，分担单笔交易的风险。在一定时间内，能够让投资者有利可图，有所回报。以往的研究表明中国的股票市场仍然不成熟，而美国市场相对成熟。这为本文的研究提供了一个契机。本文将成熟市场中总结的理论应用于不成熟市场，从而加速不成熟市场的发展。另一方面，市场的成熟能够引导投资者变得理性，也让投资行为变得合理。现有的研究表明，通过大量的交易，市场中的价格可以近似地反映其内在价值，这可以促进真实的交易变得更加频繁，从而更快地创造社会价值。在过去十年中，投资者在有限的市场内做出了太多不成熟的投资行为，比以往任何时期都多。这些不成熟的投资行为浪费了对社会至关重要的证券的实际价值，导致股票价格波动率太高，波动幅度太大。如果能够在现有的交易市场经验的基础上完善现代投资理论，那么随着物联网、数字货币、虚拟现实等人类技术的发展，交易市场将更加规范化、智能化。如果能够进一步实现交易的规范化和智能化，人们的任何买卖行为都可以在机器的帮助下立即完成，不必担心买卖价格偏离其公平价格。真正实现想买就买，想卖就卖。在这种环境下，社会创造价值的能力也将爆炸式增长。本文所能做出的贡献是尽可能准确地预测价格变化趋势，减少价值预测的偏差，为推进成熟的交易市场提供一些参考依据。

第二节 国内外文献综述

一、基于国外市场的文献综述

近几十年来，大量的选股因子模型和因子挖掘方法被提出。相较于国内市场的研究，国外市场的研究时间更长，研究内容更丰富，构建的多因子选股模型更为多样。Markowitz[10]提出在特定水平的风险下投资组合的期望收益可以基于历史数据进行预测，并引入均值和方差量化收益和风险，建立了资产组合的基本模型。Sharpe 等[11, 15, 16, 17]通过研究收益与风险条件的相互关系，提出了资本资产定价模型（Capital Asset Pricing Model, CAPM）。CAPM 模型认为所有证券的收

益率都与市场证券组合的收益率这一因子存在着线性关系。Ross[13]认为可以利用资产的预期收益和一系列代表系统风险的宏观指标的线性组合预测资产的真实收益，即收益与多个因子线性相关，提出了套利定价理论（Arbitrage Pricing Theory, APT）。然而 APT 并没有指出与证券收益率相关的具体因子，后续学者在上述现代投资理论的基础上，发现了各类影响收益率的因子。Fama 和 French[2]提出 Fama-French 三因子模型，该模型选用上市公司的市值、账面市值比、市盈率三个因子来解释股票收益率的差异。之后，Fama 和 French[3]针对金融市场的发展情况，通过引入盈利因子和投资因子改进三因子模型，提出了 Fama-French 五因子模型。Fama-French 三因子和五因子模型被后人广泛使用，是多因子选股最经典的模型。在之后的研究中，大部分学者仍在 APT、Fama-French 因子模型进行不断地完善。Schmidt 等[14]依据经典模型理论并基于 23 个国家的数据，挖掘出了各个国家的系统风险因子。Guerard 等[7]根据基本面数据和动量数据完善了 Fama-French 三因子模型并建立了股票筛选模型。Feng 等[5]结合了 LSTM 和图神经网络构建了股票收益排名的预测模型，其中 LSTM 用于分别处理单支股票输入的时序数据，图神经网络用于整合 LSTM 处理后股票之间的相互关系并构建选股因子。实验表明该模型在纽约证券交易所和纳斯达克证券交易所的部分数据集上表现优秀。然而该模型需要估计大量的参数，在大规模数据集上的训练速度比较缓慢。Kakushadze[8]在前人研究的基础上基于遗传规划算法挖掘选股因子，总结和归纳了已挖掘的选股因子，并最终发布了 101 个 alpha 因子。

二、基于国内市场的文献综述

国内学者针对中国证券交易市场对选股因子挖掘进行了大量的研究。陈小悦等[21]针对上海交易所股票的收益对 CAPM 模型进行验证，发现中美股市并非完全相同。施东晖[31]基于上证 50 家 A 股数据，发现系统风险对 A 股市场的收益影响较大。贾权等[24]提出流通市值、账面与市场价值等因子对收益率具有很强的解释性。陆静等[30]认为证券流动性风险也是影响价格的因子。与此同时，各类机器学习优化算法也被逐步应用于因子挖掘任务。舒时克等[32]引入弹性网络模型对传统的线性模型进行正则化约束，构建了因子筛选策略。林晓明等[28, 29]通过

自定义函数集，利用遗传规划挖掘日频选股因子。吴先兴等[33]基于遗传规划，利用互信息进一步优化目标函数，重点对股票价量因子进行挖掘。周渐[35]基于 SVM 对沪深 300 成分股的截面数据构造选股模型。各类决策树相关算法[1, 6, 9, 12]也被引入因子构建任务中。曹正凤等[20]将随机森林算法应用于选股任务。李云翔[27]利用 GBDT 模型对表现优秀和一般的成分股进行分类，构造了能够筛选出排名较高的股票的相关因子。霍丽佳[23]利用 AdaBoost 模型对现有因子库的因子进行了进阶的筛选和组合。李想[26]基于 XGBoost 模型从财务、红利、动量、规模、估值、宏观、债券和楼市八个方面构建了共计 307 个选股因子。祝养豹[36]对多种树算法进行了对比，发现 XGBoost 和 LightGBM 在多因子选股任务上的效果优于随机森林、GBDT、AdaBoost。喻术奇[34]结合已知的高频因子和低频因子，利用 LightGBM 模型构建选股因子，但是该方法每期都需要更新大量的模型参数，无法实现即时的大规模交易。陈玄玄[22]将回归问题简化为二分类问题，引入 CatBoost 模型对股票进行分类，筛选出预计超额收益排名较高的股票。李文宇[25]根据收益将股票分为十类，利用神经网络构建分类模型，结果表明该模型能够减少投资组合的最大回撤。Yang 等[19]利用 LSTM 对同一股票的时序量价因子进行挖掘，利用 CNN 对单一时间不同股票间的量价因子进行挖掘。Fang 等[4]对中国 A 股市场不同类型的时序数据分别利用 XGBoost、CNN、RNN、BERT 等深度学习方法构造因子。

第三节 研究内容与研究思路

本文的研究对象为中国 A 股市场股票交易数据，研究目标为预测个股未来 10 个交易日的超额收益率。本文的研究内容分为三个部分。

第一部分是对数据进行清洗、预处理使数据形式符合后续模型的输入要求。本文利用基本面分析、技术分析从原始交易数据中提取人工构造的日频特征数据，如账面市值比、营业额等特征，并将这些特征作为模型所需的自变量。在清洗过程中剔除缺失值和异常值，剔除 ST、PT 股票，剔除每个截面期下一交易日涨停和停牌的股票。在预处理中对离群值进行约束，并按交易日将因变量进行规范化。经过处理后的因变量的各个变量在各个交易日均能处在同一量纲下，为后

续的建模提供了便利。本文的因变量来自于股票未来 10 个交易日的收益率。本文将基于回归分析利用市值因子和行业因子估计股票的预期收益率，将总收益率减去预期收益率得到超额收益率。将超额收益率按照交易日进行规范化，便于模型预测每个交易日上股票的顺序关系。规范化后的超额收益率即为模型所需的因变量。至此，预处理后的数据已经符合模型输入的要求。

第二部分是对经典 Transformer 的改进，使其能够完成选股因子挖掘任务。本文将从内部四个模块分别改进经典模型。在模型输入模块中，将自变量复制一遍形成相同的两组，并分别对两组进行正弦编码和余弦编码，从而实现输入嵌入和位置编码的过程。在编码模块，将一只股票的样本看作为经典模型中的词向量，一批样本看作为经典模型中的一句话，并通过填充掩码完善数据格式。之后模型引入多头注意力机制使得模型能够学习到不同自变量、不同样本之间的关系，有助于提炼信息并构造有效的选股因子。在解码模块，通过模仿多因子模型构造线性层使得模型输出符合套利定价理论。将编码模块的输出结果向量化处理后作为新的选股因子，利用多因子模型实现对股票超额收益率的预测。在模型输出模块，计算未来 10 个交易日的超额收益率真实值和估计值的均方误差，并反向传播给各层网络从而实现 Transformer 中参数的更新策略。至此，选股因子挖掘的 Transformer 模型已经构建完成，后续可以应用于实证分析中。

第三部分是根据实际数据进行实验，并对模型的有效性进行对比分析。本文的数据来源于中国 A 股市场 2017 年至 2020 年的交易数据。在划分数据集方面，本文选择将前两年的数据作为训练集，第三年的数据作为测试集。在经过第一部分提到的数据处理流程和第二部分构建的模型求解后，得到了 Transformer 模型的实验结果。通过建立均方误差、平均绝对误差、相关系数三种评价指标对预测结果进行定量分析。本文将对比经典模型的实验结果，例如线性回归、深度神经网络、XGBoost，判断模型的有效性和实用性。本文还将从交易日的角度评估模型预测的稳定性。

经过上述三部分的研究和分析，本文能够利用改进后的 Transformer 模型完成选股因子挖掘和超额收益预测任务，之后将对研究进行总结和展望。

第四节 创新点

本文构建的基于 Transformer 的选股因子挖掘方法主要有以下三点创新。

第一个创新点是改进模型输入模块，使模型能够分析股票交易数据。经典的 Transformer 模型将单词转化为词向量，它的输入数据为句子中所有词向量构成的矩阵。本文选用一条样本的观测值向量代替词向量，一批样本的观测值作为输入矩阵。通过对输入模块的改进，模型能够对股票交易数据进行特征提取，并且能够对一批样本中股票之间的相关信息进行提炼。

第二个创新点是引入多头注意力机制，让模型自发地构造新的因子。注意力机制能够衡量输入特征之间的关系，多头能够对相同的特征构造不同的映射方式。相比于传统启发式算法从人工定义的函数集合中选取函数构造因子表达式的思想，Transformer 模型能够从多头注意力学习特征之间的相关关系出发，根据相关关系构造非线性的因子表达式。因此，新的因子能够摆脱人工的束缚，由机器对输入的特征进行训练和学习得出。在此基础上，机器所构造的新因子能够扩充现有的选股因子库。

第三个创新点是利用多因子模型改进 Transformer 模型的解码模块，使模型满足因子挖掘任务的需求。一方面，经典 Transformer 模型的解码器必须按照位置顺序将编码后的数据进行还原，但是股票数据不存在复杂的位置关系，因此需要对此限制进行改进。另一方面，模型编码后的数据可以被看作为选股因子，此时解码模块仅需要利用选股因子预测股票的超额收益。而多因子模型不仅能实现从因子到收益的映射，还符合套利定价理论，所以能够使得 Transformer 模型满足选股因子挖掘任务的需求。

第五节 结构安排

本文的组织结构安排如下：

第一章，主要介绍了选股因子挖掘的研究背景，提出研究问题为挖掘新的选股因子。本章分别对国外和国内市场的研究进行归纳和综述，在确定研究对象为中国 A 股市场股票数据之后提出了研究思路和方法。

第二章，主要介绍了选股因子挖掘所需理论基础。本章详细解释了选股因子

可来源于基本面分析、技术分析、机器学习分析三种方法，展示了多因子模型的发展过程，阐述了经典 Transformer 模型内部四个模块各自的原理。

第三章，主要提出应用于选股因子挖掘的 Transformer 模型。本章对自变量和因变量分别进行预处理，对异常值、离群值进行约束，分别从四个模块改进 Transformer 模型，使得编码和解码模块能够完成选股因子挖掘任务，输出模块能够完成股票超额收益预测任务。

