单位代码	10445
学　　号	2021313009
分 类 号	O213

山东师范大学

硕士专业学位论文

基于深度学习的养老行业选股策略研究

Research on Stock Selection Strategy of Pension Industry Based on Deep Learning

学位类别：应用统计硕士

领　　域：应用统计

学习方式：全日制

研 究 生：王文静

指导教师：王海洋

提交时间：2023年5月

目　录

摘要

随着经济的不断发展，我国的金融资本市场不断壮大，金融软实力也成为我国综合国力的重要组成部分。因此，无论是学术研究还是投资实战，证券投资领域都是人们关注的对象。量化投资作为新兴的投资方式，随着计算机理论和技术的发展开始受到投资者的关注，在我国的股票市场中已然成为流行的投资方式。无论是公募还是私募领域，能够获得较高收益并具有较低风险的投资策略都受到广泛而长久的欢迎。

本文以中证养老指数成分股为研究对象，运用机器学习和深度学习算法构建选股模型，从而制订并完善多因子量化选股策略，实现可观且稳定的策略收益。多因子选股策略是量化投资领域常见的策略，其核心是挖掘影响收益的指标，根据指标构建投资组合，以期实现超越指数的收益。机器学习算法所做的工作就是筛选因子数据集，建立选股模型。

具体地说，我们首先通过因子重要性程度和相关性分析来对股票初始的因子池进行初步的筛选，并确定有效因子，形成最终的因子池。然后基于因子数据，进行选股模型的构建。选股模型策略的构建主要分为两个部分。一是由多因子机器选股模型在预测精度上进行模型的选取，然后对股票进行筛选。二是由深度神经网络学习历史数据，对股票价格进行预测。然后基于选股模型，每月进行多因子轮动选股，筛选出表现优异的股票。最终对选出的优秀股票进行交易，通过量化平台形成交易回测的结果，计算相关评价指标，对输出的结果进行评价。

基于上述研究内容，本文利用 2014 年至 2021 年的数据进行因子筛选和选股模型的构建，使用 2021 年至 2022 年的数据进行交易回测以验证策略的效果。回测结果表明本文所构建的多因子选股模型能够稳健地获取超额收益，其中 CNN-LSTM 选股模型的表现最为突出，能够在 2022 年金融市场普遍低迷的情况下实现 32%的策略收益，抵抗住了大环境的动荡。另外，本文还在选股模型的基础上加入了大盘的择时，结果表明此时的收益可以进一步提高，其各项回测指标也都有进一步的改善。

关键词：量化投资；多因子选股；策略回测；机器学习；深度学习

Abstract

At present, the financial capital market is growing along with China's economic development regardless, and the soft power of finance is an important part of our national comprehensive national power. As a result, the field of securities investment is the object of attention, whether in academic research or in investment practice. And with the development of computer theory and technology, quantitative investment as an emerging investment approach began to enter the focus of investors, quantitative investment combined with big data and financial theory, in China's stock market has become a popular investment approach. And having a more high-yield, low-risk investment strategy, both in the private and public sectors, can gain wider and longer-term support.

This thesis constructs a quantitative investment strategy using the CSI Pension Index constituents as the research object. The main research is to use machine learning and deep learning algorithms to build stock selection models to select and improve multi-factor quantitative stock selection strategies, and to achieve considerable and stable strategy returns. Multi-factor stock selection is a common strategy in the field of quantitative investment. The core of the strategy is to find the indicators that affect the return and build a portfolio based on the indicators in order to expect to achieve the return of outperforming the index. Its theoretical explanation has evolved with the introduction of machine learning, and multi-factor regression combined with machine learning is widely used in the practice of the strategy. What machine learning does is to sift through existing research for factor datasets and build models to fit them.

Specifically, the initial pool of factors for a stock is first screened by factor importance and correlation analysis and the final valid factors are identified to form the final pool of factors. Based on the factor data, stock selection models are then selected. There are two main bodies in the construction of the stock selection model strategy. One is the selection of a model by a multi-factor machine stock selection model in terms of prediction accuracy, and then the screening of stocks. The second is a deep network that learns historical data to make predictions on stock price data and make stock selection model screenings. Then based on the stock

selection model, a multi-factor rotation stock selection is made each month to screen out the top performing stocks. The selected outstanding stocks are eventually traded and the output is evaluated by forming trading backtest results through the quantitative platform statistics and calculating relevant evaluation indicators.

Based on the above research, this paper uses the data from 2014 to 2021 for factor screening and stock selection model construction, and the data from 2021 to 2022 for trade backtesting to verify the effectiveness of the strategy. The backtest results show that the multi-factor stock selection constructed in this paper is able to obtain good excess returns in a robust manner, and the CNN-LSTM stock selection model performs better under the evaluation of each backtest indicator, and is able to achieve a strategy return of 32% in 2022 despite the general downturn in the financial market, carrying the turbulence of the general environment. In addition, this paper also wanted to further improve the returns based on the stock picking model, so it added broad market timing, and the results showed further improvements in all its backtest indicators.

Keywords: quantitative investment;deep learning;machine learning;multi-factor stock selection

第一章 绪论

1.1 研究背景及意义

1.1.1 研究背景

金融领域以其强大的收益性吸引着越来越多的投资者，他们时刻关注着股票市场及其金融衍生产品的价格走向。由于历史周期性的发现，投资者期望能更好地预测金融产品价格变化情况，从而能够通过自己的交易策略来获得超额收益。量化投资的方法就是一种通过研究历史走势的数量特征来构建有效的投资方法，属于定量分析方法。随着数字技术的发展，量化投资逐渐利用计算机技术来建立投资量化分析模型，发现资产价格背后走势，指导制定投资策略，由此量化投资逐渐在投资领域得到了广泛应用。量化交易是金融领域最受关注的热点，近几年在海外及国内的发展也日益成熟。目前的研究方法主要在选股方法的结合和创新以及择时上的加入，目的主要是提高预测精度，来实现策略更大的收益以及较小的回撤。

机器学习在金融领域的应用非常广泛，机器学习可以挖掘股价走势潜在的信息，适合与金融市场高频而多维度的数据。由于传统的机器学习在处理多规模数据时存在维度灾难，存储灾难等问题，于是当前人们对于选股模型的研究转向了深度学习网络。利用深度学习进行多因子量化选股是指利用神经网络的学习方法来识别，分析和预测具有经济价值的特征，然后通过大量的学习建立交易策略，来执行自动交易，获得超额收益。其核心算法围绕卷积神经网络和循环神经网络展开。卷积神经网络（Convolutional Neural Networks,CNN）是一种前馈神经网络，由于卷积层和池化层的存在，能够对数据特征进行充分的提取并简化维度提高运算效率。循环神经网络（Recurrent Neural Network,RNN）因为其一个序列的输出与前面的输出有关而被称为循环神经网络。其具体表现为网络会对最前面的信息进行记忆并应用于对当前输出的计算中，即隐藏层之间的节点是有连接的，并且隐藏层的输入信息不仅包括当前输入层的输入还包括上一时刻隐藏层的输出，并由此对有长度的序列数据进行了处理。而在方法的运用上，由于目前人们研究领域的细化，对于股票价格预测上的

适用并无最好的定论，人们更偏向于对于自己所研究的数据结合主观策略去探索能实现预测精度，持有收益和持有风险三者结合的量化选股策略。

结合当下的老龄化背景，不少投资研究人员预测养老行业是未来一段的朝阳产业。根据相关数据表明，我国 2050 年老年人口将达到惊人的 4.8 亿，由此也将成为养老服务市场潜力最大的国家。随着我国老龄化程度越来越严重，各种养老问题应运而生，由此也催生了养老产业的持续发展[1]。在老龄化加剧的背景和相关政府文件的发布下，养老概念开始兴起并成为众多投资者的研究对象。本文的研究对象就是养老行业中的成分股。

1.1.2 研究意义

理论意义上看，在当前使用非线性方法对股价预测的研究中，还是以传统的机器学习方法为主。而作为机器学习的分支，基于深度学习的神经网络模型则主要用于对图像分类以及文本分类，即使在金融时序预测中也仅是以 BP 神经网络及其变形为主。在深度学习的几个常用的神经网络模型中，由于 RNN 相比 CNN 更适合对时间序列进行建模，而金融数据是一组时间序列数据，因此大多研究都是围绕 RNN 及其变体长短期记忆网络(Long Short-Term Memory,LSTM)开展的，将 CNN 与 LSTM 两种神经网络架构结合起来的研究不多。故而本文将 CNN-LSTM 神经网络相结合对股价建模并进行趋势判断更具有创新性。此外，由于 CNN-LSTM 神经网络模型输入变量是二维结构，如何将一维时间序列数据二维化以及具体的使用效果同样值得探究。现实意义上看，本文在选股中依据行业指数选股，可以通过行业发展趋势来减少风险，并且在股票持有中加入了限制条件，大盘择时能使量化选股的表现提高，从而进一步避免风险，规范金融投资市场。

1.2 文献综述

1.2.1 国内外研究现状

由于股票价格存在频率高，记忆性长等特点，吸引了一众经济学家和数学家对其的研究。随着虚拟经济的发展和数据时代的到来，国内外学者对其的研究也愈加成熟和多样，从传统的时间序列方法到深度学习网络模型，从单因子发展为多因子，随着研究的深入，股票预测的精度也有所突破。如今，量化投资已经形成了一个较为成熟的理念，人们从中

寻找适合于数据特征的预测方法和主观策略，从而减少投资的盲目性，将金融投资带入到一个更为高阶而稳定的领域。

量化选股研究已经经历了长时期的发展。最早由 Jules Regnault(1863)运用量化方法，从证券市场中总结出了一条具有普遍意义的规律，并就此规律展开金融市场投资而获取超额利润，在此之后开始了量化理论的发展[2]。自此基于对收益的渴望，量化领域开始涌进大量人才。他们针对金融市场的内在规律展开研究，于是因子选股，量化择时和高频策略等多方面的量化投资理论和策略得以快速的发展。而无论是怎样的策略组合，量化选股都是优秀策略的第一步，即如何在庞大的股票池中选择未来高收益的股票并持有。因为投资者首先期望是选出稳定的，潜在收益高的优质股票，从而才能更好地进行投资策略制定。量化选股中多因子模型的理论最早源于二十世纪，Markowitz(1952)提出了投资组合理论，他运用均值方差模型来衡量收益与风险，是最早的风险投资模型，也是首次在投资组合的选择中运用了数理统计的方法[3]。之后 Shar(1964)等人通过大量研究最终提出了资本资产定价模型，此模型阐述了风险资产和预期收益的内在联系,其研究的重点是通过探究二者的数量关系，来补偿风险换取收益[4]。自此量化投资理论逐步完善并发展。Ross(1976)在研究中发现了套利定价理论，即根据多种可能因素来解释风险资产的回报的差异性，他认为当市场处于不均衡状态时，是会有潜在的无风险套利机会存在的，这就为多因子模型提供相应的理论基础[5]。到了二十世纪末，Farm 和 French(1992)通过研究资本资产定价模型和套利定价理论，结合相关已有研究，模拟三种风险来构造三因子模型。他所发现的风险因子为三类，分别是账面市值比，市场风险和市值风险，可以一定程度上解释股票收益的内在变化的规律[6]。研究者在此之后不断研究因子理论，后来加入了波动率因子和公司质量因子提出了五因子模型[7]。随着因子指标不断地被发现，多因子模型理论逐渐普遍被运用于金融市场的投资决策中，并从中寻找股票市场收益与风险的变化规律。近年来，伴随着人工智能的发展，多因子模型理论开始基于各种机器学习算法理论逐步发展成多种多样的量化选股模型。

当前广泛用于金融领域票券价格预测的方法大致分为三个分支，一种是将其当作时间序列，一种是多因子机器学习模型预测，另外一种是深度学习预测。这些预测方法是时间上的发展，但在面对具体数据却表现各异，很难用一种方法去定义最优，故而人们在研究时，对于方法的结合和细节的把握更为重要。

首先是传统的时间序列预测，我国对于此类的研究开展较早也较多。其中，主要包括自回归滑动平均(Auto-Regressive Moving Average Model,ARMA)、广义自回归条件异方差(Auto-Regressive Conditional Heteroskedasticity,ARCH)等。于志军等(2013)引入了误差校正的方法,建立了结合 GARCH 和回归方法的股价预测模型，为了检验模型的性能，他利用该模型来预测上证指数，实证分析后发现，引入误差校正后的模型预测精度得到了显著的提高[8]。杨琦和曹显兵(2016)研究发现 ARMA 模型与 GARCH 模型的结合可以更准确地用于预测股价，于是使用 ARMA-GARCH 模型对大众公用的股价进行了分析预测,结果表现符合预期[9]。然而由于股价数据的庞大和不稳定性，采用传统的时间序列分析存在很大的弊端[10]。

目前被广泛用于股价预测的是机器学习模型，其理论支撑于资产定价因子模型，随着人们对于因子的发现越来越多，从而为多因子机器学习模型提供了基础。Bruno 等(2008)通过建立支持向量机回归模型预测了资本在三个不同市场中的股票价格，并与随机游走模型进行比较，结果发现支持向量机回归模型在较低的波动期间内，这里他将资本分类，小资本的波动幅度较低，也有较高的预测精度[11]。Cao lijuan(2001)选取芝加哥期货数据进行研究，对比发现支持向量机模型在泛化能力上优于 BP 神经网络[12]。同样，我国的杨新斌和黄晓娟(2010)也基于支持向量机和留一法进行特征筛选建立股票价格预测的模型，实证分析得出支持向量机模型适用于证券市场中股价的预测和分析[13]。Huang(2012)通过基于遗传算法改进的支持向量机回归模型对台湾股票市场进行量化选股，研究表明可以通过遗传算法进行特征选择和参数调优来提升模型的预测效果[14]。吴微等(2001)利用 BP 神经网络模型对沪市综指涨跌做预测，预测准确度达到 70%[15]。但是 BP 神经网络模型存在局部最优的问题，刘恒和候越利用贝叶斯正则化算法对此进行改进，提升了模型的泛化能力[16]。张潇和韦增欣(2018)基于价值成长投资策略选取特征，利用随机森林模型预测股票涨跌，并在回测后得到了较高的超额收益[17]。

另外由于股票数据结构的特殊性，深度学习对于处理后的数据也有很强的适配性。李志杰(2002)利用主成分分析对输入的原始数据进行降维，并利用神经网络模型预测，结果表明，将数据降维后的神经网络模型效果更好[18]。随着时间的发展，RNN 因其能够有效地挖掘股价数据中的时间和语义信息而受到越来越多量化研究者的关注[19]。然而,普通的循环神经网络不能克服其长期依赖性,因此目前使用较多的是其变体长短期记忆网络(LSTM)。

曾安、聂文俊(2019)基于过去和未来的信息建立了双向 LSTM 网络,结果显示该模型可以更精确地反映股指变化,且具有更好的泛化能力[20]。黄媛(2019)基于多因子模型发现 LSTM 模型比支持向量机更适用于股票时间序列数据，并且在回测中验证该策略可超过基准收益[21]。欧阳明哲(2020)结合了门控循环神经网络(Gated Recurrent Neural Network,GRU)构建了多因子量化选股策略，并与 RNN 的预测效果进行对比，发现 GRU 拥有更好的预测效果和更好的策略表现[22]。张虎等(2020)运用自注意力神经网络结合多因子选股，相较于沪深 300 指数获得了更高的收益且风险更小[23]。

1.2.2 文献评述

综合以上文献可以发现，目前大多数研究者将量化选股转换为股票的分类或者股票价格的回归来进行研究，并且借助于数据挖掘算法的不断发展，在金融领域关于量化选股已经有了丰硕的成果。但是模型不可能完全预测未来的发生，有模型预测就有拟合误差，存在主观策略就有策略风险，这都是无法避免的。于是人们开始研究如何提高模型收益及其稳定性，因此很多学者通过引入有效指标和指标评价体系来进行改进。

当前的研究者大多数借助于机器学习模型，仅将每条股票数据作为单独的样本来看待，而忽视了股票数据的时序性特征。而通过对经典量化模型的研究可以发现股票时序性特征对股票价值的估算有很大的影响[24]。因此能否识别股票价格金融时序的特征，将其纳入预测模型，对于选股模型有一定影响。

不可否认，近年来随着深度学习的崛起，给研究者带来了更为优化的预测模型，使得对于高频股价数据的预测精度提上一个阶梯。但是，近年来对于量化选股领域的研究集中于方法上的创新和结合。而股价的数据特征复杂，引发波动的因素多样，其主要核心产业分级，行业分化等等都在给研究学者抛出一个新的问题。即如何在交易策略的设计上，考虑到这些复杂性，来选择相对最优方法保证预测精度的同时，兼顾更多的环节来提高收益和减少风险。

1.3 论文组织结构

1.3.1 研究内容

本文以中证养老指数 80 支成分股为样本，选取多个因子，截取 2017 年 1 月 1 日至 2021 年 11 月 30 日数据进行模型训练，并在 2021 年 12 月进行选股模型的预测评价，然后在 2021 年 12 月到 2022 年 12 月一年的区间内进行回测。即基于机器学习和深度学习，构建出多因子选股模型，通过机器学习算法对个股进行预测分类，从而筛选出可能获得超额收益的股票，构造有效的前沿组合。并模拟回测证实其有效性，做出进一步的策略评价，以及加入对大盘的 XGBoost 择时，通过回测判断其相对于单选股策略回测表现的提升是否明显，最终构建出一个完整的量化选股流程。即择时+深度学习选股来实现表现更为优势的量化策略。本文研究中所用的方法有，首先利用随机森林提取因子重要性，然后在选股模型中利用逻辑分类，高斯贝叶斯，随机森林和支持向量机进行涨跌预测和模型评价，再利用硬投票融合改善模型预测结果。利用 CNN，LSTM, CNN-LSTM 等深度学习算法进行股票预测和模型评价。最后再借助量化平台完成收益与风险的评价，最后利用 XGBoost 对大盘预测进行择时，提高收益指标并减少模型风险。

1.3.2 文章章节安排

第一章，绪论。该章在量化理论和模型选择方面介绍了当前的研究背景和研究现状，结合当前研究来对文章的内容进行安排和框架介绍。

第二章，相关概念和理论基础。该章节介绍了量化选股理论层面的合理性和量化选股模型的介绍，主要有机器学习中的高斯贝叶斯模型，支持向量机模型，随机森林模型，集成模型，深度学习中的 CNN，RNN，LSTM，GRU 以及 CNN-LSTM 和 CNN-GRU 模型。

第三章，数据预处理。主要对数据进行了异常值、空值处理，以及重点样本筛选，因子池的构建。

第四章，多因子选股模型构建。本章对机器学习分类模型进行了比较，其中有单一模型在验证集上的准确率评价以及集成模型的准确率评价，深度学习模型进行了股票收盘价格预测，选出预测模型后分别设置了轮动选股规则，另外，所有进入评价指标的数据都没有进入模型训练，其结果更具有代表性。

第五章，量化选股模型回测。上文内容是对选股模型预测精度的评价，而判断量化策略好不好，在金融领域就要对策略进行回测，来从收益、风险多维度的评价本文策略，并加入择时来看策略是否提升。本文的流程框架如图 1-1 所示。

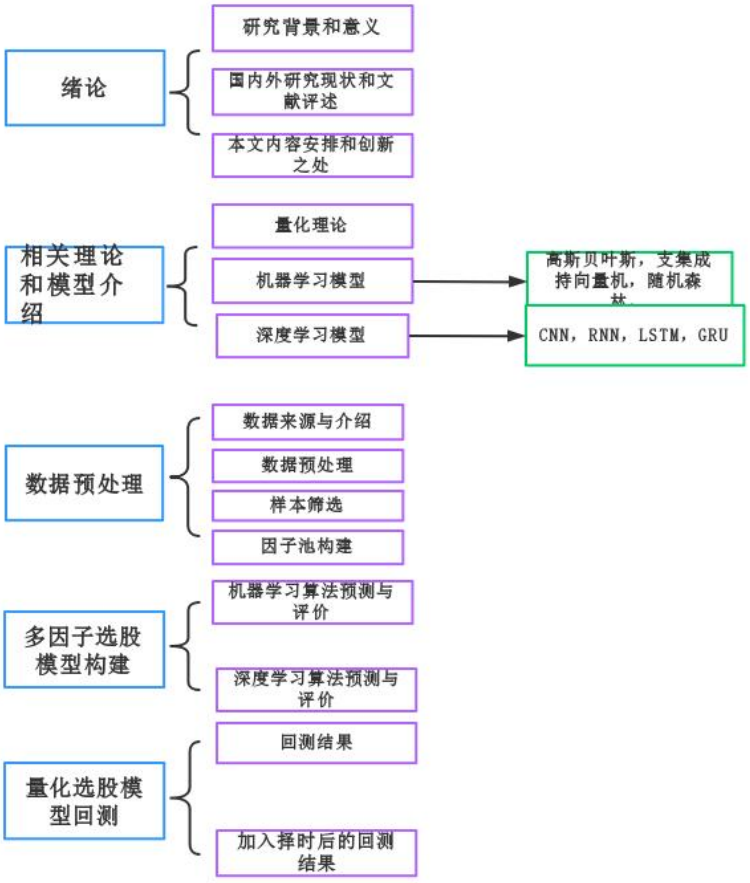


图 1-1 研究框架

1.3.3 研究重难点与创新

本文在研究过程中遇到了许多的问题，也相应的做出了处理。首先是预测标签要选取重点样本。因为数值差异微弱而计入分类标签会影响模型的预测，于是在实际模型训练中进行了参考，只选取了今日收盘较昨日收盘为增的前 20%标签为 1（涨），反之为 0（跌）。然后是策略设计考虑多方面，本文从选择行业内股票，模型选择和评价在单独验证集进行，到考虑加入量化择时，都以降低风险，提高收益为最终目的。最后深度学习较机器学习模型更为优势，但 LSTM 为当前主流涉及时间序列的预测模型，能否再此基础上得到更为优化的模型也是本文研究的重点。另外本文没有只参考当前的选股模型比较，而是在针对自己的股票池完整构建一套量化策略。并实现了 CNN 结合 LSTM 以及 GRU 针对金融数据的预测，而在历史文献中，CNN 结合 GRU 还没有被用于金融数据的研究中。

第二章 相关概念和理论基础

本章主要介绍在构建多因子量化选股模型中用到的模型算法的理论。首先是构建多因子策略的理论基础。然后解释高斯朴素贝叶斯理论，随机森林和支持向量机机器学习模型以及模型集成。最后阐述深度学习中的 CNN，RNN，LSTM 算法理论。

2.1 多因子选股模型

多因子选股策略是当今股票选股策略最广泛应用的策略之一，主要是通过选择会影响收益率的因子，确定收益率与选定因子的关系来选择未来持有股票。其原理上的解释由资本资产定价模型发展而来[25]。该模型主要研究资产风险与期望收益之间的关系，只引入量化后的资产风险一个因子。随着市场异象的出现，人们发现股票市场存在超额收益，而这种超额收益并不能简单由市场因子所解释，由于越来越多的因子开始被研究和发现。KROPE(2012)的研究中又发现了市盈率因子与开发支出[27] 。HULTEN. C .R 和 HAO. X(2015)在论文中将股票市值进行分解，分解的变量有行业虚拟变量，个股市值，公司利润，股票净资产[26]。随着多种影响因子的发现，多因子模型也逐渐丰富和完善[28]。

多因子选股目前被分为两类，基于基本面因子进行选股和基于技术面因子进行选股[29]。技术面因子在量化后表现为对历史日值，三日值，七日值相关指标的计算，主要是历史交易的反映。基本面选股则是针对上市公司的发展指标来进行研究，主要是该公司生产经营情况的反映。

构建一个多因子选股模型先要获取大量因子数据，然后处理后选择部分因子加入因子库，再建立模型确定股票收益率与因子之间的关系，从而以获取尽可能高的收益为原则创建一个选股策略，并且通过进行回测的结果判断策略的好坏。

2.1.2 因子选择方法

多因子选股模型认为股票的收益率与一系列因子相关，这些因子分为规模因子、估值因子、成长因子、技术因子几大类。通过专业的金融平台可以获得包含多类因子历史数据，但是因子数量众多，需要通过对因子数据进行预处理、因子有效性分析，多因子相关性分析等步骤确定出几个相关性强的因子列入因子库。传统的多因子选股有打分法和回归法，

打分法是根据各个因子数值影响收益率的大小进行打分，打分法确定因子权重时过于主观，对结果的准确率有较大的影响。回归法是建立线性回归模型确定因子数据与股票收益率关系，然后进行预测，从而选出收益率较高的股票进行投资。回归法在确定权重系数时比较准确，但是收益率与因子之间的关系并不一定呈线性相关。随着各类机器学习方法的发展，因子的选择也被用于量化选股模型之中，本文的因子筛选通过随机森林来确定因子的重要性程度。在随机森林中，特征对目标变量预测的相对重要性可以通过树中的决策节点特征使用的深度来进行评估。由于决策树顶部使用的特征对更大一部分输入样本的最终预测决策做出贡献，因此，可以使用接受每个特征对最终预测的贡献的样本比例来评估该特征的相对重要性。

2.2 机器学习算法与集成

2.2.1 高斯朴素贝叶斯

一种常用的朴素贝叶斯实现算法，高斯朴素贝叶斯假设条件概率在输出为第 k 个标签下特征满足的输出，即 $P(X=x\,|\,Y=c)$ 是多元高斯分布。由于特征之间具有独立性，我们可以通过每个特征的条件概率建模，第 i 个特征的条件概率服从 $N(\mu_i,\sigma_i^2)$。在 c 类下第 i 个特征对应的高斯分布为:

$$g(x_i)=\frac{1}{\sqrt{2\pi}\sigma_{i,c}}\exp\left\{-\frac{(x_i-\mu_{i,c})^2}{2\sigma^2{}_{i,c}}\right\}. \tag{2-1}$$

由此我们可以得到 c 类的条件概率：

$$P(X=x\,|\,Y=c)=\prod_{i=1}g(x_i;\mu_{i,c},\sigma_{i,c}), \tag{2-2}$$

此时朴素贝叶斯模型变为：

$$P(Y=c_k\,|\,X=x)=\frac{P(Y=c_k)P(X=x\,|\,Y=c_k)}{\sum_k P(Y=c_k)P(X=x\,|\,Y=c_k)}\times P(Y=c_k)P(X=x\,|\,Y=c_k), \tag{2-3}$$

输出 y 的标签为：

$$y=\arg\max_{c_k}P(Y=c_k)P(X=x\,|\,Y=c_k). \tag{2-4}$$

2.2.2 随机森林

随机森林算法将决策树与自助采样法结合，并以最开始的样本集为基础，通过有放回的抽样，抽取部分样本来组成若干个训练样本集，再将样本集放入到决策树的训练之中，一直重复上述步骤来产生更多的决策树，最终定义终止阈值结束[31]。随机森林由多个互相独立不相同的决策树组成。决策树在构建时，样本每次回选取有放回的重采样，并选取部分特征进入训练。这样每棵树使用的样本不同，特征不同，由此训练出的结果不同。这样的随机过程降低了异常点和相关程度弱的特征的影响对于分类预测结果的影响。随机森林采用了集成的算法，结合重采样，提高了准确性。由于选取的样本和特征是随机的，可以避免发生过拟合。不容易受噪声数据影响，处理高维数据有优势，由于决策树之间独立，可以在处理高维数据时提高效率,但是随机森林内部运行没有办法解释，不能输出预测模型。随机森林可以用于分类和回归，随机森林实际上就是用多个决策树分类的分类器，它以单一分类器的结果经过投票后得到最后的分类结果。其中，训练每棵树的节点使用的特征是从所有特征中根据一定比例随机无放回选取的。其内部流程如下图所示。

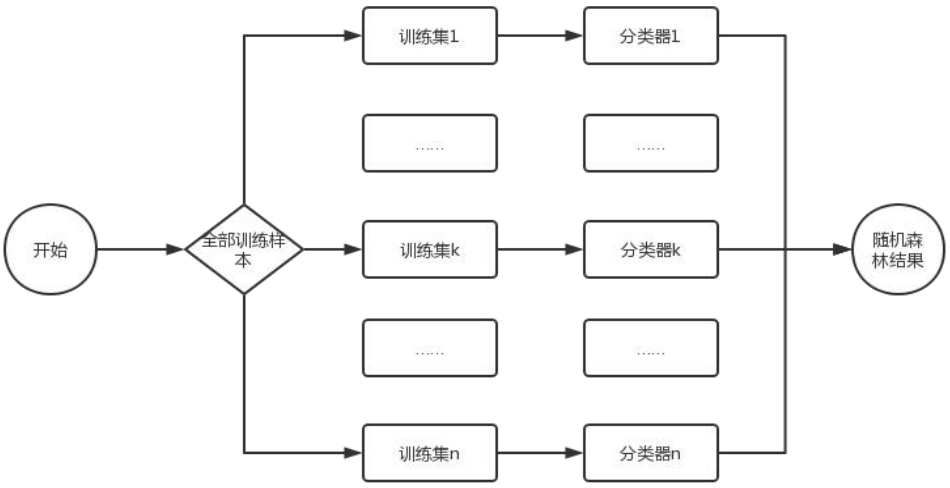


图 2-1 随机森林原理图

2.2.3 支持向量机

支持向量机是当前二分类问题的主流机器学习算法，其基本思想是找到一条线，使其与最近样本点的间隔最大，并由此判断新的样本点所属类别。这条间隔最大的线性分类器就是支持向量机超平面。此方法适用于中小型数据量，非线性，多维的数据结构，属于监

督学习。

令 $\boldsymbol{w}$ 和 b 分别表示权重向量和最优超平面偏移，则相应的超平面可以被定义为：

$$\boldsymbol{w}^T X + b = 0, \tag{2-5}$$

样本 $\boldsymbol{X}$ 到最优超平面的几何距离是：

$$r = \frac{\boldsymbol{w}^T \boldsymbol{X} + b}{\| \boldsymbol{w} \|}. \tag{2-6}$$