第四章，主要基于中国 A 股市场交易数据进行实验分析。本章根据时间顺序划分数据集，构建评价指标，并根据预测结果对 Transformer 和三种经典模型进行定量分析，分别从样本总体和单个交易日的角度评估模型预测的准确度。

第五章，主要概括了全文的流程和结果，提出总结性的结论，并展望了未来的研究方向。

第二章 理论基础

第一节 股票市场分析方法

股票市场预测是试图确定股票在交易所交易的未来价值的行为。股票市场的分析方法常被分为三大类，即基本面分析、技术分析和机器学习分析。这三种方法可以同时使用，并且能够相辅相成。基本面分析和技术分析经常被用于提供多因子模型所需的基本面因子和技术因子，机器学习分析经常被用于辅助人工快速有效地构造因子。

一、基本面分析

股票市场中基本面分析的目的是找出股票的真实价值，然后将其与市场上交易价格进行比较，从而判断当前市场上的股票是被低估还是被高估。相对于技术分析，基本面分析注重长期的策略，它的主要目标是估计一个公司未来的所有利润，在将未来的利润贴现到当前时刻之后，对预估价值和实际价格进行对比分析。基本面分析的思想是人类社会需要资本才能取得进步，如果一家公司经营良好，那么它会得到额外的资本回报，并导致其股价显著提升。因为基本面分析是最合理、最客观的，并且是可以从公开可用的信息，比如公司财务报表中得出的，所以它被以基金经理为代表的分析师群体广泛使用。

基本面分析关注的对象是与股票直接相关的公司。分析师会评估这家公司过去的业绩、会计数据、信用。许多衡量公司业绩的比率都是通过基本面分析构造的，比如市盈率，这些比率能够帮助分析师评估股票的有效性。除了比率之外，基本面分析还包括公司的基本信息和会计数据，比如市值、行业等。通过基本面分析得到的因子被称为基本面因子，这一类因子是多因子模型的重要组成部分。

二、技术分析

技术分析通过利用过去价格的趋势来估计股票的未来价格。相对于基本面分析，技术分析并不考虑公司任何的基本面。技术分析师通常还会绘制图表来表述历史价格时间序列的变化以及预测未来的发展趋势。技术分析师会计算和分

析各类量价指标，例如指数移动平均值（Exponential Moving Average, EMA）、柱状垂直线图、支撑位、阻力位、动量等指标。

技术分析更多地应用于短期策略，而非长期策略。由于股票交易市场价格波动幅度较大，波动频率较快，因此难以像预测传统时间序列一样预测股票价格长期变化规律，这也是导致技术分析经常与短期调仓相结合的重要原因。在这种条件下，因为交易员关注的是短期价格波动，所以技术分析在大宗商品和外汇交易市场中的应用更为普遍。

为了保证技术分析的有效性和实用性，市场需要遵从三个基本假设：第一是关于一家公司的所有重要信息都已经被定价到了它的股票中；第二是价格会按照趋势中移动；第三是由于市场心理，历史价格往往会重复出现。尽管这三个假设在实际市场中难以完全地实现，但技术分析仍能为多因子模型提供大量的技术指标。

三、机器学习分析

随着计算机的出现，股票市场分析和预测进入了新的技术领域。在机器学习分析中，最常用的方法是遗传算法（Genetic Algorithm, GA）和神经网络（Neural Network, NN）这两大类分析方法。一方面对于遗传算法，学者们不断发展，通过各类启发式算法和规划方式不断挖掘选股因子。这类因子的优势在于能够完整地写出表达式，并且一部分因子可以联系实际得到基于经济学理论的解释。另一方面，神经网络被认为是一种优秀的数学函数逼近器，能够挖掘各式各样的因子。前馈神经网络是用于股市预测最常见的形式，它利用误差的反向传播来更新网络权重。与此同时，循环神经网络（Recurrent Neural Network, RNN）和时延神经网络（Time-Delay Neural Network, TDNN）也常应用于预测股票的时间序列数据。

机器学习分析的任务可以分为两类。第一类是分类任务，对未来股票的买入、卖出、不操作这三种行为进行预测。第二类是回归任务，对未来股票的超额收益进行回归拟合。相比于回归任务，分类任务更为简单。现有的实验表明，神经网络在分类任务上的效果更加优秀，但是分类任务无法构建实用可行的选股因子，

因此回归任务仍是机器学习分析中不可或缺的一部分。本文选择完成回归任务，进一步完善现有多因子模型在预测准确度上的不足。

机器学习分析的输入数据可以是股票的历史价格，也可以是公司的财务数据，甚至可以是基本面分析和技术分析后得到的因子。因此机器学习分析经常和基本面分析、技术分析一同进行。本文将延续这种思路对现有的日频数据和日频因子建立选股因子挖掘和超额收益预测模型。

本文将基于上述三种分析方式，选择合适的交易数据、基本面因子、技术因子作为模型的自变量，利用深度学习算法完成对选股因子的挖掘。

第二节 多因子模型

多因子模型是一种常用的金融模型，它利用多个因素组成的数学表达式来解释证券市场现象和均衡资产价格。多因子模型通过比较两个或更多因素来分析变量之间的关系以及由此产生的收益变化。

一、资本资产定价模型

资本资产定价模型（CAPM）[15]是金融市场试图为证券定价并由此确定资本投资的预期收益的理想化描述。该模型提供了一种量化风险并将风险转化为预期股本回报率的方法。CARM 给定风险时资产的预期收益计算表达式如下：

$$E(R_i) = R_f + \beta_i\big(E(R_m) - R_f\big), \tag{2.1}$$

其中，$E(R_i)$表示资产 i 的预期收益，R_f表示无风险收益，R_m表示市场收益，$E(R_m) - R_f$表示市场风险溢价。β_i表示资产 i 的β值，它的表达式如下：

$$\beta_i = \frac{Cov(R_i, R_M)}{Var(R_M)}. \tag{2.2}$$

根据公式(2.2)可以看出β_i为资产 i 预期收益和市场收益的协方差除以市场收益的方差。β值是一种关于投资组合风险与市场风险相似程度的度量。如果一只股票的风险比市场风险更大，那么它对应的β值将会大于 1。反之如果一只股票的风险比市场风险要小，那么它对应的β值小于 1。CAPM 虽然比较简洁，但是它提出了β值并表明投资组合的超额收益来自于系统风险和非系统风险，为后续的研

究提供了理论依据。

二、套利定价理论

套利定价理论（APT）[13]是一种多因子资产定价模型，从价值投资角度来看，APT 是分析投资组合实用有效的工具，能够识别暂时被低估或高估的证券。它的基本思想是，资产未来的收益可以通过使用资产的预期收益与表示系统风险的一组宏观经济变量的线性组合来预测。根据这一思想，可以表示出 APT 模型中基础的因子模型，相应的表达式如下：

$$r_i = a_i + \sum_{j=1}^{K} b_{ij}F_j + \varepsilon_i, i = 1,2,\cdots,N, \tag{2.3}$$

其中， r_i表示资产 i 的预期收益，a_i为常数截距项，b_{ij}为资产 i 在因子 j 上的因子载荷，也可以称作是因子 j 上资产 i 的因子暴露程度，F_j表示因子 j，ε_i表示随机误差并满足$E(\varepsilon_i) = 0, i = 1,2,\cdots,N$。$N$ 为资产的总数，K 为因子的总数。进一步地，可以将因子模型简化成矩阵形式，表达式如下：

$$r = a + BF + \varepsilon, \tag{2.4}$$

其中，$r = (r_1,\cdots,r_N)^T$表示 N 个资产收益构成的列向量，$a = (a_1,\cdots,a_N)^T$表示截距列向量，$B = \left(b_{ij}\right)_{N\times K}$表示因子暴露度矩阵，$F = (F_1,\cdots,F_K)^T$表示 K 个因子构成的列向量，$\varepsilon = (\varepsilon_1,\cdots,\varepsilon_N)^T$表示随机误差向量。

公式(2.4)即为多因子模型最基本最简洁的表达式。多因子模型又被称作为多因素模型，它用一组因子的线性组合表示投资组合的未来收益。从表达式上来看，多因子模型与线性回归模型非常近似，需要预测的未来收益r可以看作因变量，构造完成的一组因子F可以看作是自变量，此时截距项a和因子暴露矩阵B可以看作是需要求解的参数。因此，在给定因子的情况下，未来收益的预测问题可以转化为回归拟合问题。

为了进一步说明多因子模型的运作机制，本文以 Fama-French 三因子模型[2]为例对因子F进行具体的介绍。Fama-French 三因子模型是最经典的多因子模型之一，它通过在市场风险因子中引入规模风险和价值风险因子来扩展经典的 CAPM。Fama 和 French 发现了价值型股票和小盘股经常跑赢市场平均水平这一现象。他们通过引入两个新的因子，使得模型对两类股票表现优秀的现象进行了

修正。Fama-French 三因子模型的表达式如下：

$$R_{it} - R_{ft} = a_{it} + b_1\left(R_{Mt} - R_{ft}\right) + b_2 SMB_t + b_3 HML_t + \varepsilon_{it}, \qquad (2.5)$$

其中，R_{it}表示投资 i 在时刻 t 的总收益，R_{ft}表示在时刻 t 的无风险收益，R_{Mt}表示时刻 t 的市场投资组合总收益，$R_{it} - R_{ft}$表示期望的超额收益，因子$R_{Mt} - R_{ft}$表示市场投资组合的超额收益，因子SMB_t表示时刻 t 的市值投资组合收益率（Small Minus Big），因子HML_t表示时刻 t 的账面市值比投资组合收益率（High Minus Low），ε_{it}表示投资 i 在时刻 t 的随机误差并且满足$E(\varepsilon_{it}) = 0$。

根据上述公式，Fama-French 模型包含三个因子，即公司规模、账面市值比和市场超额收益。Fama-French 三因子模型仅仅利用三个因子就完成了对股票超额收益的预测，为后续的多因子模型研究提供了长久的参考意义。随着市场的发展，经典的三因子模型难以满足人们对预测准确度更高的需求。为了进一步提升对超额收益$R_{it} - R_{ft}$的预测准确度，本文将延续多因子模型的思路，利用 Transformer 模型构造并计算合适的因子组合F，并且在模型输出模块建立线性回归方程完成对股票超额收益的预测。

第三节 Transformer 模型

Transformer 是一种采用自注意力（Self-attention）机制的深度学习模型，它对输入数据的各部分进行差分加权，是一种常用的提取特征方法。Vaswani 等[18]提出 Transformer 模型并率先将其应用于机器翻译领域。在之后的研究中，Transformer 被广泛地应用于自然语言处理和计算机视觉两大领域中，并取得了目前最高水平的效果。本文利用 Transformer 在提取特征时的优势，将其应用于股票因子挖掘任务中，并预测股票未来的超额收益。

经典 Transformer 模型的示意图如图 2.1 所示。图中左侧的灰色矩形框内为模型的编码（Encoder）模块，右侧的灰色矩形框内为模型的解码（Decoder）模块，两个模块的下方为模型的输入数据，两个模块的上方为模型的输出结果。由于经典 Transformer 解决的是自然语言处理里的翻译任务，因此输出的是属于一组词向量的概率。本文将依次从模型输入、编码模块、解码模块、模型输出四个方面介绍 Transformer 模型的具体理论。