我们将泛函距离固定为 1，那么对于给定的训练集 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n \in \boldsymbol{R}^m \times \{\pm 1\}$ 有：

$$\begin{cases} \boldsymbol{w}^T \boldsymbol{x}_i + b \geq 1 \ , \ \mathrm{y}_i = +1, \\ \boldsymbol{w}^T \boldsymbol{x}_i + b \leq -1 , \ \mathrm{y}_i = -1. \end{cases} \tag{2-7}$$

支持向量 $\boldsymbol{X}^*$ 到最优超平面的几何距离为：

$$r^* = \frac{\boldsymbol{w}^T \boldsymbol{X}^* + b}{\| \boldsymbol{w} \|} = \begin{cases} \dfrac{1}{\|\boldsymbol{w}\|} \ , \ \mathrm{y}^* = +1, \\ -\dfrac{1}{\|\boldsymbol{w}\|} \ , \ \mathrm{y}^* = -1. \end{cases} \tag{2-8}$$

分离的间隔 ρ 可以表示成：

$$\rho = 2r^* = \frac{2}{\|\boldsymbol{w}\|}, \tag{2-9}$$

最大化间隔即：

$$\begin{aligned} &\min \ \frac{1}{2}\|\boldsymbol{w}\|^2, \\ &s,t. \ \ y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) \geq 1, \left(i = 1, 2, \ldots \ldots, n\right). \end{aligned} \tag{2-10}$$

用拉格朗日乘数法表示为：

$$L_{(\boldsymbol{w},b,a)} = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + b \right) - 1 \right], \tag{2-11}$$

对 $\boldsymbol{w}$ 和 b 求偏导并使之为零，得到最优化条件：

$$\begin{cases} \dfrac{\partial L_{(w,b,a)}}{\partial w} = 0, \\ \dfrac{\partial L_{(w,b,a)}}{\partial b} = 0, \end{cases} \Longrightarrow \quad \begin{cases} w = \sum_{i=1}^{n} \alpha_i y_i x_i, \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases} \tag{2-12}$$

相应的对偶问题为：

$$\max W_{(\alpha)} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j,$$
$$\sum_{i=1}^{n} \alpha_i y_y = 0, \tag{2-13}$$
$$\alpha_i \geq 0, (i = 1, 2\cdots\cdots，\ \mathrm{n}）.$$

补充 KKT 条件：

$$\alpha_i \left[y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + b \right) - 1 \right] = 0. \tag{2-14}$$

在确定最优拉格朗日乘子α_i^*后，我们可以计算最优权重向量$\boldsymbol{w}^*$和最优偏置b^*为：

$$\begin{cases} \boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i, \\ b^* = 1 - \boldsymbol{w}^{*T} \boldsymbol{x}_s, y_s = +1. \end{cases} \tag{2-15}$$

2.2.4 集成学习

集成方法是将两个或多个单独的机器学习算法的结果结合在一起，并试图产生比任何单个算法都准确的结果。在软投票中，每个类别的概率被平均以产生结果。在硬投票中，每个算法的预测都被认为是选择具有最高票数的类的集合。即：软投票是概率的集成，硬投票是结果标签的集成。集成学习集合多个算法，算法可相同可不同，由训练数据构成一组基分类器，模拟多个决策者同时进行一个角色，输出可加权（软投票）可少数服从多数（硬投票）。即图 2-2 所示。

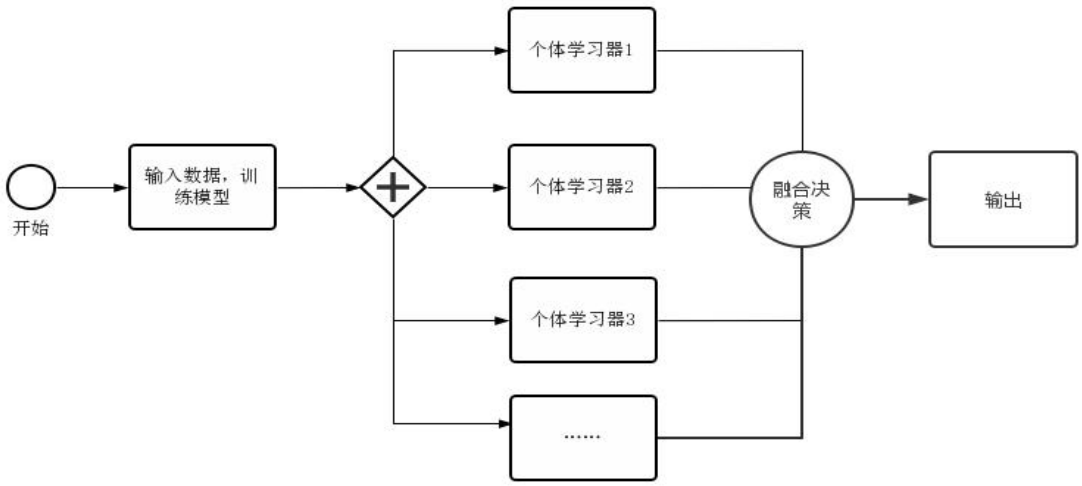


图 2-2 模型集成原理图

2.3 深度学习算法

2.3.1 CNN

CNN 是一个常见的卷积神经网络结构，其能够对因子特征进行充分的提取，由输入层，卷积层，池化层，全连接层，输出层组成，各层有着不同的功能与原理。输入层输入的是初始数据，为矩阵形式。卷积层能够使用卷积核进行特征提取和特征映射，采用 RELU 及激活函数可以对输入的数据进行非线性变换。池化层能够通过最大池化，平均池化，进行下采样，作稀疏处理，减少运算量。全连接层通常在 CNN 的尾部进行重新拟合，让该层的所有节点与上一层相连接，减少特征信息的损失。输出层是全连接层激活后的结果，能够输出分类结果。

其中卷积层和池化层是 CNN 网络特有的网络层，其大小和参数可以根据实际问题调整，卷积层中的超参数有卷积核大小和数目，其实际工作就是将原输入与卷积核做乘积运算，边界多以 0 填充，由此不改变输入大小，卷积核数目会改变输入深度。池化层参数会改变输入大小，只用于减少运算。另外 CNN 的网络层还要设置 Dropout 层，来随机砍掉一些样本，防止过拟合。以上就是 CNN 网络层中的前向传播，若输出结果与预期效果相差较大，则需要通过损失函数梯度下降来进行反向传播，寻找最优参数。CNN 工作过程如下图。

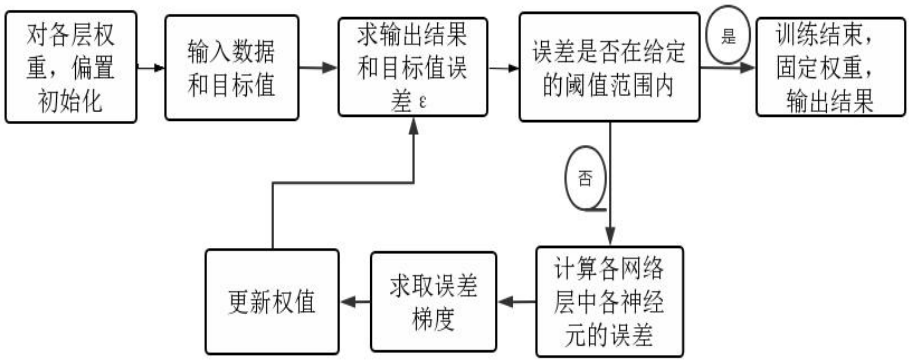


图 2-3 CNN 工作流程图

2.3.2 RNN

传统的神经网络对于时间序列的数据处理效果不理想，比如声音信号等，因为传统单向无反馈连接使神经网络只能处理输入信号所包含时间段的信号，而对于其他信号在处理

本段信号时几乎不起到效果。这便与现实情况相驳，人们需要的是能够将前后时间段的信号连接起来的网络模型，由此诞生了循环神经网络。RNN 之所以能被称之为循环神经网络，主要依据了“一个序列的此刻输出与之前的输出也是相关的”。表现在下一层数的输入值要加入前面层的输出值，隐藏层之间存在连接的。RNN 适用于处理序列数据，例如我们在预测下一次天发生什么时，往往需要用到前面很多天的数据。循环神经网络其具体表现为网络在提取当前信息进行预测处理时会参考和记忆前面的数值信息，即隐藏层之间的节点变成了有连接的，由此对序列数据进行处理。展开的 RNN 结构如下图所示。

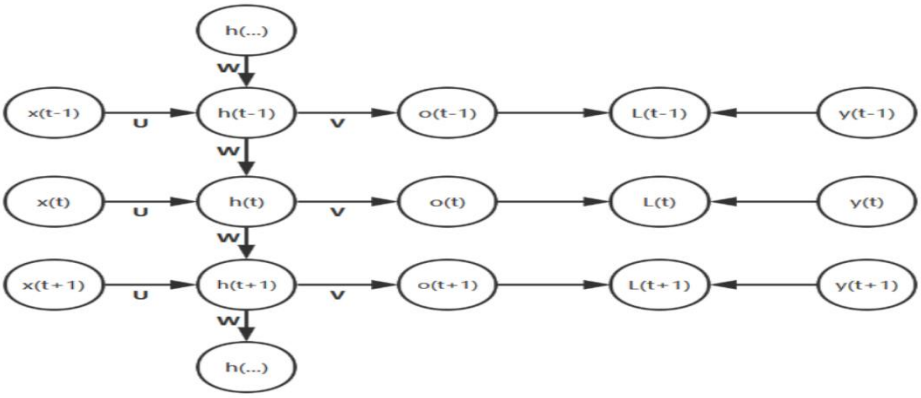


图 2-4 展开的 RNN 结构图

在每个时间状态，网络层接受当前时间状态的输入 $\boldsymbol{x}_t$ 和上一个时间状态的网络状态向量 $\boldsymbol{h}_{t-1}$，经过

$$\boldsymbol{h}_t = f_\theta(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) \tag{2-16}$$

变换后得到当前时刻的新状态向量 $\boldsymbol{h}_t$，并写入内存状态中，在每个时间状态上，网络层均有输出产生：

$$\boldsymbol{o}_t = \phi(\boldsymbol{h}_t), \tag{2-17}$$

我们可以用权重向量 $\boldsymbol{w}_{xh}$ 和 $\boldsymbol{w}_{hh}$ 来参数化 f_θ，此时内存状态按照如下公式更新：

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{w}_{xh}\boldsymbol{x}_t + \boldsymbol{w}_{hh}\boldsymbol{h}_{t-1} + b\right). \tag{2-18}$$

通过此表达式我们可以看出，$\boldsymbol{w}_{xh}$ 和 $\boldsymbol{w}_{hh}$ 均是可导的，由此可以求解梯度，寻找最优参数。

2.3.3 LSTM

长短期记忆神经网络，是一种特殊的循环神经网络。循环神经 RNN 较 CNN 更能解释

时间序列的问题，它是一类以时间序列为输入，经隐藏层和内部记忆单元对数据进行提取，且所有循环单元链式连接的循环神经网络。它的特点是权重共享，以下一时刻的输出为预测值，且这个预测值由当期输入和往期预测构成，但其缺点是往期信息经过长时间的激活，会损失信息，造成梯度消失，不能长期记忆，这就是长期依赖问题。为此 LSTM 作为一种 RNN 的改进算法应运而生，LSTM 模型存储块包含 3 个门。三个门的作用和位置可以由下图表示。

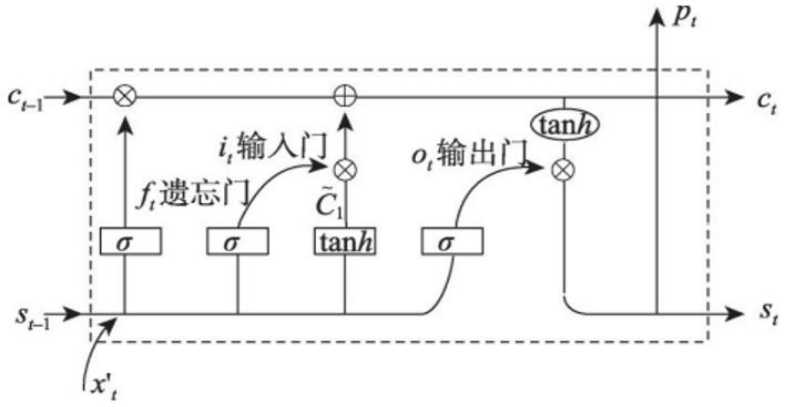


图 2-5 LSTM 内部结构图

其中遗忘门为*f*，它经激活函数输出为区间(0,1)内的值，由此决定了上一期的内部记忆单元 *c* 有多少可以输出到这一期。同样，输入门和输出门也是同样的门控原理，其分别决定了内部记忆单元的更新状态有多少进入到新一期的记忆单元，以及新一期的记忆单元有多少输出到最新一期的预测。其取值同样在(0,1)内，其中 0 代表全部不通过，而反之 1 代表全部通过。其结合上图各个门控的运算以及内部记忆单元的更新如下：

$$\begin{aligned} f_t &= \sigma\left(\boldsymbol{w}_f \cdot \left[\boldsymbol{s}_{t-1}, \boldsymbol{x}_t\right] + b_f\right), \\ i_t &= \sigma\left(\boldsymbol{w}_i \cdot \left[\boldsymbol{s}_{t-1}, \boldsymbol{x}_t\right] + b_i\right), \\ o_t &= \sigma\left(\boldsymbol{w}_o \cdot \left[\boldsymbol{s}_{t-1}, \boldsymbol{x}_t\right] + b_o\right), \\ \tilde{c}_t &= \tanh\left(\boldsymbol{w}_c \cdot \left[\boldsymbol{s}_{t-1}, \boldsymbol{x}_t\right] + b_c\right), \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t, \\ s_t &= o_t \cdot \tanh\left(c_t\right). \end{aligned} \tag{2-19}$$

LSTM 模型网络训练可以分为三个步骤进行。先进行网络初始化，初始化权值和偏置向量，设定存储块的总个数 N，窗口长度 L，学习率η，迭代的阈值ε，迭代次数，初始记忆，初始输出。然后进行数据的划分，划分训练样本和测试样本。最后利用训练样本训练建立 LSTM 网络。

2.3.4 GRU

虽然 LSTM 对处理长序列具有很好的记忆能力，在很多情况下都比基础的 RNN 有更好的性能表现，最关键的是不易出现提督弥散情况，但是 LSTM 的缺点也十分明显，结构较复杂、计算的代价较高、模型参数也容易较多。因此 GRU 门控神经网络很好地解决了这些问题。由于 LSTM 里最重要的结构门是遗忘门，GRU 将门控数量减少到 2 个，分别是复位门与更新门，并只使用一个状态向量，大大减少了模型的参数数量，简化了模型结构。

复位门用来控制上一个时间戳的状态$\boldsymbol{h}_{t-1}$进入 GRU 神经网络的量，向量由当前输入与上一时间状态得到，具体公式如下：

$$f_r = \sigma(\boldsymbol{w}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_r). \tag{2-20}$$

其中$\boldsymbol{w}_r$、b_r为复位门的参数，优化由反向传播算法，σ为该门的激活函数，常使用的是 Sigmoid 函数，上述f_r不会控制输入只控制状态$\boldsymbol{h}_{t-1}$，由此输出$\tilde{\boldsymbol{h}}_t$。

复位门的工作门结构原理可由下图解释。

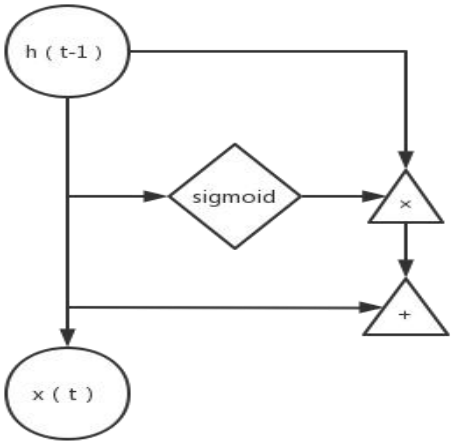


图 2-6 GRU 复位门工作图

更新门用于控制上一个时间段的状态$\boldsymbol{h}_{t-1}$与新输入$\tilde{\boldsymbol{h}}_t$对当前状态向量$\boldsymbol{h}_t$的影响程度，向量设置为

$$f_z = \sigma(\boldsymbol{w}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_z). \tag{2-21}$$

其中$\boldsymbol{w}_z$、b_z为更新门的参数，优化由反向传播算法，σ为该门的激活函数，常使用的是

Sigmoid 函数，上述 f_z 控制状态 $\tilde{\boldsymbol{h}}_t$，$1-f_z$ 控制状态 $\boldsymbol{h}_{t-1}$。由此可以清晰得出，$\tilde{\boldsymbol{h}}_t$ 与 $\boldsymbol{h}_{t-1}$ 对 $\boldsymbol{h}_t$ 的更新量多少为相互竞争的关系。复位门和更新门在 GRU 中的完整工作流程如下图，其中黄色为上文复位门内容。

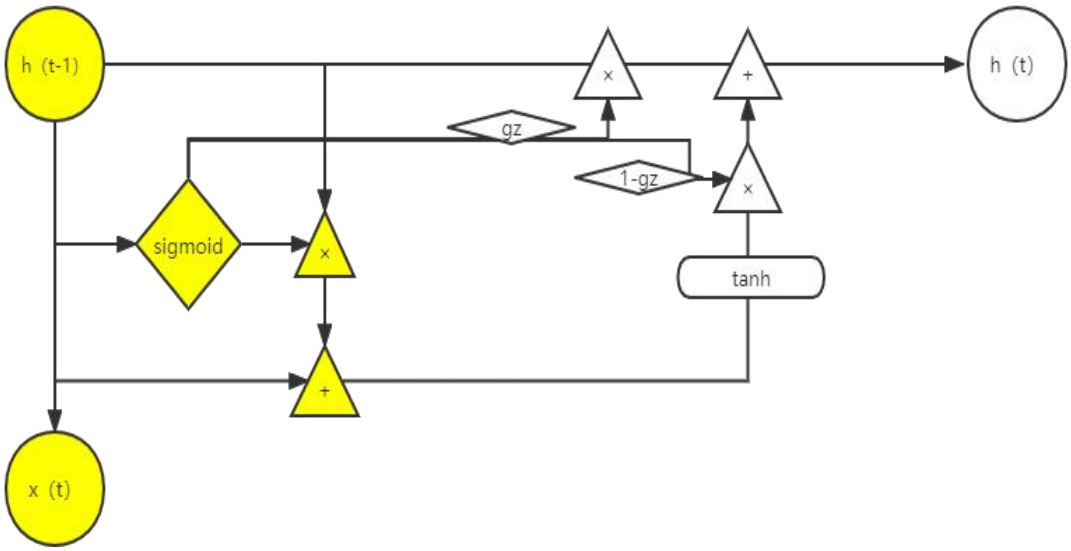


图 2-7 GRU 内部工作流程图

2.3.5 神经网络应用于股票价格预测的可行性

股票价格数据本身是非线性非平稳的时间序列，因此早期的传统时间序列分析方法可以适用于它，并解决它的非平稳性，但其本质上还是线性回归。而影响股票价格的因素众多且复杂，相关性强，在一定程度上限制了传统的线性预测的预测结果。神经网络模型可以通过对人脑工作机制的模拟，神经元之间连接权重，架构网络模型，之中每个神经元内都有偏置项和激活函数来对输入信息进行非线性变换，便克服了股票价格时序的非线性和非平稳性。以此在数据结构上神经网络模型便适用于金融股票价格，从而突破传统限制，更大程度上能够拟合实际的股价。此外，神经网络结构中，各层神经元节点之间相互连接，可供训练的参数众多，可以更有效地发现数据之间的隐藏关系，从而强化模型的表达能力。

然而，如果数据量不足的情况下，过多的训练参数反而不会达到很好的效果，这是因为出现了过拟合现象，这会使模型丧失对未来数据的预测能力。而且，网络层数和节点越多，训练花费时间越长，因此在利用神经网络时要保证数据的充足，而大数据时代和众多金融数据的发展，让人们获取海量的金融数据信息更为便捷。由此，在国内外有多名学者已将神经网络用于做股票预测，并取得较好的预测效果。

第三章 数据预处理

本章主要是数据的介绍和预处理。对于原始数据的处理工作，重点在于因子数据的处理。首先处理因子的缺失值，极端值，然后对因子做一个重要性和相关性筛选。利用搜集到的多因子选股基础因子，结合机器学习重要特征的提取和因子池之间的热力图，来筛选进入最终预测模型的因子，形成最终的模型因子指标。

3.1 数据来源与介绍

本文选取的股票池为中证养老指数中的 80 支成分股，中证养老指数的指数代码为 399813，指数编制从 2004 年 12 月 31 日开始，基准指数为一千点。它的成分股是从酒店旅游，文化传媒，人寿保险，医药卫生等相关产业选取的股票，用来反映养老相关上市公司的整体表现，如今中证养老指数已突破七千点，远高于沪深 300 指数的增长速度[30]。本文选取养老行业成分股一是为了更好的保证股票池的优良性质，纳入行业指数中的股票多为相关行业中业绩较好，持续表现更优的企业，从而避免了上市交易时间短，被警告有退市风险等相关股票。其次，养老概念股是顺应国家政策风向的股票概念股，其本身就隐藏着持续且强势的盈利能力。由此看来，挑选养老概念股进行研究，在股票的筛选上就体现了降低风险，提高收益的量化选股宗旨。

本文数据来源于优矿平台，并结合中证养老指数中的 80 支成分股代码，选取了所有成分股自 2017 年 01 月 03 日（首月第一个开盘日）至 2022 年 12 月 01 全部日频收盘价以及对应平台因子，其中 2021 年最后一个月 23 个开盘日数据作为验证集，来评价模型好坏。平台因子总计 244 个，包括技术因子，财务利润，管理因子，成长因子等，其中技术指标多为量价因子，其相关指标的计算采用的是除权去息后的前一日收盘价，股票交易数据为当日收盘价，原始数据样本如下表。

表 3-1 原始数据展示

ticker	secShortName	tradeDate	closePrice	AccountsPayablesTDays	AccountsPayablesTRate	…	JDQS20
000526	紫光学大	2017/1/3	39.8	-7.8101	-46.0942	…	0.4

3.2 空值和异常值处理

在中证养老指数成分股及因子数据中，同一只股票之间，价格数据与交易量数据，公司财务数据等不在一个量级上，不同股票之间，数据差异跨度也较大，因此需要对数据进行预处理。首先是空值的处理，ST 股票由于财务亏损或经营异常从而存在退市风险，在正常情况下的投资收益非常低且风险大，应该剔除此类数据，原始数据中股票成分较好，只有 ST 紫学和 ST 慧球两支 ST 股票，将其被 ST 的时间段数据剔除。除此之外，由于因子众多，部分因子数据缺失，这类因子多为公司不公开披露，给予剔除。对于异常值本文利用 3σ原则，将距离样本平均值 3σ之外的数据样本定义为异常值，由于它是金融时间序列数据的一部分，不能随意将其剔除，于是用上 95%分位点和下 95%分位点代替来降低异常值的影响。

针对于数据量纲问题，采用数据归一化的方式。数据归一化就是将不同规模，不同单位的数据做统一的变换处理，使得不同维度的数据在数值上变得可比，便于模型训练时进行特征提取。在模型训练过程中，经过无量纲化之后的数据特征对于模型的求解有加速作用，特别是对于需要计算梯度和矩阵的模型，因此将数据进行归一化是模型训练的重要环节。数据归一化指将一列数据变化到某个固定区间中，通常，这个区间是[0, 1]，当然广义的讲可以是各种区间，其他情况可能映射到[-1,1]。在做数据归一化时，本文希望所有数据范围在 0 到 1 之间，同时保持原有数据结构。因此采用 Max-Min 归一化方法对特征因子下的数据做无量纲化处理，归一化具体公式如下：

$$y = \frac{(x - x_{\min})}{(x - x_{\max})}. \tag{3-1}$$

因子标准化之后我们可以统一量级，使因子之间更具可比性，也可以进一步削弱异常值的影响，从而有助于提高模型训练效率和精度。

3.3 样本筛选

下载后将全部数据拼接发现有 85294 条数据，其中新加了一列今日收盘价比昨日收盘价来定义涨跌程度，并提取了涨跌数据中的前 20%来训练模型，以减少变化不大的股票表现对模型的影响，也提高了模型训练的效率。

本文章选取的是中证养老指数下 80 支成分股的全部数据，总计 8 万多条，由于涨跌

幅度会对模型训练产生影响，一些涨跌幅度比较小的样本会影响模型判断。于是本文首先对放入模型的样本做了一个筛选，挑选了 2017 年 01 月 01 日至 2021 年 12 月 01 日全部股票数据样本，对（今日收盘价-昨日收盘价）/昨日收盘价定义了一个新的涨跌比率，并调取了上涨幅度的前 20%的股票，标记为 1。同理，对于下跌幅度的前 20%做一个筛选，标记为 0，最后输入 16194 个股票数据标签，做模型训练的股票池。

3.4 因子池构建

初始因子有 244 个，样本标签按照上文方法标记为上涨和下跌。由于不是所有的因子都会对股票最终的涨跌有显著的影响，因此本文先做了因子池的筛选，来确定最终放入选股模型的有效因子。随机森林可以在生成的多个决策树中寻找最优的那一棵，在此过程中决策树的不同是由特征维度的不同划分的，因此最优的决策树对应的特征排列顺序可以代表其重要程度。本文首先利用随机森林做因子重要程度的计算，其中因子重要性程度排序如下图所示。

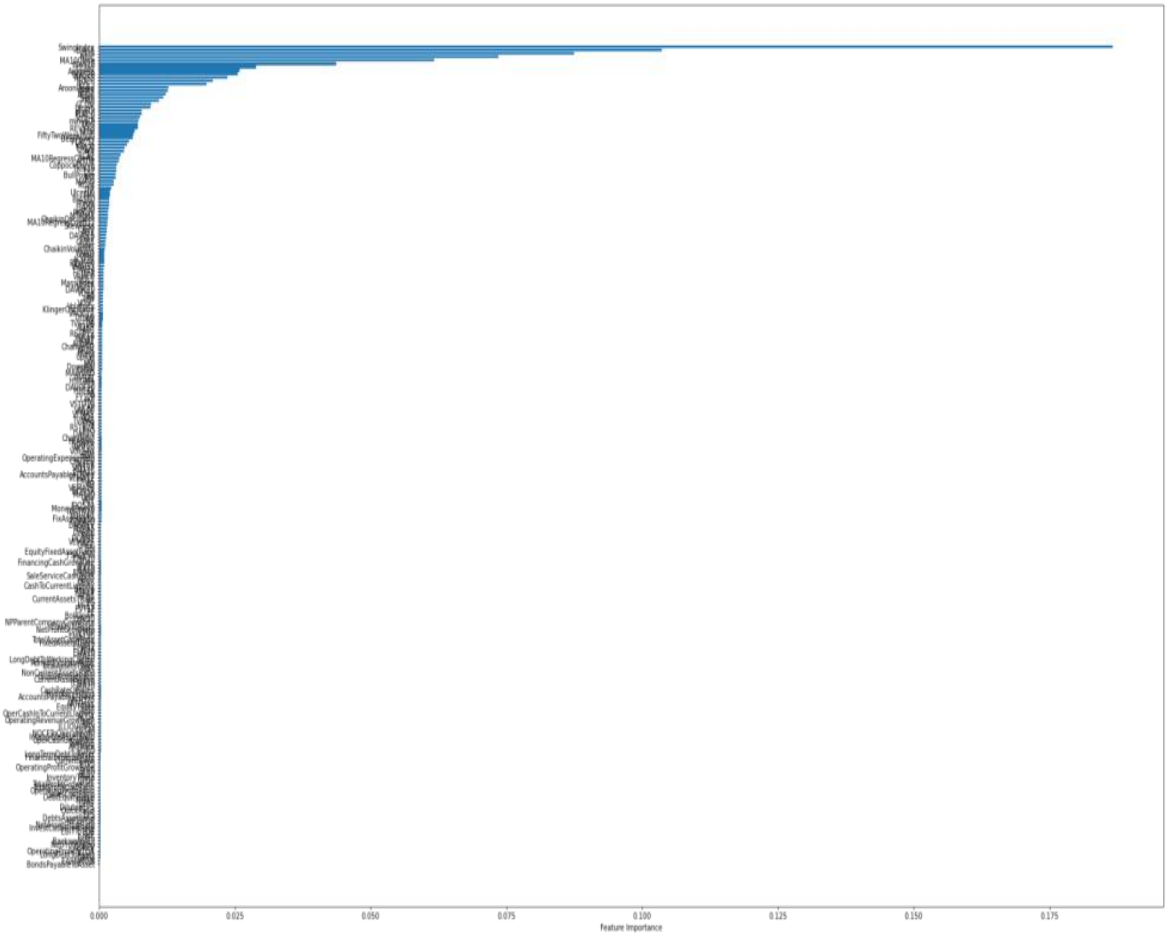


图 3-1 初始因子重要性程度

选取重要性程度大于 0.005 的特征作为模型的预输入，总共 27 个特征。这些特征的名称和重要性程度如下表所示。

表 3-2 预输入因子展示

因子序号	因子名称	因子重要性程度	因子序号	因子名称	因子重要性程度
217	Swingindex	0.1662	210	TRIX5	0.0133
138	BIASS	0.1513	219	UIcer5	0.0128
144	CCI5	0.0867	198	AroonDown	0.0121
227	BBIC	0.0681	155	RVI	0.0117
230	MA10Close	0.0451	215	PLRC6	0.01
136	BIAS10	0.042	212	UOS	0.0093
86	REVS5	0.0365	193	plusDI	0.0092
142	CCI10	0.0299	156	SRMI	0.0076
199	AroonUp	0.0256	127	FiftyTwoWeekHigh	0.0069
137	BIAS20	0.0208	221	ACD6	0.0062
148	KDJ_J	0.0206	84	REVS10	0.006
149	ROC6	0.0179	194	minusDi	0.006
143	CCI20	0.0139	147	KDJ_D	0.0056
			146	KDJ_K	0.0055