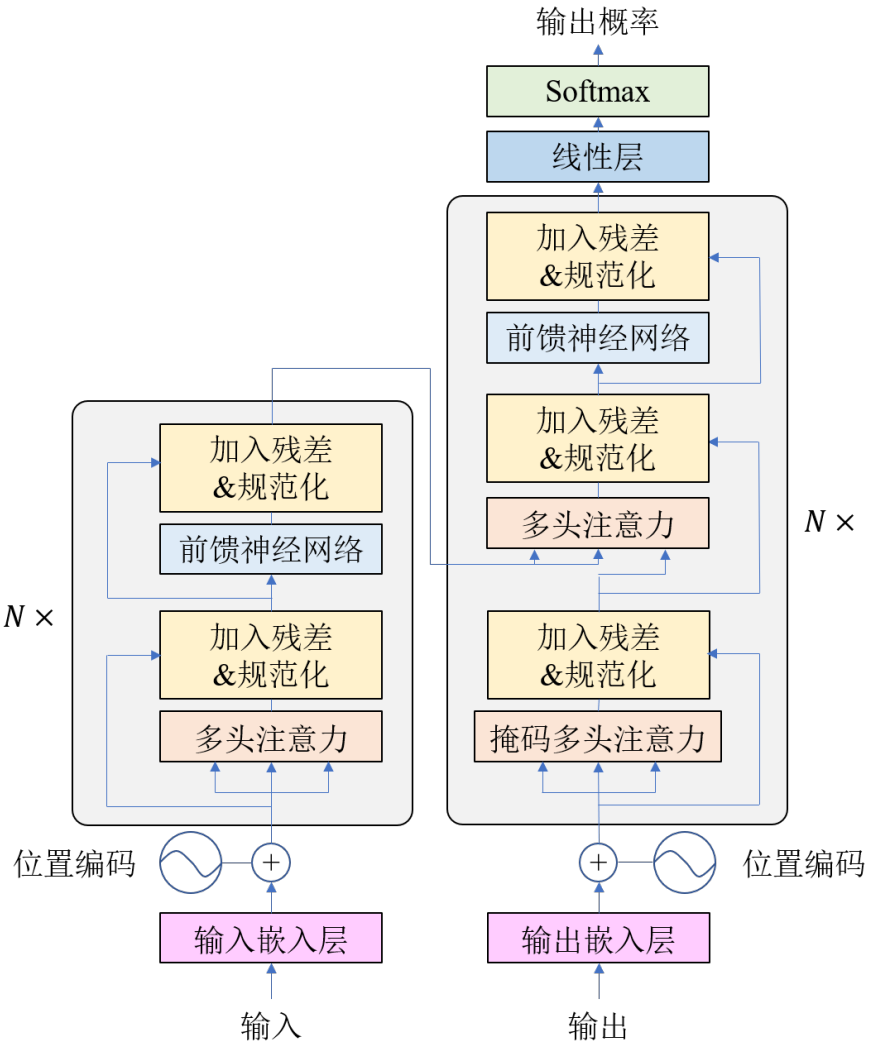

图 2.1 Transformer 模型示意图

一、模型输入

经典 Transformer 模型的输入数据为序列数据，如图 2.1 中下方模块所示。在自然语言处理任务中，输入的数据为一段话分词后的词向量，分词这一变化即为从输入数据变为输入嵌入层的过程。在本文的因子挖掘任务中，输入的数据为不同股票不同时间的日频特征数据。为了后续位置编码的便利，本文将输入的特征复制一遍，完成从输入数据到输入嵌入层的过渡。

位置编码（Positional Encodings，PE）是为了提取不同词语或特征之间相关关系。位置编码前的位置关系可称为绝对位置，编码后的位置关系可称为相对位

置。传统的时序模型是按数据出现的前后顺序进行训练和预测的，所以这些模型数据之间的相关性是随着时间增加而递减的。在实际任务中，两个特征即使绝对位置相距较远仍可能存在较近的相对位置。例如，在机器翻译任务中，后一句话中某一代词可能指代上一句话的主语，但此时代词和主语名词的绝对位置距离较远，传统的时序方法可能会因为忽略它们之间的相对关系而导致翻译结果指向性错误。在因子挖掘任务中，输入的特征是同等重要的，在建模前难以根据绝对位置而判断特征之间的相关关系。因此需要利用位置编码弱化其原有的位置关系，构建新的相对位置，为提取任意两个变量之间的相关关系提供理论可能。Transformer 所使用的位置编码是正弦编码和余弦编码，编码表达式如下：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \tag{2.6}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \tag{2.7}$$

其中，pos为样本所在的绝对位置($pos = 0,1,2,\cdots$)，在机器翻译任务中d_{model}表示词向量的维度，i表示词向量中的第i维，此时$2i$和$2i+1$分别表示特征的奇偶性。在本文根据实际任务需求对上述变量含义略作更改。因子挖掘任务中d_{model}表示输入的特征数量，即输入的日频基本面因子、技术因子、机器学习构造因子的总数。在将原始特征复制成两组后，对其中一组的绝对位置做正弦编码，对另一组做余弦编码，从而实现表达式(2.6)和(2.7)中$2i$和$2i+1$对应的奇偶性关系。

在此之后，将输入嵌入层的数据和位置编码进行相加，从而实现模输入数据的预处理和相对位置关系的提取。输出和输出嵌入层的变换方式与输入和输入嵌入层同理，因此同理可以完成这一部分的嵌入和位置编码。根据上述变换操作，可以完成 Transformer 模型输入模块的内容，为后续的编码和解码模块提供数据来源。

二、编码模块

编码是指将输入数据映射到潜在空间的变换方式，在潜在空间中机器更容易处理、分析、归纳数据的信息和关系。在选股因子挖掘任务中，编码也可以被

看作是构造因子的过程，编码后的数据可以作为新的选股因子。Transformer 的编码模块由 N 个相同的编码器堆叠组成，即上一个编码器的输出结果会作为下一个编码器的输入数据。其中 N 的取值可以为任意正整数，在机器翻译任务中，通常取 N 为 6。每一层编码器都还有两个子层，即图 2.1 中的多头自注意力机制以及全连接前馈神经网络。在每个子层之后都会引入一个残差块，在对子层的输出结果加上这一残差矩阵之后再规范化。换言之，每一个子层的输出结果的表达式可以写作$\text{LayerNorm}\big(x + \text{Sublayer}(x)\big)$，其中$\text{Sublayer}(x)$表示子层的计算结果。为了实现从输入数据到各个编码器之间的连接，模型中所有的子层和嵌入层的维度均为d_{model}。下面将介绍多头注意力和前馈神经网络两个子层。

注意力（Attention）是指模型对信息的关注程度，是自然语言处理领域的一个基本概念。一个效果优秀的模型应当对重要信息进行重点关注并充分学习。在 Transformer 中，注意力可以理解为模型对某一或某些特征信息的重点关注。注意力函数可以表示为查询 Quiry 和一组键 Key、值 Value 到输出值的映射，其中查询 Quiry、键 Key、值 Value, 输出值均为向量。输出可以看作是 Value 的加权求和，其中每个权重由查询 Quiry 与相应键 Key 的乘积计算得出。因为每一个样本的每个特征都至少包含两个维度，例如词向量就至少包含两个维度，所以按对应位置维度将向量查询 Quiry、键 Key、值 Value 拼接起来可以得到三个矩阵，分别将它们记为 Q, K, V。那么，注意力的计算过程可以分为以下四个步骤。

步骤一：将 Q 和 K 中对应的向量进行点积，计算每个特征之间的相关性得分 $score$，矩阵表达式为$score = Q \cdot K^T$。

步骤二：将上述得分进行归一化，使得训练时不会出现梯度变化过大的现象。矩阵表达式为$score^* = score/\sqrt{d_k}$，其中$d_k$为矩阵 K 的维度。

步骤三：利用 softmax 函数将得分映射到区间[0, 1]内。

步骤四：对值 Value 进行加权平均，权重为步骤三处理后的分数。矩阵表达式为 $\text{softmax}(score^*) \cdot V$。

根据上述步骤，可以将注意力函数 Attention 矩阵形式的表达式完整地写出来，表达式如下：

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.8}$$

在上述注意力机制的基础上可以引入多头注意力（Multi-Head Attention）机制的概念。相比于单头注意力，多头注意力能够让模型关注不同的位置，并且给出多个潜在子空间。在多头注意力机制中，模型会对每个头设置独立的权重矩阵，因此训练出的查询 Quiry、键 Key、值 Value 均不相同。多头注意力可以看作是单头注意力的加权拼接，矩阵形式的表达如下：

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)W^O, \tag{2.9}$$

$$\text{head}_i=\text{Attention}\left(QW_i^Q,KW_i^K,VW_i^K\right), \tag{2.10}$$

其中，h表示多头注意力头的个数，$W_i^Q\in\mathbb{R}^{d_{model}\times d_k},W_i^K\in\mathbb{R}^{d_{model}\times d_k},W_i^V\in\mathbb{R}^{d_{model}\times d_v},W^O\in\mathbb{R}^{hd_v\times d_{model}}$ 均为参数矩阵并且满足 $d_k=d_v=d_{model}/h$。$\text{Concat}(\text{head}_1,\cdots,\text{head}_h)$表示将内部的矩阵$head_i(i=1,2,\cdots,h)$依次按列拼接。至此多头注意力机制子层的基本内容均可实现，下面将介绍前馈神经网络子层。

每个编码器不仅包含注意力子层，还包含按照位置计算的全连接前馈神经网络子层。编码模块中每个编码器的全连接前馈神经网络的框架都是一样的，它们均选用了两个线性变换和一个 ReLU 激活函数，表达式如下：

$$\text{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2, \tag{2.11}$$

其中，W_1,W_2,b_1,b_2均为需要训练的参数，并且同一层线性变换的参数均相同，不同层之间的参数是不同的。

根据上述变换和映射，可以完成 Transformer 模型编码部分的运算过程和结果，这些结果将用于解码模块的一部分输入数据。

三、解码模块

解码是指将潜在空间内的数据还原成初始空间的中的数据，并且在还原的同时可以完成回归拟合或分类等任务。如图 2.1 所示，和左侧的编码模块类似，右侧的解码模块同样含有 N 个依次堆叠的相同的解码器。每个解码器含有三个子层，除了和编码器类似的多头注意力层和前馈神经网络层，解码器还包含了掩码多头注意力（Masked Multi-Head Attention）层。和编码器同理，在每个子层里

都包含了引入残差矩阵和规范化。在掩码多头注意力层中，为了满足编码器仅已知当前的信息以及不知道未来信息的条件，需要对部分数据进行掩码。在实际的操作中，需要对每一个序列进行一个额外的掩码变换。操作方法是生成一个方形矩阵，并且矩阵的维度大于任意一条序列的长度，上三角位置的值均为 0，其余位置的值为一个很大的负值或$-\infty$，将其作用在每一个序列上，即可完成掩码。

解码模块的具体计算过程与编码模块基本一致，至此可以完成模型主体部分的编码和解码中的多次映射，并且实现了选股因子挖掘领域中构建因子的任务。

四、模型输出

模型输出是指将模型提取的特征映射到实际任务需求的过程。根据不同的任务背景，模型输出的结果也会不同。经典 Transformer 解决的是机器翻译任务，因此模型输出模块如图 2.1 上方所示，先将提取的特征进行线性变换，再利用 Softmax 将其映射成概率。通过字典可以将输出概率最大的单词作为模型预测的实际结果。

然而，在选股因子挖掘的背景下，并不需要输出一个向量中每一个元素对应的概率，仅需要输出股票未来的超额收益。因此，本文在实际的建模中改进了经典的 Transformer 模型，将解码模块输出的结果作为构造的因子，并将结果矩阵拉直成一个向量。之后，以向量中的元素为输入数据，对超额收益进行线性回归，即可完成模型的输出部分。相比于经典 Transformer 模型，选股因子挖掘领域最终的线性层的输入维度更大，并且输入维度与编码器的数量有关，因此在参数设置中，需要对编码器的个数进行限制。