绘制预输入的 27 个特征的热力图，以看选入模型的因子之间的相关性，其因子之间相关程度如下图。

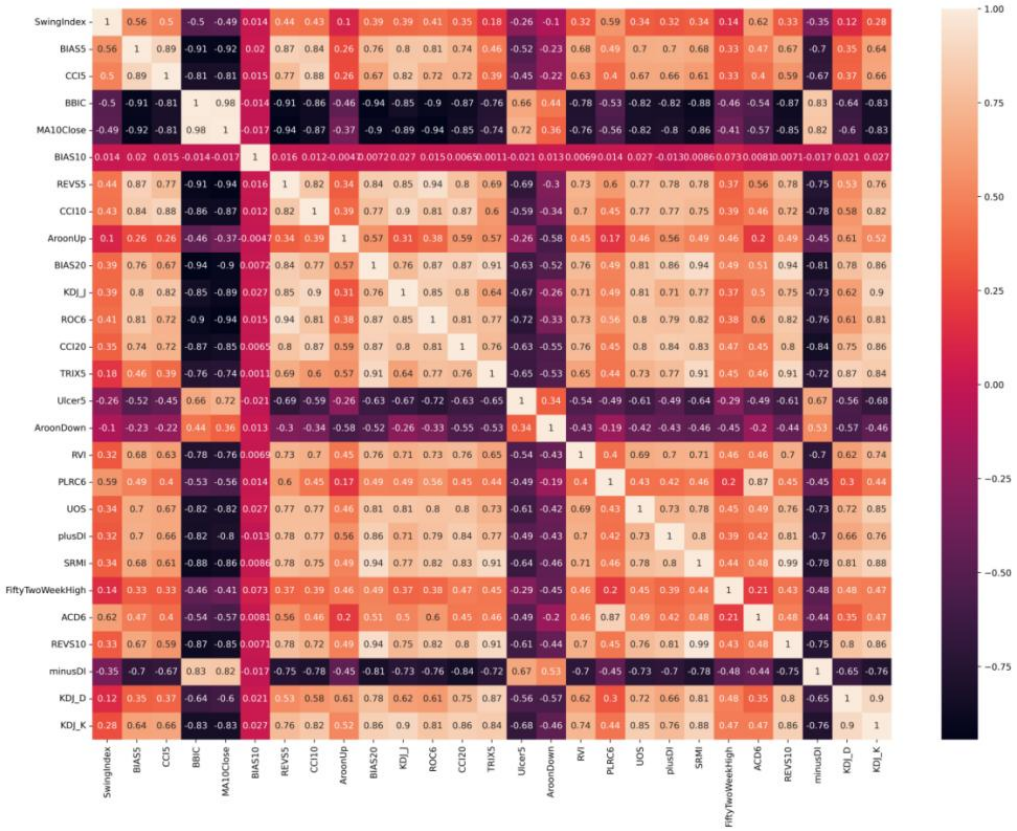


图 3-2 预输入因子相关性热力图

结合热力图剔除高度自相关的因子 7 个，分别是 MA10close，minusDi，KDJ_K，KDJ_D,SRMI,REVS10,ACD6。剔除后热力图绘制如图 3-3，最终因子为表 3-3。

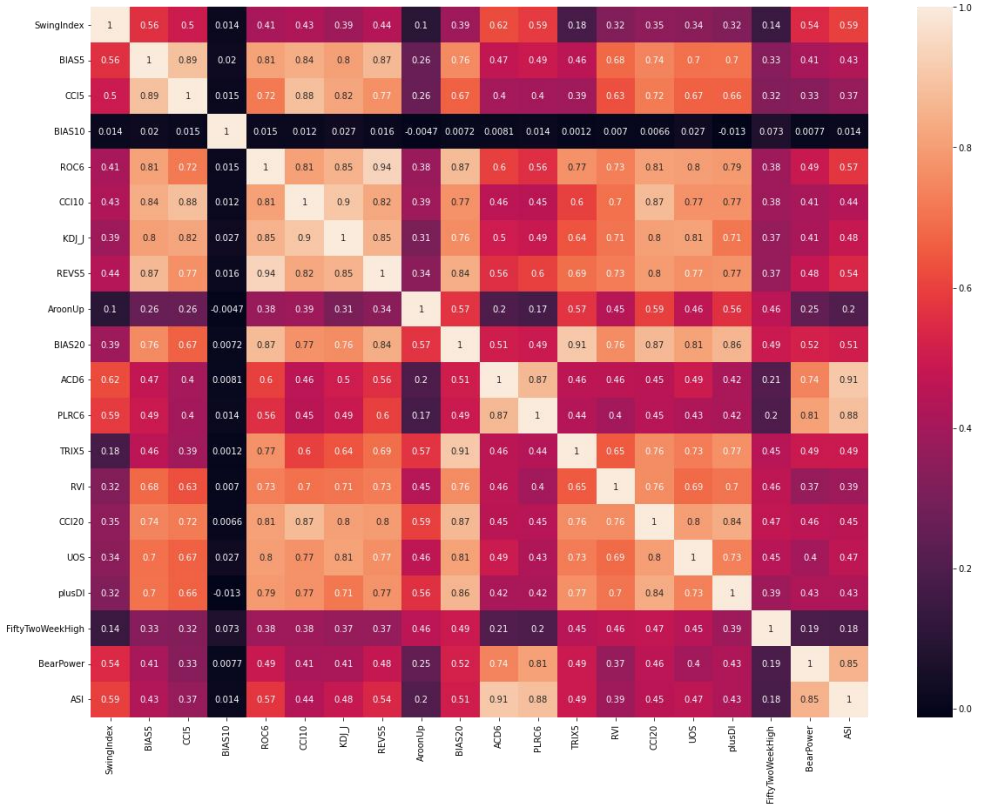


图 3-3 最终因子相关性热力图

表 3-3 最终输入因子展示

指标名称	所属类别
swingindex（摆动指标）	技术指标
KDJ（随机指标）	技术指标
BIASS（偏离率）	技术指标
CCI 指标（顺势指标）	技术指标
BBIC（多空指标）	均线型指标
REVS5（5 日收益率）	技术指标
……	……
Aroonup（阿隆上升）	技术指标
ROC（变动率）	技术指标
TRIX5（属于长线指标）	技术指标
UICER5（溃疡指数）	技术指标

本文可以在因子筛选中发现两个问题，在预测因子中，流动性技术指标表现较强。首先，流动性低的股票本来就比较与流动性高的股票难达成交易，这也是流动性因子的溢价来源，当面临股市衰退时，流动性低的资产无法处置，从而损失加大，因而风险溢价加大。

另外，由于中国二级股票市场的管理仍需完善，公司随意停盘和涨跌停板制度虽能保持金融体系稳定性，但也加大了低流动性股票的溢价程度，由此流动性特征对于未来股票价格的预测有着较大的影响。另外，由于财务类，成长类因子在数据更新变化频率低，故而对其进入重要因子有重大影响，这并不说明因子的不重要。而是在模型训练中，由于特殊的筛选机制本文暂时忽略它们，这为模型预测不好时进一步改进提供了一个方向。除此之外，即使人为的去除了相关性高的因子，由于技术指标本身的相互关联，还是无法忽视因子之间的高相关性，这也能解释为何线性预测的精度不高。由于多重共线性的存在，本文在评价预测效果时，不能用进入模型的数据去评价，于是本文单独提取 2021 年 12 月数据作为不参与模型训练的验证集，以此来保证模型结果的可靠性。

3.5 本章小结

本文只选取了编制进中证养老指数的 80 只成分股数据，剔除空值和替换异常值后筛选重点样本形成最终的股票池。能够编制进指数数据中的股票，其相较于沪深 300 的股票更为优质，所以，对于原始数据的处理工作，重点在于因子数据的处理。由于因子众多，在对缺失值，极端值进行处理后，就要对因子做一个筛选。这里是用随机森林模型对因子重要性程度进行计算并作出取舍，选取因子重要性程度>0.005 的因子。然后计算因子之间的相关程度，剔除相关程度高的因子，形成最终的因子池。

第四章 多因子选股模型构建

本章选取逻辑回归，高斯朴素贝叶斯，随机森林和支持向量机四个单一模型对其进行训练和涨跌的分类预测，并生成混淆矩阵，对正确率进行评价。最后尝试用硬投票集成来对模型进行融合提升，最终得到一个相对较好的选股模型。然后通过 CNN，RNN，LSTM，GRU 模型来对 2021 年 12 月份开盘日股票数据进行预测，对比单一模型和改进后的模型进行比较，选取最优的选股模型。

4.1 处理后的数据介绍

建立机器学习算法样本股票池。在数据处理阶段，筛选出剔除掉 ST 股票的中证养老指数成分股自 2017 年 1 月至 2021 年 12 月全部开盘日日期，股票代码及当日收盘价，当日涨跌数据，再提取涨跌程度前 20%做重点样本构建模型。样本标签为 0 和 1。

建立深度学习算法样本股票池。经过数据预处理后，可以得到个股数据以及指标因子数据的归一化处理后数据。但这不符合深度学习的样本结构，需要进一步处理，首先将单只股票按时间顺序数据对齐，然后通过滑动窗口采样的方式生成输入样本。以 20 天为窗口，1 天为步长进行采样，遍历所有样本后得到“N×20×T”的数据样本矩阵，N 为样本个数，20 为前 20 个交易日，T 为样本特征。

在建立模型之前需要对数据进行划分工作，首先人为划分数据集。为了策略严谨性和减少指标维度单一的影响，后续模型评价验证集均采用 2021 年 12 月开盘日数据，而用于模型训练的数据截至到 2021 年 11 月最后一个开盘日。这样可以检验模型的泛化能力，即其在样本外数据预测评价。用于模型训练的内部数据再通过定义划分样本内训练集和测试集，来进行模型内部的评价。

4.2 机器学习算法预测及评价

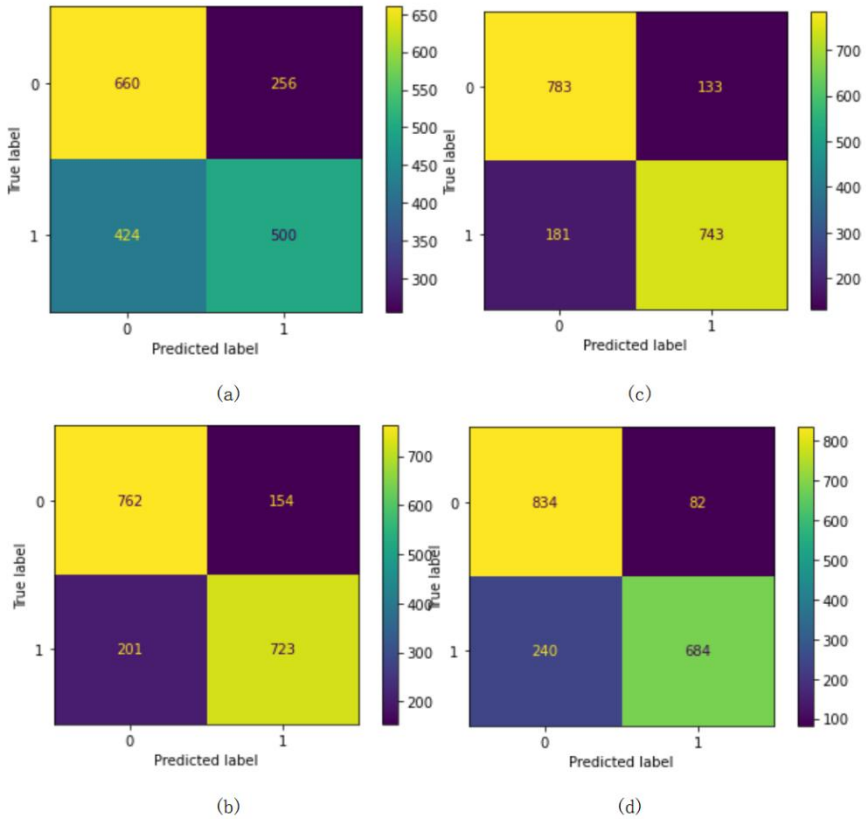
由于选入的股票数据部分因子有缺失，首先进行了缺失值的处理，最终保留了 15755 个样本，20 个因子，和 1 列涨跌标签放入模型训练。最终的模型评价在 2021 年 12 月份全部股票开盘日 1840 个数据上进行，输出正确率和预测样本的混淆矩阵。混淆矩阵经常用

来描述分类模型在已知真实值的一组测试数据上的性能，下表是一个二分类问题的混淆矩阵定义，分类问题正确率定义用（TP+TN)/全部预测样本数来定义。

表 4-1 混淆矩阵定义

	预测为+	预测为—
实际为+	TP（被正确划分到正例的个数）	FN（被错误地划分到负例的个数）
实际为—	FP（被错误地划分为正例的个数）	TN（被正确划分到负例的个数）

将放入模型的全部股票涨跌标签和其因子数据进行模型训练集与测试集的划分，划分比例为 7:3。并利用其训练出的模型在最初的验证区间内进行验证，即 2021 年 12 月份全部股票开盘日。由于这些数据从未进入模型训练，故其结果有很强的可靠性，最终验证集有 80×23 个数据。得到四个单一模型的预测结果，混淆矩阵如下图所示。



(a) 朴素贝叶斯分类预测结果 (b) 逻辑回归分类预测结果 (c) 随机森林分类预测结果 (d) 支持向量机分类预测结果

图 4-1 单一模型预测结果

依据正确率的计算公式，对这四个单一模型分类预测结果进行了对比，结果如下表所示。

表 4-2 单一模型预测准确率比较

模型名称	准确率
高斯朴素贝叶斯	0.6304347826086957
逻辑分类	0.8070652173913043
随机森林	0.8252314357658709
支持向量机	0.8293478260869566

对比发现单一模型预测最好的是支持向量机，准确率为 0.829，其与随机森林预测精度差不多，这个结果初步可以。这说明样本的处理和因子选取不存在太大问题，下面将这四个单一学习器融合在一起分析。由于是分类问题，本文采取了最简单的硬投票尝试是否会对结果有提升。硬投票采用的是少数服从多数的原则，即多个学习器有一半以上学习器输出为同一结果，则输出该结果。本文尝试了多种学习器组合下的预测，最终的预测表现如下表所示。

表 4-3 硬投票集成学习在预测样本上的表现

模型名称	准确率
逻辑分类-朴素贝叶斯-随机森林	0.85217
逻辑分类-随机森林-支持向量机	0.86793
朴素贝叶斯-逻辑分类-支持向量机	0.85435
朴素贝叶斯-随机森林-支持向量机	0.86196

比较准确率可以确定，逻辑分类-随机森林-支持向量机的集成硬投票效果最好，且较单一预测模型有了明显提升，其输出混淆矩阵如图 4-2 所示。

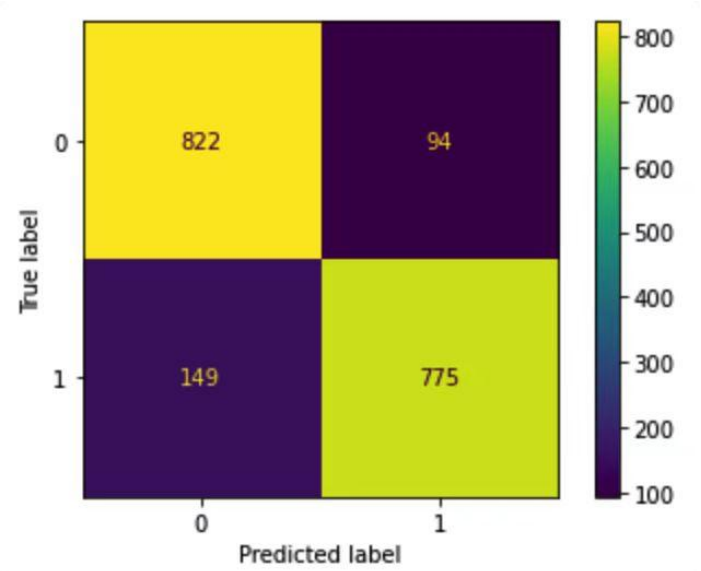


图 4-2 逻辑分类-随机森林-支持向量机的硬投票预测展示

计算后输出的正确率为：0.86793。本文将此模型用于在验证集 2021 年 12 月的预测，将预测结果数据提取。预测结果部分展示如下表所示。

表 4-4 验证集预测结果

股票代码	预测日期	涨跌标签
600054	2021/12/1	1
600054	2021/12/2	0
600054	2021/12/3	0
600054	2021/12/6	0
600054	2021/12/7	0
600054	2021/12/8	1
600054	2021/12/9	1
600054	2021/12/10	1
600054	2021/12/13	0
600054	2021/12/14	0
600054	2021/12/15	0
600054	2021/12/16	0
……	……	……

本文想要得到未来一年的预测数据，时间区间为 2021 年 12 月到 2022 年 12 月。于是将上述时间区间内相同因子数据下载用于预测，以最终预测效果最好的机器学习模型作为最终的机器学习选股模型，其选出的建仓股票流程如下：

1) 输出未来一个月开盘日涨跌标签，提取其概率；

2) 计算每只股票未来一个月开盘日预测为 1 的概率均值；

3) 提取均值最高的十只股票作为初次建仓股票；

4) 轮动选股，继续预测下一个月开盘日涨跌即提取概率，并同样筛选出建仓的十只股票；

5) 对比上次的建仓股票，两次都在继续持有，不在新的建仓股票池做平仓处理，新出现的股票开新仓处理。

按照上述方法，本文最终选出 2021 年 12 月到 2022 年 12 月每月优秀股票代码如下表所示。

表 4-5 按月选取的股票代码

1	2	3	4	5	6	7	8	9	10	11	12
603392	601601	603486	002511	605499	605499	300142	600054	605499	688169	300741	300251
603345	002001	002511	300957	688363	300896	603896	600138	603345	600745	300142	300413
300760	300144	603444	300529	688036	603486	300122	300144	002602	603345	605499	603896
300676	603156	688169	300896	600655	600436	300957	600754	600763	002624	300832	000963
002739	300122	300957	300122	603486	688363	603392	300142	002555	002242	002032	689009
688169	600138	300529	688169	300015	300741	600196	603156	601628	600754	2624	002739
002925	000963	600754	688363	603882	603392	600161	002602	600887	002007	600196	300418
603605	688036	000661	603605	600436	603605	600436	603087	689009	603882	300251	000538
600745	300142	002867	000963	300146	300957	300418	300418	601888	300896	300418	002555
603486	600763	603605	300142	600196	688036	300015	601601	002867	002705	603444	600754

4.3 深度学习算法预测及评价

输入样本矩阵 20×20，其代表同一只股票连续 20 日，20 个因子的数据。样本标签为其第 21 日的收盘价，由此滑动选股，即同一只股票第一日到地二十日因子数据为一个样本，第二日到第二十一数据为一个样本，由此类推并同样的方法提取全部股票的样本矩阵。利用深度学习 RNN，LSTM，GRU 进行预测，选择最好的深度学习方法做选股模型，预测全部股票自 2021 年 12 月的收盘价日行情。本文会基于 RNN，LSTM，GRU 和改进模型来实现股票单收盘价指标的训练和预测。

4.3.1 数据处理与模型装配

选取最终的 20 个因子和收盘价作为标签输入，并以 20 天为一个时间步，滑动窗口取

样，即第一天至第二十天，20 个因子作为一个输入矩阵，第二天至第二十一天，20 个因子作为一个输入矩阵……由此滑动采样，遍历样本生成。特征构建完成后的数据结构为 N 个“20×20”像素的特征图像，其中前一个 20 为标签日前 20 个交易日，后一个 20 为加入因子数据，N 为滑动窗口取样后的样本个数，输入全部 2017 年 01 月 03 日至 2021 年 11 月 30 日的股票数据，以 7：3 划分内部数据集来对模型的拟合和泛化能力进行评价。

由于不同股票涉及到的因子量不同，会造成部分数据的缺失，为了不影响后续的模型构建，将缺失值全部填充为 0.0001。

为了后续模型训练，需要对模型进行划分。以单只股票 000526 为列，将日期在 2021 年 12 月 01 日之后的数据作为模型的测试集，用于后续检验模型预测效果。由于模型的构建过程中也需要检验模型的配置，以及训练程度是过拟合还是欠拟合，日期在 2021 年 12 月 01 日之前将在后续拆分为训练集与验证集。训练集用于训练得到神经网络模型，然后用验证集验证模型的有效性，挑选获得最佳效果的模型。验证集可以重复使用，主要是用来辅助构建模型的。最后，当模型“通过”验证集之后，再使用测试集测试模型的最终效果，评估模型的准确率，以及误差等。此外，不能将测试集用于模型训练，因为随着模型训练的进行，网络会慢慢过拟合测试集，导致测试集的结果没有参考意义。将训练机划按照 trade date 参数顺序分为训练集与验证集，划分比例 7:3 划分好数据后将 trade date 时间因子删除，因为该因子表示的是日期，后续不能放入网络模型里。

定义分区函数，给定一个参数 n，训练集与验证集的最后那个参数作为模型训练时预测的值，每个值都由训练集与验证集的数量个数减去参数 n 的数据来预测原始数据，同理测试集。故将训练集与验证集通过分区函数，训练集将会有 len(训练集)-size 对数据用与模型训练，并用 len(验证集)-size 对数据进行验证，最后通过 len(测试集)-size 组数据进行测试。

在训练网络时，一般的流程是通过前向计算获得网络的输出值，再通过损失函数计算网络误差，然后通过自动求导工具计算梯度并更新，同时间隔性地测试网络的性能。对于此类训练逻辑，可以直接通过 Keras 提供的模型装配与训练等高层接口实现。

基础的 LeNet5 模型有卷积层，池化层，全连接层，网络层，参数设置参考已有研究，采用 Adam 优化器，采用均方误差作为模型的损失函数，将训练集通过分区函数后分别放入 fit()函数进行模型训练，epoch 设置为 10，validation_data 设置为验证集通过分区函数得

到的数据组合。训练完成后即可进行模型测试。

由于真实数据的分布往往是未知且复杂的，无法推断出其分布函数的类型和相关参数。因此在选择学习模型的容量时，往往会根据经验值选择稍微大的模型容量，当模型的容量过大时，网络模型除了学习到训练集数据的模态之外，还把额外的观测误差也学习进来，导致学习的模型在训练集上表现较好，但是在未见的样本上表现不佳，也就是模型泛化能力偏弱，即造成了过拟合现象。在本次研究出现过拟合现象时，采用提前停止与正则化方法来处理[32]。

4.3.2 RNN 模型建模及预测结果

RNN 模型建模预测分为如下步骤进行：

1) 输入测试集，输入数据前 2426-300=2126 天数据；

2) 利用 for 循环，遍历整个训练集，提取训练集中连续 20 日的收盘价和因子特征作为输入特征，弟 21 天的数据作为标签，共计 78×2126 组数据；

3) 对训练集进行打乱，将训练集由 list 格式改为 array 格式；

4) 使送入训练集符合 RNN 输入要求：送入样本数×循环核事件展开步数×每个时间步输入特征个数。

RNN 模型装配共分为五层。使用 RNN 模型时共使用了两层 SimpleRNN 层与一层全连接层，参数设置下图所示，SimpleRNN 的输出维度为 80 与 100，为了防止过拟合在层之间装配 Dropout 层，参数为 0.2，通过 RNN 循环网络数据后连接全连接层，节点数为 1。模型使用 Adam 优化器，学习率设置为 0.001，损失函数为均方误差，网络层如下表所示。

表 4-6 RNN 网络层设置

Layer	参数设置
SimpleRNN	80
Dropout	0.2
SimpleRNN	100
Dropout	0.2
Dense	1

在模型训练过程中随迭代次数绘制损失函数图 4-3，在经历约 12 个 epoch 后，损失函数值平均损失函数值已达到 loss: 2.0044e-04 - val_loss: 5.4920e-04。

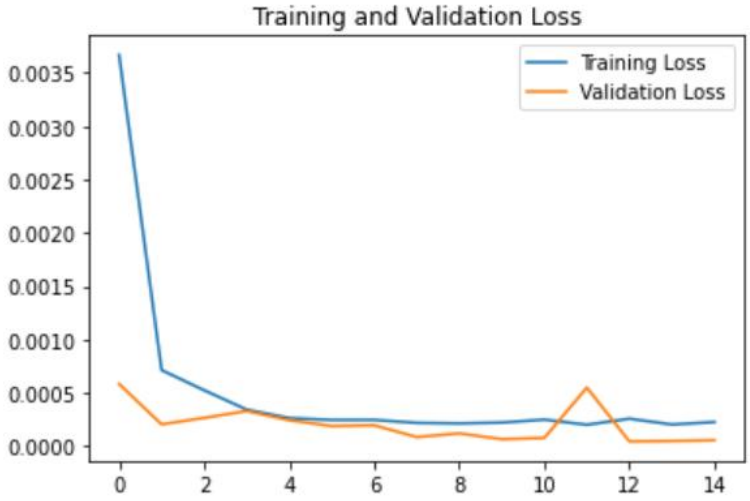


图 4-3 RNN 损失函数图

将预测结果在验证集上进行比较，结果在图 4-4 中进行展示。其中红色是真实股价，蓝色是预测股价，展示的是全部成分股在 2021 年 12 月全部开盘日的预测与真实情况。可以看到，二者走势基本吻合，除个别极值不精准以外，这可能因素是引起股价大幅增长的因素众多，我们未将其纳入因子池中。

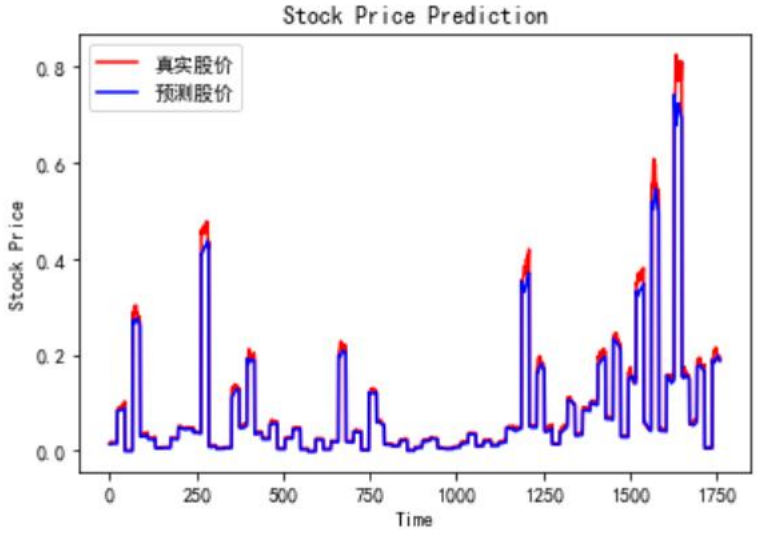


图 4-4 RNN 预测结果与真实值比较展示

RNN 模型的预测评价指标依旧用平均绝对误差和平均绝对百分比误差来比较。输出的误差结果如下表所示。

表 4-7 RNN 误差结果

评价指标	结果
MAE 平均绝对误差	46.28698236744972
MAPE 平均绝对百分比误差	40.72311775445892

4.3.3 LSTM 模型建模及预测结果

与 RNN 同理，使用 LSTM 训练模型时装配两层 LSTM 神经网络与一层全连接层网络，模型整体结构与参数与 RNN 保持一致。LSTM 中设有 LSTM 层，全连接层，Dropout 层，分别用来对输入数据进行时间记忆和丢弃，学习输出，随机剪枝防止过拟合。

表 4-8 LSTM 网络层设置

Layer	参数设置
LSTM	80
Dropout	0.2
LSTM	100
Dropout	0.2
Dense	1

损失函数同样使用均方误差，损失函数随每个 epoch 迭代变化如图下所示，观察可知相较于 RNN 神经网络，训练集的优化速度更快，测试集的起伏趋势较小，LSTM 能很完美的做到提取有用信息遗忘无用信息。

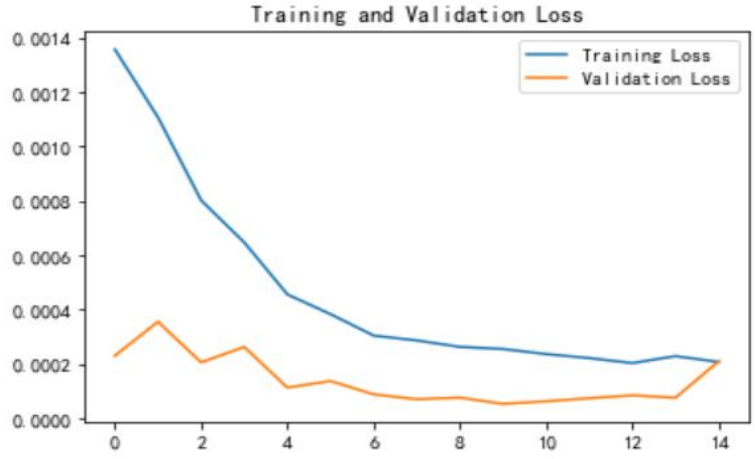


图 4-5 LSTM 损失函数图

测试集预测数据对比如图 4-6 所示，可知 LSTM 依然可以根据 20 天数据预测出股价的走势，但拟合效果较强于 RNN，评价指标值也优于 RNN 神经网络。

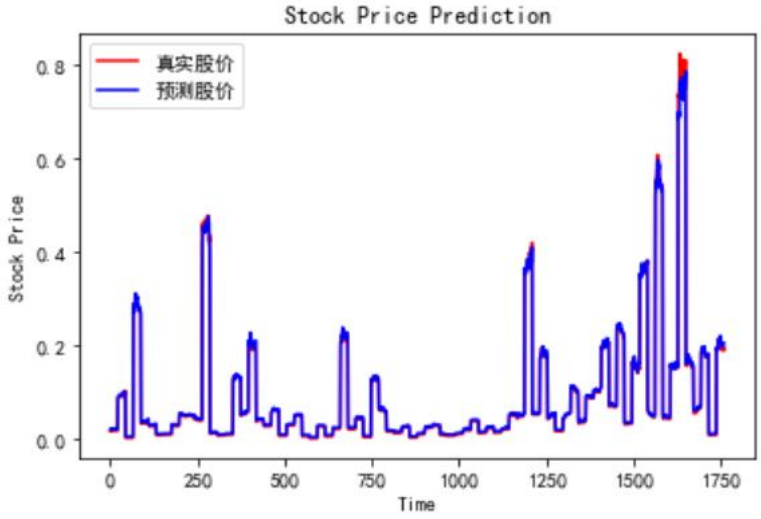


图 4-6 LSTM 预测结果与真实值比较展示

LSTM 模型输出的误差结果如下表所示。

表 4-9 LSTM 误差结果

评价指标	结果
MAE 平均绝对误差	25.908147765520066
MAPE 平均绝对百分比误差	39.779455749141995

4.3.4 GRU 模型建模及预测结果

模型装配与 RNN 与 LSTM 上述相同，网络层换成 GRU 层，为了更好地区分效果，网络层参数结构保持不变。GRU 只有两个门控开关，参数数量小于 LSTM 但是可以实现与 LSTM 接近的效果，并且可以解决 RNN 不能长期依赖的问题。损失函数用均方误差，损失函数随每个 epoch 迭代变化如图 4-7 所示。可知训练集的优化速度与 LSTM 大致保持一致，可以看到在十次迭代之后，损失程度趋于稳定在 0.03 左右，但橙色的验证集会出现损失程度上涨的现象，训练集的优化出现了随着 epoch 的增长损失函数起伏逐渐增大的情况。