上述内容即为经典 Transformer 模型的理论介绍和其在选股因子挖掘领域需要做的改进。下一章将结合实际数据，对 Transformer 模型进行改进，分析并求解选股因子。

第三章 选股因子挖掘的 Transformer 模型

第一节 数据预处理

本文的研究对象为中国 A 股市场的股票基本信息和日频交易数据，研究目标是预测个股未来 10 个交易日的超额收益率。本文首先对一部分股票的异常值进行剔除，即剔除 ST、PT 股票，剔除每个截面期下一交易日涨停和停牌的股票。其次本文对训练数据集和测试数据集的时序长度进行约束，利用前两年的数据集作为训练集，之后一年的数据作为测试集。因此，挖掘出的因子可用于次年的选股策略，等到次年结束时再重新对选股因子进行更新。本文的自变量为基本面分析和技术分析得到的日频因子，因变量为未来 10 个交易日的超额收益。在这一节里，本文将从自变量和因变量两个部分介绍预处理流程。本节的数据预处理仅为从原始交易数据到 Transformer 输入数据的过程，Transformer 输入模块的嵌入和位置编码将在下一节介绍，根据实际数据的实验操作细节将在下一章进行补充。

一、自变量预处理

本文的样本数据来源于中国 A 股市场剔除异常值后的数据，原数据仅包含交易正常，所属公司不存在问题的股票。本文的第二章第一节股票市场分析方法中提到了基本面分析和技术分析，根据证券公司从业者在这两种主要分析方法上的研究，本文将初始的逐笔交易数据转化为日频交易特征数据，这些特征包含了常用的基本面因子、高频技术因子、低频技术因子三类共计 50 个因子特征。下面将从这三个类别介绍特征数据。

在所有的因子中，基本面因子的数量占比最高，共有 33 个。基本面因子包括四种账面市值比（book-to-market ratio）特征，即内在价值与市值比、研发费用与市值比、研发销售管理总费用与市值比、留存收益与市值比。基本面因子还包含了收益增长率、收益与价格比率、营运收入、季度盈利增长率、季度盈利与价格比率、季度营运收入、收入增长等特征。这些基本面因子关注的对象是与股票直接相关的公司。这些因子能够评估公司过去的业绩、会计数据、信用，尤其

能够为长期的预测提供较高的参考价值。

技术因子是根据股票的历史交易数据计算得出，根据用于计算的数据的频率又可以将技术因子细分为高频因子和低频因子。如果数据的频率不低于日频，例如分钟频、30 分钟频、日频等，则可以将计算求解的因子归纳为高频技术因子；如果数据的品读低于日频，例如年度频、季度频、月频等，则可以将计算求解的因子归纳为低频技术因子。本文共计算得出了 14 个高频技术因子，包括了价格和成交量的相关系数、价格的偏度、盘前 30 分钟时间的成交量占比、最后 30 分钟的成交量占比、开盘后 60 分钟的成交量占比等特征。同时，本文也计算获得了 3 个低频技术因子，即年度营业额、月度营业额、季度营业额。其中的年月并不是自然年和自然月，而是去年相应日期到今天、上个对应日期到今天的营业额，因此这个低频技术因子的取值也是每天变化的。

在计算获得所有的因子特征之后，本文对各个因子进行标准化。标准化的方式为按照日期将观测值减去对应因子的均值后除以对应因子的标准差，表达式如下：

$$\tilde{x}_{jt} = \frac{x_{jt} - \bar{x}_{jt}}{\mathrm{std}(x_{jt})}, \tag{3.1}$$

其中，x_{jt}为交易日 t 特征 j 初始的取值，$\tilde{x}_{jt}$为交易日 t 特征 j 标准化后的取值，$\bar{x}_{jt}$为交易日 t 所有股票特征 j 的取值的平均值，$\mathrm{std}(x_{jt})$为交易日 t 所有股票特征 j 的取值的标准差，j 为特征编号，t 为日期编号。标准化能够消除取值过大或过小的离群值，有助于模型预测的稳健性。此外，本文按照日期进行标准化有利于对不同的交易日在同一量纲下比较股票之间的排名。因为后续的预测是对每一只股票每一个交易日进行预测，所以按照日期标准化不会因为某一个交易日行情的好与坏而导致参数训练产生偏向性。在此自变量预处理的基础上，后续模型能够在不同交易日对比各只股票的排名，为筛选业绩表现优秀的股票提供了数据基础，也为准确预测超额收益提供了参考依据。

二、因变量预处理

本文选择的因变量为未来 10 个交易日的超额收益。因变量超额收益是本文

Transformer 模型的预测目标，也是判断挖掘的选股因子是否有效的重要依据。因为自变量经过了标准化变换，所以因变量也需要相应地进行合理的规范化约束。以往的基本面分析研究表明，公司的市值和行业是影响股票收益较大的两个因子。对于市值较大股票，其收益的涨跌幅度一般较大。因此本文会根据市值因子对股票超额收益进行规范化。此外，对于不同的两个行业，它们的股票的涨跌趋势一般并不同步。而对于同一行业中的各只股票，可能出现股价同时上涨或同时下跌的情况，为了进一步比较行业内的各只股票，本文还将根据行业因子对股票超额收益进行规范化。在实际的选股投资中，投资不同行业的风险是不相同的，因此常常会对各个行业买入的股票数量进行约束，比如每个行业至少买入一只股票。在这种情况下，建立的多因子选股模型需要对行业内的股票进行细化分析，所以需要构造能够对比行业内股票的因子。基于上述对市值因子和行业因子的分析，本文将未来 10 个交易日的收益率对市值和行业回归的残差作为超额收益，表达式如下：

$$\frac{price_{t+10}}{price_t} = c_1 G_1 + \sum_{m=1}^{M} d_m H_m + r_\alpha, \tag{3.2}$$

其中，$price_t$表示时刻 t 的股票价格，$price_{t+10}$表示时刻 t 未来 10 个交易日的股票价格，G_1表示股票的市值。H_m表示股票行业的独热编码，用于判断股票是否属于行业 m，如果股票属于行业 m，那么$H_m = 1$，如果股票不属于行业 m，那么$H_m = 0$。r_α表示股票的超额收益，为本文模型预测的对象。M为股票市场行业的总数，c_1和$d_m (m = 1,2,\cdots,M)$均为系数。公式(3.2)其实为多因子模型公式(2.3)的进一步推算形式，本文将市值和行业两个因子单独取出来用于回归预测，基于该回归方程的收益率的估计值可以看作为预期收益率。结合公式(3.2)和公式(2.3)表达式，本文未来 10 个交易日的收益率计算公式表达如下：

$$\frac{price_{t+10}}{price_t} = c_1 G_1 + \sum_{m=1}^{M} d_m H_m + a + \sum_{j=1}^{K} b_j F_j + \varepsilon, \tag{3.3}$$

其中，G_1, H_m, F_j均可以被看作为选股因子，a为截距项，c_1, d_m, b_j为参数，ε为随机误差，$m = 1,2,\cdots,M$，$j = 1,2,\cdots,K$。进一步地，可以将收益分为由市场、行业影响产生收益r_β以及由其他因子产生的收益r_α两部分，给定时刻 t 两部分收益

的形式可以改写为$r_{t\alpha}$和$r_{t\beta}$，两者的表达式如下：

$$r_{t\alpha} = a + \sum_{j=1}^{K} b_j F_j + \varepsilon, \tag{3.4}$$

$$r_{t\beta} = c_1 G_1 + \sum_{m=1}^{M} d_m H_m . \tag{3.5}$$

此时，未来 10 个交易日的收益率可以表示为两项之和，简化后的计算公式如下：

$$\frac{price_{t+10}}{price_t} = r_{t\alpha} + r_{t\beta}, \tag{3.6}$$

其中，$r_{t\beta}$表示时刻 t 由市场、行业影响产生的收益，这一部分为股票的客观数据，不需要人工挖掘，因此本文将这一项单独计算作为预期收益。$r_{t\alpha}$表示时刻 t 由其他因子影响产生的收益，本文将这一项作为股票在时刻$(t+10)$的超额收益，并且将其作为后续 Transformer 模型预测的目标。

基于上述预处理操作，已经初步完成了对超额收益率的计算。通过消除市值因子和行业因子的影响，能够尽可能地减少收益率出现离群值的情况。尽管如此，但在实际的股市中仍难以避免一只股票价格的骤增或骤减，比如股票初次发行，企业严重失信等情况都会造成股价的剧烈变动，进而导致计算后的超额收益率存在离群值。为了进一步规范化因变量中的离群值，本文按照日期对超额收益率进行修正，大于 95%分位数的收益率均替换为 95%分位数，小于 5%分位数的收益率均替换为 5%分位数，表达式如下：

$$y_{it} = \begin{cases} r_{t\alpha,0.05}, & r_{it\alpha} < r_{t\alpha,0.05}, \\ r_{it\alpha}, & r_{t\alpha,0.05} \le r_{it\alpha} \le r_{t\alpha,0.95}, \\ r_{t\alpha,0.95}, & r_{it\alpha} \le r_{t\alpha,0.95}, \end{cases} \tag{3.7}$$

其中，$r_{it\alpha}$表示股票 i 在时刻 t 的超额收益率，$r_{t\alpha,0.05}$表示在交易日 t 的所有股票超额收益率的 5%分位数，$r_{t\alpha,0.95}$表示在交易日 t 的所有股票超额收益率的 95%分位数，y_{it}表示股票 i 在时刻 t 修正后的超额收益率，也是本文模型的因变量。与自变量同理，因变量同样从交易日日期层面进行规范化。本文按照日期修正离群值有利于在近似的量纲下对不同交易日的股票进行排名。因为后续的预测是对每一只股票每一个交易日进行预测，所以按照日期标准化不会因为某一个交易日某一只股票行情的好与坏而导致参数训练产生偏向性。至此，可以得到模型

的因变量为修正后的未来 10 个交易日的超额收益率y，以及它在时刻 t 股票 i 上的取值y_{it}。

基于上述数据预处理，可以得到本文用于建模的自变量和因变量，自变量为中国 A 股市场股票去除离群值并且规范化后的日频因子特征数据，因变量为修正后的未来 10 个交易日的超额收益率。

第二节 模型框架改进

Transformer 是一种采用自注意力机制的深度学习模型，本文的第二章第三节对经典的 Transformer 模型原理进行了介绍。传统的 Transformer 模型被率先应用于机器翻译领域[18]，之后也被经常用于自然语言处理领域。为了利用 Transformer 在提取特征时的优势，本文将其模型框架进行改进，使其能够应用于股票因子挖掘任务中，并完成对股票未来的超额收益率的预测。改进后的 Transformer 模型示意图如图 3.1 所示。

经典 Transformer 模型的可以分为模型输入、编码模块、解码模块、模型输出四个模块，本文将分别从这四个模块介绍对 Transformer 模型框架的改进。

一、模型输入

在模型输入模块，本文对经典 Transformer 的输入没有进行较大的改进，仍然沿用了位置编码这一处理方式，框架示意图如图 3.1 最下方所示。相比于经典 Transformer 中输入嵌入层将词语转化为词向量，本文将一只股票在某一交易日的所有自变量的集合看作是一个向量，不同的股票看作是不同位置的词语。与经典模型不同的是，本文的自变量均为有效的信息，如果直接进行位置编码会导致信息没有充分被利用，所以需要将输入的自变量转换为嵌入输入层。本文将原本的 50 个日频因子特征数据依次复制一遍，使其成为一个 100 维的向量，之后对其进行位置编码。若初始的向量为(特征 1，特征 2，…，特征 50)，则经过复制变换后的向量为(特征 1，特征 1，特征 2，特征 2，…，特征 50，特征 50)。