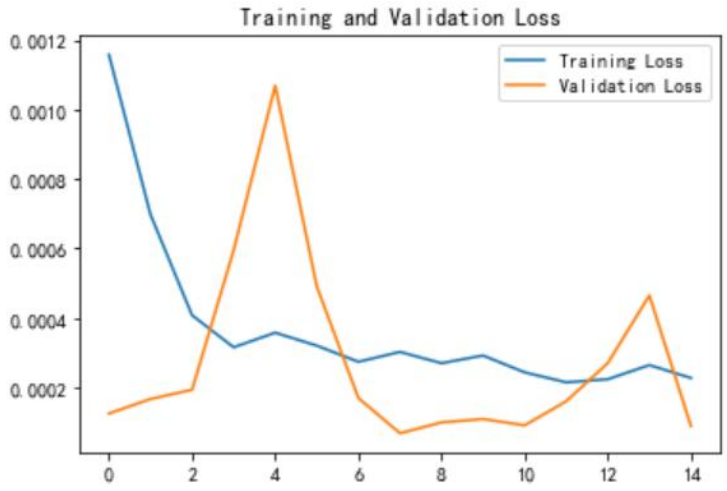


图 4-7 GRU 损失函数图

测试集预测数据对比图如下，可知使用 GRU 神经网络预测的拟合的效果较差，预测效果优于 RNN，劣于 LSTM。

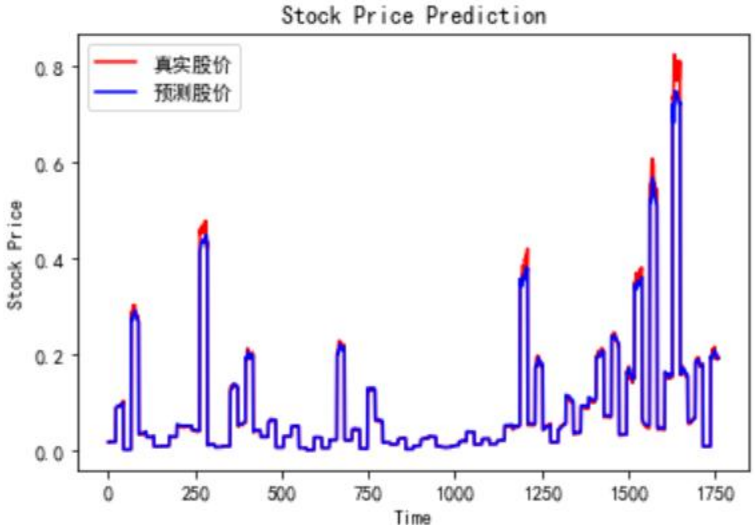


图 4-8 GRU 预测结果与真实值比较展示

为便于比较，GRU 模型的预测评价指标同样用平均绝对误差和平均绝对百分比误差来比较。输出的误差结果如下表所示。

表 4-10 GRU 误差结果

评价指标	结果
MAE 平均绝对误差	39.2771457658902
MAPE 平均绝对百分比误差	40.67909565392468

上述内容使用了三种神经网络模型对中证养老指数收盘价进行预测，从拟合效果上看

LSTM 拟合效果最好，RNN 次之，GRU 较差，模型依旧存在过拟合的问题，需要后续研究中对参数进行调整。LSTM 的拟合效果较好，可以用此方法对收盘价进行辅助预测。此外本文只考虑了历史收盘价与模型选入因子的特质影响，没有考虑到其他特征可能对收盘价的影响。

在单一模型中最终选用最好的 LSTM 来进行预测，本文先展示单只股票在 2021 年 12 月份的预测效果，由于在实际操作中无法获知当日或者未来因子数据，而输入二十个日数据可以预测第二十一天的收盘价，由此在预测时编入了循环程序。即预测出的 2021 年 12 月第一个开盘日数据后再次放入模型训练，得到第二个开盘日数据，由此类推，直至循环 22 次，得到全部 12 月份 23 个开盘日收盘价预测。我们展示模型在 2021 年 12 月份 23 个开盘日对股票代码为 000526 的预测数据如下表。

表 4-11 LSTM 预测结果

序号	预测值		
1	18.0879	13	19.9753
2	17.9866	14	19.5731
3	18.8524	15	19.2279
4	18.7082	16	19.7290
5	18.6172	17	20.0782
6	18.9082	18	18.4976
7	20.0124	19	18.6420
8	20.6909	20	18.8945
9	19.3707	21	18.7371
10	19.2234	22	19.1682
11	20.0986	23	18.9680
12	20.2287		

将预测值与真实值对比可以得到模型对单只股票 000526 在 2021 年 12 月份 23 个开盘日的预测效果，如图 4-9 所示。并计算预测的均方误差为 0.1794。

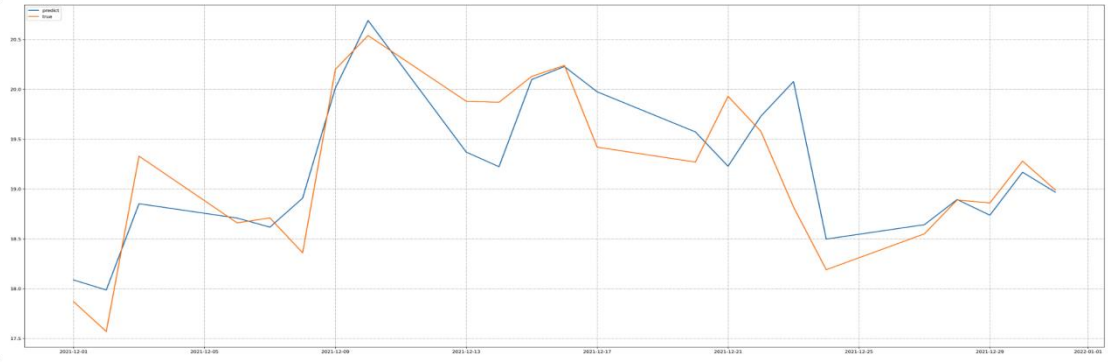


图 4-9 LSTM 预测结果与真实值比较展示

4.3.5 CNN-LSTM 模型建模及预测结果

CNN 能够通过使用卷积核从样本数据中提取出其潜在的特征，而长短期记忆网络 LSTM 能够捕捉到长期的成分。卷积神经网络由卷积层和池化层交替叠加而成，在每个卷积层与池化层之间都有 RELU 激活函数作用来加速模型的收敛。在模型中数据经过卷积神经网络的处理，所有特征融合后得到卷积神经网络的特征描述，此时传递数据给 LSTM，通常情况下此时输入的数据需要 reshape 成 LSTM 处理的类型。LSTM 得到新的输入后，确定需要保持与丢弃的，其中保持与丢弃借助 Sigmoid 激活函数完成，任何数据乘 1 都为其本身，而任何函数乘 0 都为 0，即可选择遗忘以及保存重要的数据。当数据乘 1 时则代表被保留，从输入门中获取的数据即为我们更新了状态。最后借助输出门确定携带的信息，将新的状态，以及隐藏状态转移到下个时间步。

借鉴已有研究，本文将先搭建 CNN-LSTM 模型的基本框架。该框架由两部分组成，第一部分为 CNN 神经网络，第二部分为 LSTM 神经网络，两个神经网络架构之间通过 Lambda 层来连接，Lambda 层可以实现对数据结构进行调整，让 CNN 神经网络的输出数据经处理后可以输入至 LSTM 层中。

卷积神经网络设计部分以经典的 CNN 神经网络构架 LeNet-5 为基础，由于传入 CNN 神经网络模型的样本结构为“20×20”的二维样本图片，卷积核在两个维度上提取特征。因此使用 Tensorflow.Keras 库中的二维卷积 Conv 2D 函数来构建卷积层池化层是卷积神经网络的核心要素，考虑到本文输入特征因子数量较多，因此在卷积层后面加入池化操作，LeNet-5 架构中有两次卷积操作。因此，本文同样设置两个卷积层和 1 个池化层，第二个池化层后面接一个卷积层，这个卷积层的卷积核为 1，目的是为 Lambda 层的输入做数据结构调整准备。在 Lambda 层之后接 LSTM 层，在 LSTM 层数设置上，先将 LSTM 层数暂定为 1 层，在 LSTM 层之后再接两层全连接层，第一个全连接层用于对前面各层网络提取的特征做组合，第二个全连接层用于将特征信息转换为预测输出。

在 LSTM 前加入 CNN 网络层能够对单一的 LSTM 有一定的改进，CNN 能够对因子特征进行充分的处理和提取，与 LSTM 的连接只需要 reshape 成 LSTM 需要的输入结构即可。CNN-LSTM 网络层结构如下表所示。

表 4-12 CNN-LSTM 网络层设置

Layer	参数设置
Conv2D	4
kernel_size	2
Conv2D	4
kernel_size	2
AveragePooling2D	2
Dropout	0.3
Flatten	0
LSTM	16
Dense	8
Dropout	0.3
Dense	8
Dropout	0.3
Dense	1

绘制损失函数，损失函数随每个 epoch 迭代变化如图 4-10 所示。可以观察到在六次迭代之后，验证集的损失平稳的控制在了 0.0002 以下。

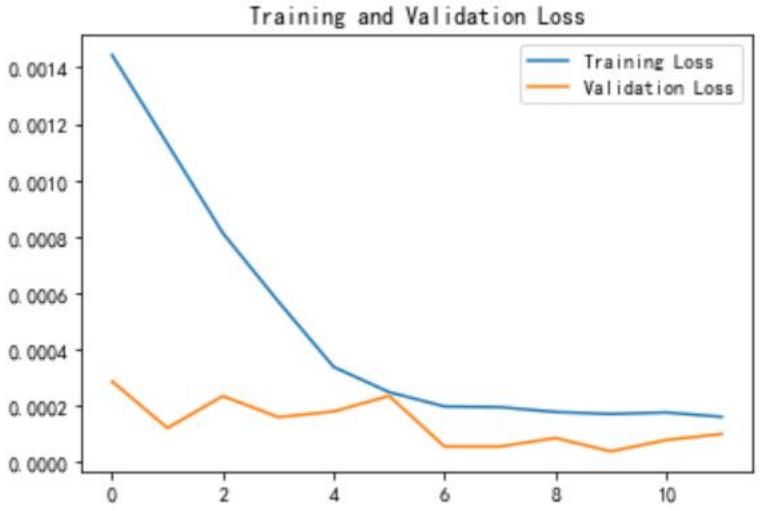


图 4-10 CNN-LSTM 损失函数图

测试集预测数据对比图如图 4-11，可知使用 CNN-LSTM 神经网络预测的拟合的效果较单一的模型直观上看更好，预测效果优于单一的 LSTM。

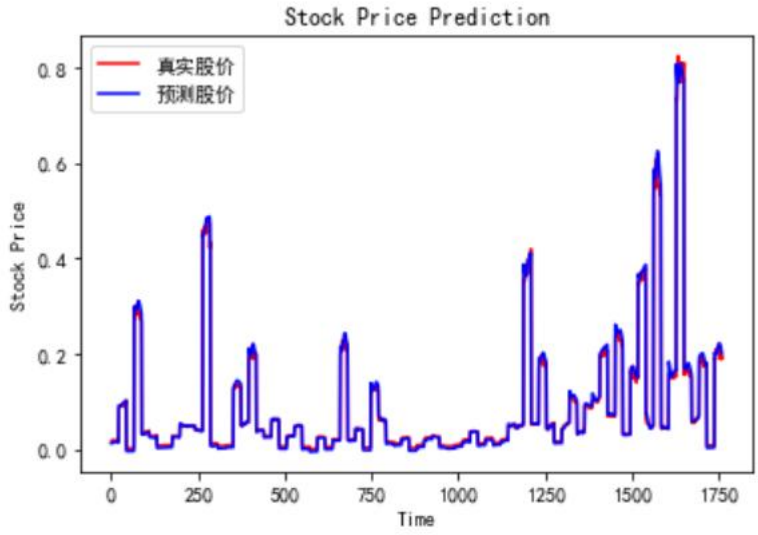


图 4-11 CNN-LSTM 预测结果与真实值比较展示

表 4-13 CNN-LSTM 预测误差

评价指标	结果
MAE 平均绝对误差	22.968040403578094
MAPE 平均绝对百分比误差	33.542938464393063

我们可以在上表中更直观的比较 CNN-LSTM 模型较单一的任一模型误差都有了明显提升。

4.3.6 CNN-GRU 模型建模及预测结果

CNN-GRU 模型是 CNN 网络的特殊表现形式，作为深度学习领域的常用网络模型之一，其搭建原理必须满足多层前馈神经调节制度[33]。CNN 网络的基础应用结构包含多个池化层与卷积层单元。其中，池化层单元负责维持原始数据信息的矩阵维度特征，并可以根据复杂度指标的数值水平，对已存储数据信息参量进行处理；卷积层单元可以根据偏执性指标的赋值结果，确定各层神经元节点之间的实时连接关系，从而在控制计算参数产出水平的同时，提升 CNN 网络对于数据信息参量的训练处理速度。与常规 CNN 网络结构相比，CNN-GRU 网络模型的池化处理能力更强，既能维护数据信息参量指标的传输完整性，也可以避免残差指标的出现，促进网络主机对于池化节点中已存储信息参量的按需处理能力。