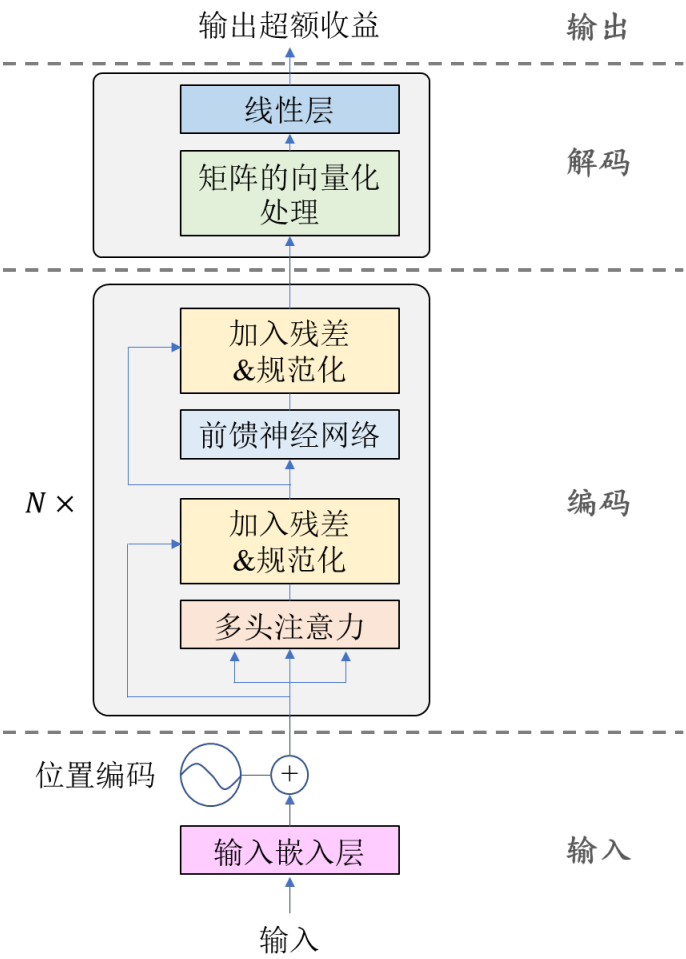

图 3.1 选股因子挖掘的 Transformer 模型示意图

在上述嵌入操作下，复制前后的两组数据会分别进行正弦编码和余弦编码。此时，向量的维度d_{model}的取值为 100。位置编码中另一个参数绝对位置pos与输入的一批数据的尺寸（batch size）有关，本文选取 512 作为一批数据的尺寸。在这种参数设置下，位置编码的表达式如下：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{i}{50}}}\right), \tag{3.8}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{i}{50}}}\right), \tag{3.9}$$

其中，pos为样本所在的绝对位置$(pos = 0,1,2,\cdots,511)$，i表示特征的位置索引$(i = 0,1,2,\cdots,49)$，此时$2i$和$2i+1$分别表示特征位置索引的奇偶性，分别对应

复制前后的两组相同的因子。在选股因子挖掘任务里，位置编码能够关联股票之间的信息，同时也可以作为正则化项对模型进行约束。

根据上述操作，本文能够完成选股因子挖掘的 Transformer 模型的输入模块的参数和编码设置，将自变量转换为编码模块需要的输入形式，为后续的模型提供了数据基础。

二、编码模块

在编码模块，选股因子挖掘的 Transformer 模型框架如图 3.1 中下方第二个模块所示。改进后的模型与经典模型十分相似，仍然将编码器分为多头注意力层和前馈神经网络层两个子层，并且对每个子层输出的结果加上残差矩阵并规范化。每一个子层输出结果的数学表达式可以写作$\mathrm{LayerNorm}\big(x+\mathrm{Sublayer}(x)\big)$，其中$\mathrm{Sublayer}(x)$表示子层的计算结果。为了实现各个子层之间的连接，需要满足编码器中所有子层的维度均和输入嵌入层的维度相同并且确保每一层输入的 x 的维度相同。根据上一小结模型输入部分的分析，所有子层的维数应当为d_{model}，即 100 维。下面将分别从两个子层的角度介绍改进后的 Transformer 模型编码模块。

首先，在多头注意力子层中，模型通过不同的头训练出不同的查询 Quiry、键 Key、值 Value 矩阵，从而实现对不同位置不同因子的注意力多样化。一方面，头的数量越多能够尽可能使模型关注更多的输入特征，有利于构造更多的有效因子。另一方面，头的数量越多考虑同一个输入特征的次数就越多，也会造成输出因子之间相关系数较高，导致因子之间存在冗余的信息量。所以需要对多头注意力头的数量进行约束。在第二章第三节中，本文曾记多头注意力头的个数为 h，此处选取多头注意力的头数 h 为 2。本文选取 h 为 2 有以下两方面原因：一方面是本文在自变量预处理中已经将初始特征数量增大一倍并实现了不同的位置编码，这一操作已经导致输入的维度d_{model}较大，因此不宜设置较大的 h。另一方面是多头注意力的效果仍比单头注意力要完善，多头更能够有效地挖掘新的选股因子，同时当头的个数为 2 时，并考虑到之前的自变量预处理，实际的特征已经被模型采样了 4 遍，足以实现因子挖掘对输入信息量的需求。基于上述分析，

本文选择双头注意力机制。此时的多头注意力子层的表达式可以表示为如下的矩阵形式：

$$\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2)W^O, \tag{3.10}$$

$$\mathrm{head}_i=\mathrm{softmax}\left(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right)VW_i^K, \tag{3.11}$$

其中，多头注意力头的索引 i 的取值为 1 或 2。$W_i^Q\in\mathbb{R}^{100\times50}, W_i^K\in\mathbb{R}^{100\times50}, W_i^V\in\mathbb{R}^{100\times50}, W^O\in\mathbb{R}^{100\times100}$均为参数矩阵。$d_k$为矩阵 K 的维度，取值为$d_{model}/h=100/2=50$。Q,K,V分别表示按对应位置维度将向量查询 Quiry、键 Key、值 Value 拼接起来可以得到三个矩阵，根据上一小节模型输入模块的改进，绝对位置的总数与输入的一批数据的尺寸相等，为 512。因此三个需要训练的矩阵为$Q\in\mathbb{R}^{512\times100}, K\in\mathbb{R}^{512\times100}, V\in\mathbb{R}^{512\times100}$。此时注意力每个头的维度为$\mathrm{head}_1\in\mathbb{R}^{512\times50}$，$\mathrm{head}_2\in\mathbb{R}^{512\times50}$。$\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2)$表示将头对应的矩阵按列从左至右依次合并起来，因此合并后的维度为$\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2)\in\mathbb{R}^{512\times100}$。进一步可以算出多头注意力子层的输出维度为$\mathrm{MultiHead}(Q,K,V)\in\mathbb{R}^{512\times100}$。

基于上述对多头注意力子层的分析，需要满足训练的三个矩阵Q,K,V的维度为512×100。但是在实际的建模过程中，可能会出现输入的一批数据的样本数量不足 512 个的现象，尤其是数据集中的最后一批数据通常少于 512 个样本。此时需要对输入的数据进行掩码操作，将样本填充至 512 个。这一操作被称作为填充掩码（padding mask）。本文将每一批样本填充至 512 个，并且所有新样本的观测值均填充为 0。经过填充掩码，本文满足了矩阵Q,K,V维度的不变性。

其次，在前馈神经网络子层中，选股因子挖掘的 Transformer 模型仍选用全连接前馈神经网络，并且变换方式仍为线性变换，激活函数选择 ReLU 函数，表达式如下：

$$\mathrm{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2, \tag{3.12}$$

其中，矩阵$W_1\in\mathbb{R}^{100\times100}, W_2\in\mathbb{R}^{100\times100}$，行向量$b_1\in\mathbb{R}^{1\times100}, b_2\in\mathbb{R}^{1\times100}$中的元素均为需要训练的参数，$x\in\mathbb{R}^{1\times100}$为某一样本在前馈神经网络子层输入的数据，形式为行向量。$\max(0,a)$为 ReLU 激活函数的表达式，若向量中元素值小于 0，则输出的向量对应位置的元素为 0，其余位置的元素取值保持不变。因为同一层

线性变换的参数均相同，不同层之间的参数是不同的，所以分别对各个样本按公式(3.12)计算，最后将结果按行拼接起来。根据公式(3.12)，每个样本输出的结果$\mathrm{FFN}(x) \in \mathbb{R}^{1\times 100}$，将这些 512 个样本的结果行向量按原始顺序依次拼接，可以得到一个512×100维的矩阵。因此前馈神经网络子层的输入和输出维度具有一致性，均为一个512×100维的矩阵。

基于上述对多头注意力和前馈神经网络两个子层的描述，可以完成一个编码器的内部框架。在 Transformer 模型的编码模块共有 N 个编码器依次相连，每个编码器的输出结果都会作为解码器的输入数据。对于一条样本，一个编码器能够输出一个 100 维的行向量，即 100 个因子，所以 N 个编码器一共能构造$N \times 100$个因子。换言之，改进后 Transformer 模型能对某一交易日中的某一只股票计算出$N \times 100$个因子观测值。在实验部分，本文将对编码器的个数参数 N 进行具体的设置。

综合上述改进分析和计算公式，改进后的 Transformer 模型编码模块可以完成选股因子挖掘的一部分任务，并能为后续解码部分提供维度合适的数据。

三、解码模块

解码模块在选股因子挖掘中的任务是将编码后潜在空间上的数据还原至初始空间，并完成超额收益率回归拟合。解码模块的示意图如图 3.1 中上方第二个模块所示。从图中可以看出本文并没有选用经典 Transformer 模型的解码框架，主要有以下四点原因：第一点为经典解码器利用掩码多头注意力使得词语按出现顺序进行预测，但本文的样本顺序不需要这种操作；第二点是掩码能够让先前预测的结果作为后续预测的输入，但本文不需要股票之间相互预测；第三点是编码器的训练结果已经可以作为因子，不需要额外的网络训练；第四点是解码器选用线性模型符合现代投资理论中多因子模型理论特点。解码模块包括矩阵的向量化处理和线性模型两部分，下面将分别介绍这两部分。

矩阵的向量化处理是指将输出的矩阵拉直为向量。在上一小节编码模块的介绍中，一条样本经过 N 个依次堆叠的编码器的输出结果为$N \times 100$维的矩阵。本文将其拉直为一个维数为$100N$的行向量，并整理出对应的数据格式。经过向

量化处理，一条样本对应$100N$个变量，即某一交易日一只股票对应$100N$个选股因子。后续则会根据这些因子建立多因子模型。

线性层指的是多因子模型，即从多个因子到超额收益的线性映射。线性层的表达式与公式(2.3)近似，具体的表达式如下：

$$y_i = a_i + \sum_{j=1}^{100N} b_{ij}x_{ij} + \varepsilon_i, \tag{3.13}$$

其中，i 表示样本的编号，y_i表示样本 i 对应的股票超额收益率，x_{ij}表示样本 i 第 j 个选股因子的观测值，a_i表示常数截距项，b_{ij}表示回归系数，也可以看作为因子 j 在样本 i 对应股票上的因子暴露度，ε_i表示随机误差并满足$E(\varepsilon_i) = 0, i = 1,2,\cdots,B$，$B$ 为样本的总数。x_{ij}为经过由矩阵拉直成的向量中的元素，即本文模型构造的选股因子。y_i为因变量预处理后超额收益率，为本文模型的预测目标。

相较于经典 Transformer 模型，本文在解码模块的改进较多。改进后的解码模块选用 APT 中基础的多因子模型，更符合现代投资理论，为证券从业者后续扩充选股因子库，筛选优质股票提供参考依据。此外，相较于卷积神经网络和梯度提升树算法等常用的机器学习方法，选用线性回归的运算速度更快，有助于更频繁地训练参数，能够满足更高频率的交易需求。

四、模型输出

模型输出是指将解码的结果转化为实际任务所需数据的过程。经典的 Transformer 需要对词语的概率进行估计，所以选用了 softmax 变换并选取概率最大的词语作为输出结果。然而本文的任务是对股票的超额收益进行回归分析，所以输出的结果不是概率而是超额收益率的估计值。改进后模型输出的示意图如图 3.1 最上方所示。

在第三章第一节中，本文对因变量进行了预处理，得到预测目标为超额收益。因此 Transformer 模型的任务是对股票的超额收益进行回归拟合。在上一小节提到的线性层中，模型能够利用多因子模型计算出超额收益y的估计值$\hat{y}$。估计值$\hat{y}$也是模型输出部分的结果。

为了计算梯度并将其反向传播回编码器和解码器的各个网络中，模型需要

定义损失函数。本文选择均方误差（Mean Square Error，MSE）作为损失函数，损失函数可以表示为$MSE(y,\hat{y})$。之所以选择 MSE 作为损失函数，是有两方面的原因的，一方面它能够使得模型对超额收益的取值更敏感，而以相关系数为代表的评价指标对收益的排名更敏感；另一方面它可以求导并且降低模型的运算时间，而以评价绝对误差（Mean Absolute Error，MAE）为代表的评价指标不易求导而造成运算时间大幅增长。关于评价指标具体的公式介绍将在本文第四章进行讨论和分析。

根据上述模型输入、编码模块、解码模块、模型输出四个部分的介绍，本文基本完成了对经典 Transformer 模型的改进。改进后的 Transformer 模型即能够利用编码器挖掘选股因子，也可以根据解码器中的线性层实现对股票未来 10 个交易日超额收益率的预测。关于数据预处理和建模方法的介绍至此基本完成，在下一章本文将结合实际数据对模型进行实验，并对输出结果机型对比分析。

第四章 选股因子挖掘实验与结果分析

第一节 数据来源

本文的数据来源于中国 A 股市场 2017 年至 2020 年的交易数据。按照第三章第一节的预处理流程，本文剔除了 ST、PT 股票，剔除每个截面期下一交易日涨停和停牌的股票。对交易数据进行基本面分析和技术分析后，得到 50 个日频因子特征数据，即模型的 50 个输入变量。此外，本文基于原始的交易数据，通过因变量的预处理，计算出未来 10 个交易日的超额收益率，并将其作为预测目标。经过处理之后，一个交易日中的一只股票可以看作为一条样本，一条样本又包含了 50 个自变量和 1 个因变量。

在划分数据集方面，本文选择将前两年的数据作为训练集，第三年的数据作为测试集。为了方便模型在训练时调整参数，本文又将训练集进行细分，将前 7 个季度的数据作为细分后的训练集，第 8 个季度的数据作为验证集，本文根据验证集上的预测结果调整超参数。以 2017 年 1 月 1 日至 2019 年 12 月 31 日的交易数据为例，2017 年 1 月 1 日至 2018 年 9 月 30 日的数据为训练集，2018 年 10 月 1 日至 2018 年 12 月 31 日的数据为验证集，2019 年 1 月 1 日至 2019 年 12 月 31 日的数据为测试集。训练集和验证集共含有样本 1409394 个，测试集共含有样本 829604 个。

本文将在同一数据集和同一种划分形式下比较经典的机器学习和本文提出的选股因子挖掘 Transformer 模型。

第二节 评价指标

本文选取的评价指标共有三种，分别为均方误差（MSE）、平均绝对误差（MAE）、Pearson 相关系数。其中 Pearson 相关系数指标又细分为两种，第一种为对训练集或测试集上所有超额收益的真实值和估计值计算相关系数，第二种为对某一交易日内所有股票超额收益的真实值和估计值计算相关系数，并按照交易日时间顺序绘制出相关系数时间序列图。以下将从数学表达式的角度分别介绍这三种评价指标。

MSE 为机器学习领域最常用的评价指标之一，也经常被应用于损失函数之中。由于 MSE 可以求导方便优化的性质，选择 MSE 作为损失函数通常能减少部分机器学习模型的训练时间。为了统一比较各种方法，本文中所有对比模型的损失函数均采用 MSE。MSE 的表达式如下：

$$MSE = \frac{1}{N-1}\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{it} - \hat{y}_{it})^2, \tag{4.1}$$

其中，$N = \sum_{t=1}^{T} n_t$表示样本的总数，n_t表示交易日 t 内股票的总数，T表示交易日的总天数，y_{it}为超额收益率的真实值，$\hat{y}_{it}$为超额收益率的估计值，$t = 1,2,\cdots,T$表示交易日的编号，$i = 1,2,\cdots,n_t$表示交易日 t 内股票的编号。MSE 取值越小表示模型的预测准确度越高。

MAE 表示估计值与真实值之差的绝对值的平均，能够反映估计值的偏差程度。本文中样本 MAE 与 MSE 的计算类似，仍计算样本全体的 MAE，因此自由度仍为$N-1$。MAE 的表达式如下：

$$MAE = \frac{1}{N-1}\sum_{t=1}^{T}\sum_{i=1}^{n_t}|y_{it} - \hat{y}_{it}|, \tag{4.2}$$

其中，$|y_{it} - \hat{y}_{it}|$表示y_{it}与$\hat{y}_{it}$之差的绝对值，其余变量的解释与 MSE 相同。同样地，MAE 取值越小表示模型的预测准确度越高。

除了 MSE 和 MAE，在选股因子挖掘领域常选用相关系数作为评价指标。相关系数能够反映两个向量之间元素排序的关系。常被用于评估预测准确度的相关系数有 Pearson 相关系数和 Spearman 相关系数。由于 Spearman 相关系数需要计算各个样本取值的排名，所以它的运算时间相对较长，而且不易于比较样本全体的估计值和真实值的相关性。因此本文仅选用 Pearson 相关系数。为了弥补没有排名的缺陷，本文分别对样本全体和每一个交易日的两组数据计算 Pearson 相关系数。样本全体的真实值和估计值的相关系数表达式如下：

$$\rho_{all} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{it} - \bar{y})(\hat{y}_{it} - \bar{\hat{y}})}{\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{it} - \bar{y})^2 \sum_{t=1}^{T}\sum_{i=1}^{n_t}(\hat{y}_{it} - \bar{\hat{y}})^2}}, \tag{4.3}$$

其中，$\bar{y}$表示所有样本超额收益率真实值的均值，$\bar{\hat{y}}$表示所有样本超额收益率估

计值的均值，(4.3)式分子部分表示真实值和估计值序列的协方差，分母部分表示二者的标准差之积。对于一个数据集，只能计算出样本全体的相关系数ρ_{all}。

ρ_{all}从总体的角度衡量了预测的准确度，下面将从每一个交易日微观的角度分析模型预测的准确度和稳定性。单个交易日真实值和估计值的相关系数ρ_t表达式如下：

$$\rho_t = \frac{\sum_{i=1}^{n_t}(y_{it} - \bar{y_t})(\hat{y}_{it} - \bar{\hat{y}_t})}{\sqrt{\sum_{i=1}^{n_t}(y_{it} - \bar{y_t})^2 \sum_{i=1}^{n_t}(\hat{y}_{it} - \bar{\hat{y}_t})^2}}, \tag{4.4}$$

其中，$\bar{y_t}$表示交易日 t 股票超额收益率真实值的平均值，$\bar{\hat{y}_t}$表示交易日 t 股票超额收益率估计值的平均值，$t = 1,2,\cdots,T$表示交易日的索引，T 表示数据中含有的交易日的总天数。对于一个数据集，能够计算出 T 个单个交易日真实值和估计值的相关系数，即$\rho_1, \rho_2, \cdots, \rho_T$。相关系数的取值越大表示模型的预测准确度越高。上述 T 个相关系数能够构成一组时间序列，本文后续将通过绘制时序图评判模型的预测准确度。

至此本文已经完成了三种模型评价指标的构建，分别为均方误差 MSE、平均绝对误差 MAE、Pearson 相关系数ρ_{all}和$\rho_t(t = 1,2,\cdots,T)$。后续将根据 MSE 进行超参数的调整，并根据所有构建的评价指标对多种机器学习因子挖掘方法进行结果对比分析。

第三节 参数设置

在第三章改进的 Transformer 模型的介绍中，已经根据数据实际情况对部分参数进行了设置，例如多头注意力头的数量设置为 2，一批样本的数量为 512，一条样本对应向量的维度为 100。除此之外，Transformer 还包含了两个重要的参数，一是编码器的数量 N，二是模型的迭代次数。编码器的数量 N 决定了模型最终构造的因子数量，模型的迭代次数则决定着模型是否能够避免欠拟合或过拟合。本文将分别对两个超参数进行调参，下面将结合训练集和验证集上的实验分析参数设置的理由。

编码器的数量决定着构造的因子数量，如果编码器的数量为 N，那么构造的

因子的数量为 100*N*，因此不宜设置过大的 *N*。本文在保证模型预测准确度的情况下，需要尽可能选择较小的 *N*。本文首先设定 *N* 的可能取值为 1，2，3，4，5，6 共六个取值，分别对六个取值下训练集和验证集上的 MSE 进行分析。此外，在这一参数选择过程中，模型设置了以下两种终止迭代的条件：第一，若训练集上最近 10 次迭代 MSE 的均值与前 11 至 20 次迭代 MSE 的均值相差不超过 1×10^{-7}，则提前结束迭代终止训练。第二，若迭代次数到达 500 次，则终止迭代。在这种限制条件下，模型能够尽可能地在不同编码器数量的情况中都获得较充分的训练，从而有利于比较各个参数下模型预测结果的 MSE。表 4.1 为设置不同编码器数量 *N* 模型在训练集和测试集上的 MSE。

表 4.1 编码器数量 *N* 与 MSE 汇总表

编码器数量 *N*	训练集 MSE	验证集 MSE
1	0.9501	0.9648
2	0.9367	**0.9557**
3	0.9854	1.0032
4	0.9732	0.9998
5	0.9311	0.9997
6	0.9775	0.9976

从表 4.1 可以看出，当编码器数量 *N* 为 2 时，验证集的 MSE 最小，为 0.9557。此外，当编码器数量大于 2 时，训练集的 MSE 也会增大，说明了较大的 *N* 不能带来较好的预测效果。并且编码器数量越大，模型中需要训练的参数会成倍增长，训练时间也相应地变长，因此需要尽可能避免选择较大的 *N*。此外，上文中也提到较小的 *N* 也能使构造的因子更加精炼。综合上述两方面考虑，本文选择设置编码器数量 *N* 为 2。一方面，当 *N* 为 2 时，模型的预测效果较好。另一方面，此时 Transformer 模型构造的新因子的数量为 200 个，能够满足多因子模型的基本需求。

根据上述分析，已经将超参数编码器数量 *N* 设置为 2，在此基础上对迭代次数进行调参。为了得到更准确地结果，本文对学习率进行了动态设置。将迭代次数等分为三个阶段，分别对三个阶段的学习率进行设置：第一个阶段的学习率为 1，第二个阶段的学习率为1×10^{-2}，第三个阶段的学习率为1×10^{-4}。如果迭代

次数的符号表达记为n_{iter}，那么第一个阶段为前$\lfloor n_{iter}/3 \rfloor$次迭代，第二阶段为第$(\lfloor n_{iter}/3 \rfloor + 1)$次至第$2\lfloor n_{iter}/3 \rfloor$次迭代，第三阶段为剩余次数的迭代，$\lfloor n_{iter}/3 \rfloor$表示迭代总次数除以 3 并向下取整。例如，如果迭代总次数n_{iter}为 100 次，那么第 1 次至第 33 次的学习率设置为 1，第 34 次至第 66 次的学习率设置为1×10^{-2}，第 67 次至第 100 次的学习率设置为1×10^{-4}。在上述学习率设置策略的基础上，本文分别设置迭代次数为 50，100，150，200，250，300，比较训练集和验证集上的 MSE，从而确定最优的迭代次数。图 4.1 记为不同迭代次数下训练集和验证上的 MSE 折线图。

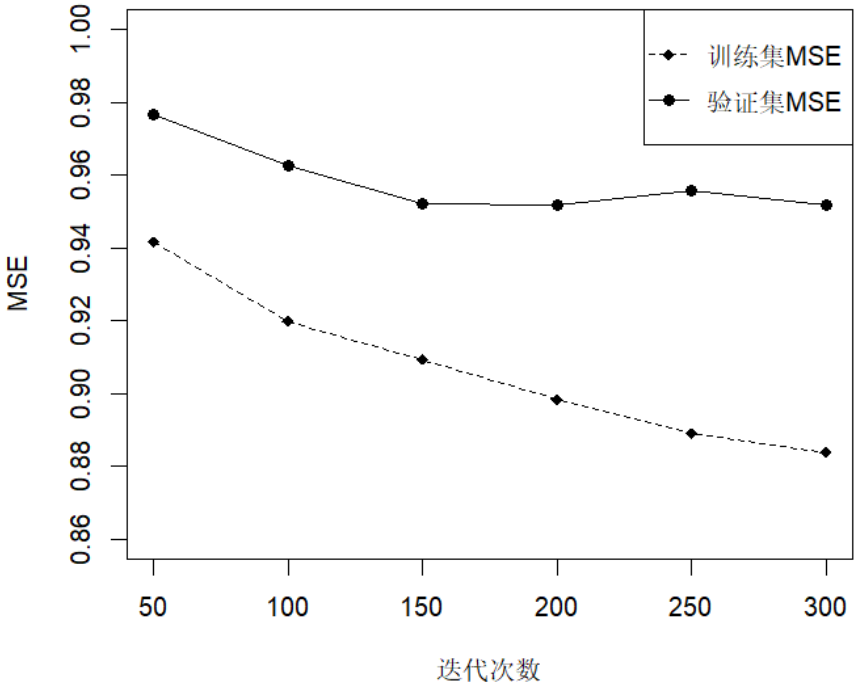

图 4.1 选股因子挖掘的 Transformer 模型示意图

从图 4.1 中可以看出，当迭代次数大于 150 时，训练集的 MSE 仍然会降低，但验证集上的 MSE 基本保持不变。此时继续进行迭代会出现过拟合现象。因此本文将参数迭代次数n_{iter}设置为 150。

根据上述参数设置的实验和分析，本文对改进后的 Transformer 模型设置编码器的数量为 2，迭代次数为 150。本文之后的分析将在此参数设置的基础上对提出的模型和经典模型进行定量的对比分析，并检验模型的有效性。

第四节 结果对比分析

本文选择线性回归、深度神经网络（Deep Neural Network，DNN）、XGBoost 作为本文的对比算法。由于本文的自变量数据均为基本面因子和技术因子，因此线性回归方法能够代表人工构建因子的预测准确度，可以作为基准水平。DNN、XGBoost、Transformer 三种方法均代表机器构造因子的预测准确度。其中 DNN 和 XGBoost 作为最常用于选股因子挖掘的方法，在实际中已有着广泛地应用，并且 DNN 可以作为神经网络方法的代表，XGBoost 可以作为树算法的代表。本文将用改进后的 Transformer 模型对这两类常用的方法进行比较。下面本文将介绍对比算法 DNN 和 XGBoost 的参数设置。

对于 DNN 模型，本文选择建立全连接前馈神经网络，并设置隐藏层的数量为 5。其中每一个隐藏层均为线性层，五个线性层的神经元的个数分别为 256，128，64，32，1。前四个线性层的结果都需要经过激活函数进行非线性变换，激活函数均选择 ReLU 函数。最终第五个线性层的输出结果即为股票未来 10 个交易日超额收益率的估计值。

对于 XGBoost 模型，本文选择网格搜索寻找最优的一组超参数。XGBoost 模型在验证集上表现最好的一组超参数作为最终选择的超参数，即使得验证集上 MSE 最小的一组超参数为最优超参数。本文重点对下列 4 个超参数进行网格搜索：第一，迭代次数 n_estimators 可能的取值为 50，100，150，200，250；第二，最大树深 max_depth 可能的取值为 6，8，10；第三，学习率 learning_rate 可能的取值为 1，0.1，0.01，0.001；第四，叶子上最少包含的样本数量 min_child_weight 可能的取值为 5，10，15。通过对 180 种可能的参数组合进行网格搜索，得到最终的超参数取值，如表 4.2 所示。

表 4.2 XGBoost 最优超参数取值

超参数名称	含义	取值
n_estimators	迭代次数	200
max_depth	最大树深	10
learning_rate	学习率	0.1
min_child_weight	叶片上最小样本数	10

XGBoost 其余重要的超参数设置如下：colsample_bytree 为 1，colsample_bytree 为 1，subsample 为 1，eval_metric 为 rmse。尽管 XGBoost 的损失函数为均方根误差（Root Mean Square Error，RMSE），其公式为 MSE 的算术平方根，实质上与 MSE 十分类似。因此在损失函数层面上，可以认为 XGBoost 和其他方法是在均等条件下建模的。

此外，用于对比的线性回归方法为带截距项的最小二乘回归。本文改进的 Transformer 模型的超参数的设置如本章第三节的内容所示，设置编码器的数量为 2，模型的迭代次数为 150，多头注意力头的个数为 2，一批样本的数量为 512，一条样本对应向量的维度为 100。

基于上述各个模型的参数设置，本文将计算 MSE、MAE、Pearson 相关系数三个指标的取值，并以此比较各个模型的预测准确度。其中 MSE、MAE 的取值越小，模型预测准确度越高；相关系数的取值越大，模型的预测准确度越高。线性回归、DNN、XGBoost、Transformer 四种模型在训练集和测试集上的评价指标取值如表 4.3 所示。

表 4.3 四种模型的 MSE、MAE、相关系数汇总表

模型名称	训练集			测试集		
	MSE	**MAE**	**相关系数**	**MSE**	**MAE**	**相关系数**
线性回归	1.5022	0.9718	0.2486	1.5424	0.9794	0.2286
DNN	0.9722	0.7802	0.2587	0.9745	0.7665	0.2349
XGBoost	0.8361	0.7223	0.4326	0.9483	0.7563	0.2302
Transformer	0.9213	0.7598	0.2816	**0.9333**	**0.7498**	**0.2579**

表 4.3 的结果显示 Transformer 模型测试集上的 MSE 为 0.9333，MAE 为 0.7498，Pearson 相关系数为 0.2579。根据表 4.3 可以发现在测试集上本文改进的 Transformer 模型在三个评价指标上的表现均为最优，尤其在相关系数上本文模型的准确度有着显著的提升。本文在训练 Transformer 模型时仅选用 MSE 作为损失函数，而预测结果在三个指标上均有提升也表明了模型具有一定的普适性。此外，DNN、XGBoost、Transformer 三种机器学习方法的预测准确度均明显优于线性回归，这也表明了机器学习分析方法能够挖掘有效的选股因子，具有一定的可行性和实用性。XGBoost 模型在训练集上表现较好，但在测试集上的表现明显

不如训练集，存在过拟合的问题。

上述分析均基于数据集样本全体的评价指标，可以发现本文改进后的选股因子挖掘 Transformer 模型的预测效果最好。为了从微观的角度进一步分析各个模型的预测准确度，本文计算了每个交易日股票真实值和预测值的 Pearson 相关系数。相关系数取值越大表示模型预测准确度越高。如果每个交易日的相关系数均较高且取值变化较为平稳，说明模型的预测效果准确且稳定。本文计算了每个交易日的相关系数，并按照交易日期绘制出了相关系数时间序列的折线图。图 4.2 表示训练集上单个交易日相关系数的时序图，图 4.3 表示测试集上单个交易日相关系数的时序图。

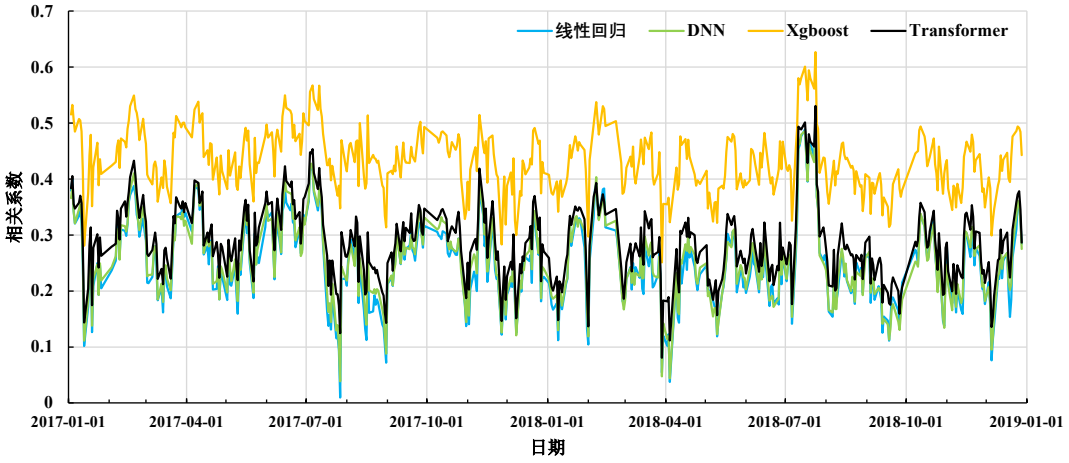

图 4.2 训练集上单个交易日相关系数的时间序列图

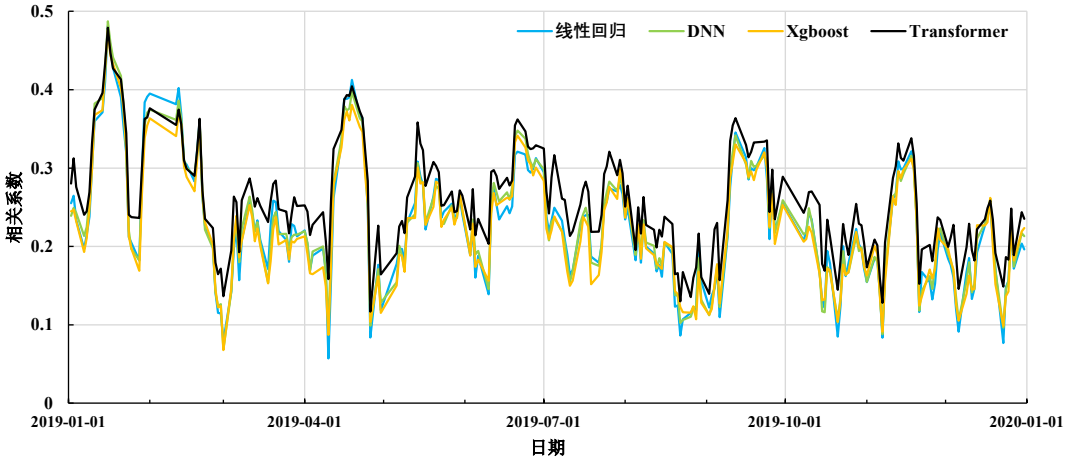

图 4.3 测试集上单个交易日相关系数的时间序列图

在图 4.2 和图 4.3 中，蓝线表示线性回归的结果，绿线表示 DNN 的结果，黄线表示 XGBoost 的结果，黑线表示 Transformer 的结果。从图 4.2 中可以看出，训练集上 XGBoost 模型表现最好，其次为本文改进的 Transformer 模型。从图 4.3 中可以看出，测试集上本文改进的 Transformer 模型表现最好，剩余三种模型表现一般。结合两张时间序列图可以发现，Transformer 模型的预测准确度较高并且比较稳定，相关系数显著优于简单线性回归。线性回归的结果代表着人工构造的因子的有效性程度，Transformer 模型预测结果优于线性回归既说明了经过 Transformer 模型编码和解码模块构造的因子优于人工构造的因子，也表明了 Transformer 模型够有效地构造新的选股因子，为扩充因子库提供参考依据。此外，从这两张图中还可以看出经过网格搜索调参的 XGBoost 模型明显存在过拟合现象，DNN 模型结果略优于线性回归但不如 Transformer 模型。

根据上述对比实验和定量分析，本文提出的选股因子挖掘 Transformer 模型能够构造新的选股因子，并且基于新因子建立的多因子模型能够更准确地预测股票的超额收益率。相比于经典的线性回归、DNN 以及 XGBoost 模型，本文的 Transformer 模型预测结果更准确更稳定，具有可行性和实用性。

第五章 总结和展望

第一节 总结

股票收益率的预测是量化交易领域最为关键的任务。多因子模型作为最成熟的选股模型之一，被广泛应用于收益率预测、股票筛选等任务。该模型建立在资本投资组合、资产定价、套利定价理论等现代投资理论的基础上，利用价格、成交量、公司会计数据等各类因子对收益率时间序列进行预测。因此，构造出包含有效信息的因子是准确预测收益率的前提和基础，选股因子挖掘是量化交易的重要步骤。本文通过改进经典的 Transformer 模型，使其能够挖掘选股因子并实现股票超额收益的回归分析。

本文的研究对象为 2017 年至 2020 年中国 A 股市场的交易数据，通过一系列的数据清洗、数据预处理、建立模型、结果可视化等流程完成了对选股因子的挖掘和对超额收益的预测，并得出本文提出的选股因子挖掘 Transformer 的预测效果优于经典多因子模型的结论。本文主要的研究结果如下：

第一，本文对 2017 年至 2020 年中国 A 股市场的交易数据进行预处理使其能够满足 Transformer 模型的输入和输出要求。首先，本文根据股市实际情况剔除了取值异常的股票。其次，对自变量日频因子特征数据进行规范化和离群值的处理。然后，通过回归分析消除市值和行业对股票收益率的影响，并计算出模型的因变量未来 10 个交易日的超额收益。接着，按照时间顺序划分数据集，整理出了包含 1409394 个样本的训练集和包含 829604 个样本的测试集。

第二，本文改进了经典的 Transformer 模型。在输入嵌入层部分，本文将一条样本类比为经典模型中的词向量来满足模型输入维数要求。在位置编码部分，本文将自变量特征复制一遍，并让其分别进行正弦和余弦编码。在编码部分，本文采用多头注意力机制，使模型尽可能多地学习到数据中地有用信息。在解码部分，本文用 APT 中的基础多因子模型替代原始的解码器，使解码思路符合多因子选股理论，并使得编码和解码模块能够实现选股因子的构造。在模型输出模块，本文以未来 10 个交易日的超额收益作为预测目标，利用回归分析的思想实现了超额收益率的预测。

第三，本文将选股因子挖掘的 Transformer 模型的预测结果与经典的线性回归、DNN 以及 XGBoost 模型进行对比，得出本文提出的模型预测准确度更高，挖掘出的选股因子具有一定的有效性。本文选择 MSE、MAE、Pearson 相关系数从样本总体上对比各个方法的预测准确度，选择单个交易日相关系数从微观上比较各个方法的实用性和稳定性。定量分析的结果表明，本文提出的选股因子挖掘 Transformer 模型在测试集上的 MSE 为 0.9333，MAE 为 0.7498，相关系数为 0.2579，均小于其他三种对比方法。另外，从四种模型单个交易日相关系数时间序列图来看，Transformer 模型在测试集上也优于其他三种对比方法。因此 Transformer 模型的预测结果准确性和稳定性较好，能够挖掘新的选股因子，具有一定的实用价值。

综上所述，本文提出的选股因子挖掘 Transformer 模型能够构造新的选股因子，并且能够基于这些新的选股因子建立多因子模型更准确地预测股票的超额收益率。

第二节 展望

尽管本文模型得到了具体可行的结果，但仍存在一些内容有待未来进一步的研究。展望未来的研究方向，本文提供以下几种可能的课题：

第一，本文在划分数据集时直接将所有的数据划分为两类，使得仅在一个训练集上对模型参数进行了一次训练。未来可以对数据集进行滑动地划分，譬如每个季度根据之前八个季度的数据进行一次参数训练，每经过一个季度实现参数的动态更新。在此基础上可以根据新数据调整参数，从而使得模型学习到新的市场风格，提升预测准确度。

第二，本文选用的自变量为日频的特征数据，未来可以使用更高频的数据，如 5 分钟频、30 分钟频数据等。更高频的数据含有更多的有效的信息，也有着更高的维度。如何从更高维的数据中提取有用信息是今后研究的一个方向。

第三，本文改进后的 Transformer 模型没有充分考虑样本的时序关系，未来可以通过 Transformer 的注意力机制跨时间维度地考虑特征之间的关系，从而构造出新的因子。当前时序 Transformer 模型的研究仍在不断地进行中，如果时序

模型能够结合第二点提到的更高维的数据，那么大量有效的因子将会被提出。基于高维数据的时序 Transformer 选股因子挖掘模型也是今后研究的重点。

参考文献

[1] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. arXiv preprint arXiv: 1810.11363.

[2] Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics, 33(1), 3-56.

[3] Fama, E. F., & French, K. R. (2015). A Five-factor Asset Pricing Model. Journal of Financial Economics, 116(1), 1-22.

[4] Fang, J., Lin, J., Xia, S., Xia, Z., Hu, S., Liu, X., & Jiang, Y. (2020). Neural Network-based Automatic Factor Construction. Quantitative Finance, 20(12), 2101-2114.

[5] Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T. S. (2019). Temporal Relational Ranking for Stock Prediction. ACM Transactions on Information Systems (TOIS), 37(2), 1-30.

[6] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189-1232.

[7] Guerard Jr, J. B., Markowitza, H., & Xu, G. (2020). Earnings Forecasting in a Global Stock Selection Model and Efficient Portfolio Construction and Management. International Journal of Forecasting 31(2015), 550-560.

[8] Kakushadze, Z. (2016). 101 Formulaic Alphas. Wilmott, 2016(84), 72-81.

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NeurIPS 2017).

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

[11] Mossin, J. (1966). Equilibrium in a Capital Asset Market. Econometrica: Journal of the Econometric Society, 34(4), 768-783.

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

[13] Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. Journal of Economic Theory, 13(3), 341-360.

[14] Schmidt, P. S., Von Arx, U., Schrimpf, A., Wagner, A. F., & Ziegler, A. (2019). Common Risk Factors in International Stock Markets. Financial Markets and Portfolio Management, 33(3), 213-241.

[15] Sharpe, W. F. (1964). Capital Asset Prices: A theory of Market Equilibrium under Conditions of Risk. The Journal of Finance, 19(3), 425-442.
[16] Treynor, J. L. (1961). Market Value, Time, and Risk. SSRN Electronic Journal. Retrieved August 14, 2015, from dx.doi.org/10.2139/ssrn.2600356.
[17] Treynor, J. L. (1961). Toward A Theory of Market Value of Risky Assets. Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics. London: Risk Books, 15-22.
[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems 30 (NeurIPS 2017).
[19] Yang, J., Li, Y., Chen, X., Cao, J., & Jiang, K. (2019). Deep Learning for Stock Selection Based on High Frequency Price-Volume Data. arXiv preprint arXiv: 1911.02502.
[20] 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(2): 21-27.
[21] 陈小悦, 李晨. 上海股市的收益与资本结构关系实证研究[J]. 北京大学学报: 哲学社会科学版, 1995(01): 72-79.
[22] 陈玄玄. 基于 CatBoost 算法的多因子量化选股策略研究[D]. 上海: 上海师范大学, 2020.
[23] 霍丽佳. 基于 AdaBoost 算法多因子选股模型的应用研究[D]. 武汉: 华中科技大学, 2019.
[24] 贾权, 陈章武. 中国股市有效性的实证分析[J]. 金融研究, 2003, 000(007): 86-92.
[25] 李文宇. 基于机器学习的多因子选股策略研究及实证分析[D]. 济南: 山东大学, 2021.
[26] 李想. 基于 XGBoost 算法的多因子量化选股方案策划[D]. 上海: 上海师范大学, 2017.
[27] 李云翔. 基于回归法与 GBDT 多因子选股研究[D]. 南昌: 南昌大学, 2020.
[28] 林晓明, 陈烨, 李子钰, 何康. 基于遗传规划的选股因子挖掘[R]. 南京: 华泰证券研究所, 2019.
[29] 林晓明, 陈烨, 李子钰, 何康. 再探基于遗传规划的选股因子挖掘[R]. 南京: 华泰证券研究所, 2019.
[30] 陆静, 李东进. 基于流动性风险的证券定价模型及其实证研究[J]. 中国软科学, 2005, 000(012): 145-150.

[31]施东晖. 上海股票市场风险性实证研究[J]. 经济研究, 1996(10): 44-48.
[32]舒时克,李路.基于 Elastic Net 惩罚的多因子选股策略[J]. 统计与决策, 2021, 37(16): 157-161.
[33]吴先兴, 杨怡玲. 基于基因表达式规划的价量因子挖掘[R]. 武汉: 天风证券, 2020.
[34]喻术奇. 基于时变加权 LightGBM 的多因子选股交易策略设计[D]. 上海: 上海师范大学, 2020.
[35]周渐. 基于 SVM 算法的多因子选股模型实证研究[D]. 杭州: 浙江工商大学, 2017.
[36]祝养豹. 基于 XGBoost 和 LightGBM 算法的多因子选股方案设计[D]. 南京: 南京大学, 2020.

后记

两年的硕士生活即将告一段落，在论文的最后一页我想对一路以来关心我帮助的师长、朋友表示感谢。

首先，我要感谢我的导师於州教授。我能够完成硕士期间学业离不开於老师的教导和帮助。於老师教会了我许多专业上的知识与本领，也为我人生的发展提供了较大的帮助。於老师已经指导我超过了两年时间，在这里我想对他说一句：“您辛苦了!”衷心地祝愿於老师在未来工作顺心，生活顺意。

其次，我要感谢多智体人工智能实验室的王祥丰副教授和校外合作企业的杨老师。从本文的开题到结题，王老师和杨老师一直以来都提供了很多帮助，感谢他们在研究方法、研究方向、数据选择上提供的宝贵建议。

然后，我要感谢我的父母。如果没有家庭的支持，那么一定就没有现在的我。感谢我的父亲和母亲对我的资助和鼓励，是他们让我能够放心地去闯荡，去选择我自己想要尝试的发展道路。未来我想要继续努力，争取早日回报他们。

接着，我要感谢一路以来许许多多帮助我的朋友。感谢他们对我论文的帮助，感谢他们对生活的帮助。希望我们未来也能齐头并进，互帮互助。

最后，我要感谢我自己。我感谢自己选择了攻读硕士的路，感谢自己选择了学术研究，感谢自己对于感受新生活的尝试，感谢自己的坚持和努力，感谢自己一路上披荆斩棘。未来我会保持自己的态度，哪有什么一夜成名，我根本等不到天亮，继续拼搏吧。