CNN-GRU 框架是股票预测网络的基础连接结构，由输入层、输出层、中间过渡层三

类单元共同组成。输入层负责对数据信息参量进行分拣处理，并可以将满足调取需求的信息指标反馈至下级单元连接结构之中。中间过渡层包括多层节点组织，同时与 CNN 单元相对应，负责记录信息的实时传输状态；第二层节点组织对应 GRU 单元，可以更改信息存储形式，在整个 CNN-GRU 框架中起到传输过渡作用。输出层能够接收由 CNN 单元制定的数据信息，并可以将这些信息参量反馈给其他结构。CNN-GRU 网络层框架设置如下表。

表 4-14 CNN-GRU 网络层

Layer	参数设置
Conv1D	32
kernel_size	1
Dropout	0.2
GRU	100
Dropout	0.2
Dense	1

损失函数随每个 epoch 迭代变化如图所示。

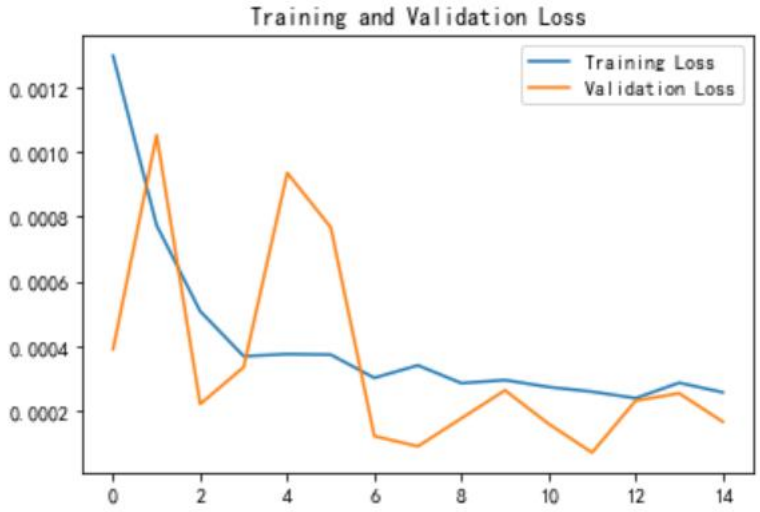


图 4-12 CNN-GRU 损失函数图

绘制测试集预测数据对比图如图 4-13 所示，可知使用 CNN-GRU 模型的拟合效果较差，尤其是对极值的预测。计算输出 CNN-GRU 的误差值如表 4-15 所示。

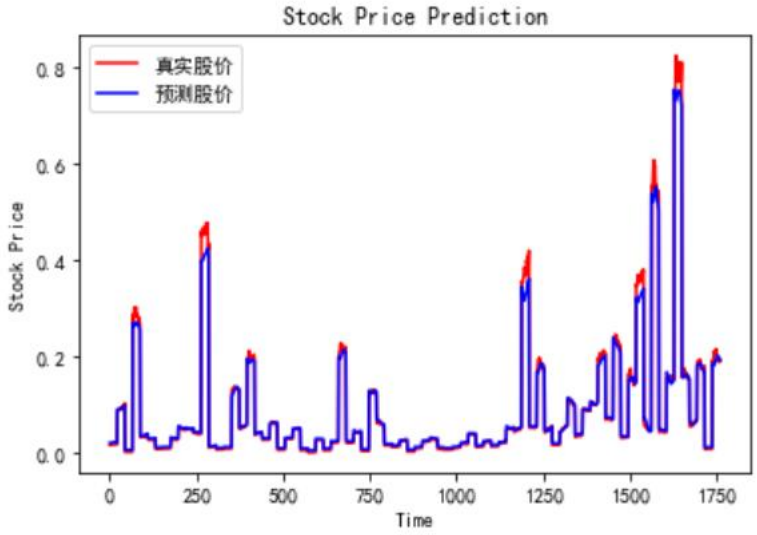


图 4-13 CNN-GRU 预测结果与真实值比较展示

表 4-15 CNN-GRU 预测误差

评价指标	结果
MAE 平均绝对误差	29.427628977395614
MAPE 平均绝对百分比误差	42.89117156861812

比较改进后的模型，可以发现引进注意力机制的 CNN-LSTM 模型在测试集上的预测误差更好，我们选择此模型进入股票预测。CNN-LSTM 模型训练参照 LSTM 实际预测，CNN-LSTM 模型在 2021 年 12 月份 23 个开盘日预测数据如下表。

表 4-13 CNN-LSTM 预测结果

序号	预测值		
1	18.0348	13	19.2190
2	18.3300	14	18.7279
3	19.4276	15	20.3841
4	19.1189	16	19.3014
5	18.5381	17	19.1916
6	18.9961	18	18.3920
7	20.5824	19	19.1016
8	20.4198	20	18.7944
9	19.7641	21	19.1214
10	19.2765	22	19.5130
11	19.9323	23	18.8162
12	20.1035		

将预测值与真实值对比得到模型对单只股票 000526 在 2021 年 12 月份 23 个开盘日的预测效果，如图 4-14。计算输出模型预测的均方误差为 0.1351。

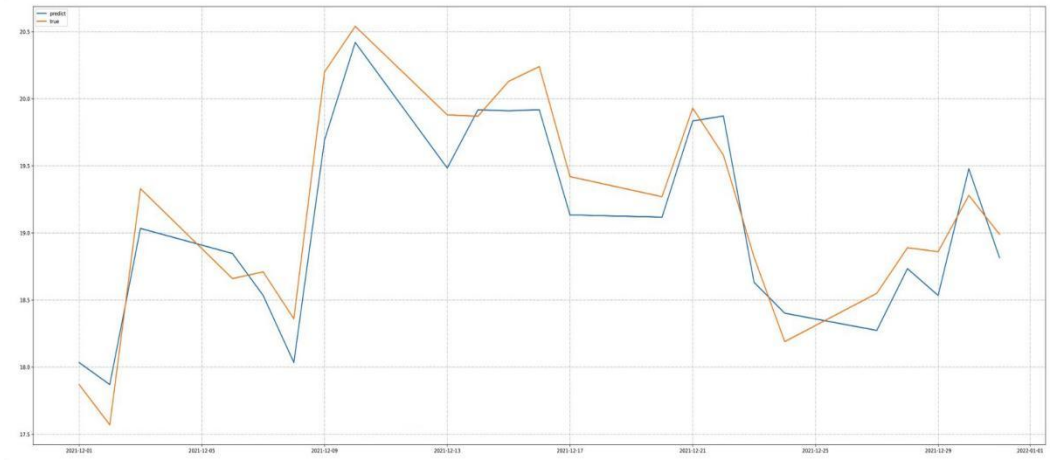


图 4-14 CNN-LSTM 预测结果与真实值比较展示

能够得出无论是模型测试还是实际预测效果，CNN-LSTM 预测模型都是较优的，所以我们选取深度学习中的 CNN-LSTM 模型来预测，并输出时间和股票代码和预测股价。我们想要得到未来一年的预测数据，时间区间为 2021 年 12 月到 2022 年 12 月。采用轮动选股的方法，即将同一只股票按照预测值持有 20 日，计算持有收益，（第 20 日收盘价-第 1 日收盘价）/第 1 日收盘价，为收益率，由此计算全部股票持有 20 日收益率，并筛选收益率最高的十只股票进行持有。轮动选股，下一个 20 日周期以同样方式筛选，对于换手的方式同机器学习。

4.4 本章小结

本章首先是对样本数据进行筛选，选取重要样本，对于深度学习模型的构建，则是采用滑动窗口采样，将数据结构调整。然后遵循机器学习做模型预测的基本流程，做数据的处理，模型的训练，模型的评估和最终的预测。并建立了模型预测的评价指标，并从中筛选了最好的单一模型支持向量机，准确率为 0.829。随后本文做了集成模型硬投票的尝试，并发现逻辑分类—随机森林—支持向量机的集成硬投票效果更优，较单一的模型预测，准确率提升到了 0.86793。于是在此选股模型上建立轮动选股策略筛选股票。然后，引入深度学习算法，意在与上述选股策略进行比较来选择最终模型。首先，由于数据结构不匹配深度学习算法，本文采用滑动窗口采样来调整数据结构。然后调整数据参数，通过 CNN，

RNN，LSTM，GRU 模型来对 2021 年 12 月份开盘日股票数据进行预测。本文中 2021 年 12 月份及以后数据在所有模型处理中都未被放入模型训练，其结果有参考价值。结果表明，LSTM 算法的 MAE,MAPE 指标都较优。本章又引入 CNN-LSTM 来对单一的算法进行改进，结果表明，其误差数据得到了改进，于是选择 CNN-LSTM 来作为深度学习选股模型。

第五章 量化选股模型回测

第四章在模型评价指标比较中完成了对选股模型的筛选来形成选股的策略，此章节将在金融领域中利用回测来对策略进行收益和风险等层面的评价，并加入择时来进一步提升量化策略的优越性。

5.1 回测与指标评价

一个策略好不好，就是验证在按照策略进行产品投资并持有一段时期后，是否可以实现超额收益，这个超额收益有多少，以及这个策略的风险有多少。而现实中不可能真的持有资产来进行验证，于是就有了策略的回测。即假设初始持有资产，设置买卖股票操作，假定手续费，并在选择的历史区间内假设持有来计算最终的评价指标。策略回测目前可以借助于金融量化平台来进行，回测指标主要包括以下内容。

累计收益率:累计收益率是经过投资股票或组合代表在回测区间内获得的收益结果，代表回测时间段中的策略表现。

最大回撤：描述策略可能出现的最糟糕的情况，即某段时间内产品价值从最高点开始回落到最低点的幅度，风险指标。

策略年化收益率：表示投资期限为 1 年的预期收益率，收益指标。是投资组合的总收益率通过换算化为以年为单位的收益率。其数值并未考虑策略风险，是纯收益指标。

阿尔法：投资中面临着系统性风险（Beta）和非系统性风险（Alpha)，Alpha 是投资者获得与市场波动无关的回报，一般用来度量投资者的投资策略好坏[34]。Alpha 衡量了股票或组合相对于市场的超额收益，可以获得的与市场波动部分无关的回报。Alpha 的正负反映了策略的表现与大盘的表现之比。若 Alpha=0 策略表现与大盘持平；若 Alpha>0，则说明策略表现优于大盘表现，该策略组合可以获取一定的超额收益。

贝塔:表示投资的系统性风险，反映了策略对大盘变化的敏感性。例如一个策略的 Beta 为 1.3，则大盘涨 1%的时候，策略可能涨 1.3%，反之亦然。

夏普比率：表示每承受一单位总风险，会产生多少的超额报酬，可以同时对策略的收益与风险进行综合考虑。

策略收益波动率：用来测量资产的风险性，表示投资组合的波动风险，波动越大代表

策略风险越高，是风险指标。

5.2 量化选股模型回测结果评价

利用多因子选股策略最终选出十只股票作为投资目标持有，持有期为一个月，轮动选股，持续持有一年，观察回测指标，回测区间为 2021 年 12 月 1 日至 2022 年 12 月 1 日。

首先我们来观察下机器学习最终的 LR-SVM-RF 集成硬投票量化选股的回测结果图 5-1。回测在聚宽平台上进行，其中红线为沪深 300 基准收益，其回测期基准收益为-19.40%，蓝线表示策略收益曲线，而我们的选股策略收益为 19.9%。

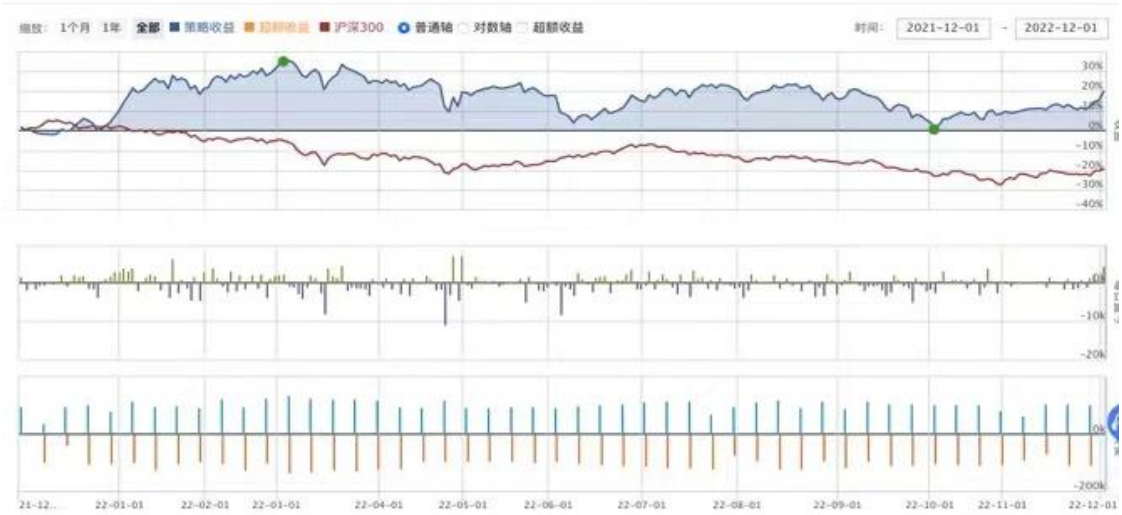


图 5-1 机器学习选股模型回测图

下面我们来观察深度学习 CNN-LSTM 量化选股模型的回测结果图 5-2，深度学习的选股策略较机器学习在收益方面更为优越，其策略收益达到了 32%。

图 5-2 深度学习选股模型回测图

为更直观，更全方面的比较两个策略，我们将回测指标汇总整理如下表 5-1 进行比较。表中可以看出，在风险指标不相上下的情况下，CNN-LSTM 模型的收益指标明显要高于机器学习模型，这可能由于模型本身的差异，也可能由于我们在选取时利用了数值预测较分类预测更为精准。

表 5-1 选股模型回测指标比较

策略名称	策略收益	策略年化收益	超额收益	α	β	夏普比	最大回撤	策略波动率
LR—SVM—RF	19.90%	20.44%	48.76%	0.324	0.671	0.529	25.16%	0.311
CNN—LSTM	32%	32.91%	63.77%	0.444	0.65	0.964	25.20%	0.3

5.3 加入量化择时进行回测

5.3.1 量化择时理论与研究

量化择时与量化选股都是结合历史数据研究来预测资产价格走势，但是两者目的不同。量化选股目的在于选出表现优异的股票来持有，而量化择时却是确定投资产品组合的买卖时机。金融市场变化莫测，动荡剧烈，买卖时机的确定尤为重要。

量化择时其中的一个关键步骤，在于模型的选择，根据所获取的历史数据选取合适的模型进行预测，来判断买卖时机，从而改善策略。关于量化择时方法和理论的研究，国内外也有众多。KIM 最早在 2003 年就选用技术指标作为特征进行择时研究，并且在回测中得到了显著的验证[34]。Wei Huang 等在 2005 年建立支持向量机模型预测单只股价走势构建择时策略，并于线性预测进行比较，结果要优于线性模型[35]。近年来，量化择时的研究也转向了神经网络模型。Deng Y 等人在 2016 年建立了金融信号的递归神经网络，并在期货市场上证实了其鲁棒性[36]。而我国学者对于量化择时的研究相对较晚，王春丽等 2018 年对上证 180 指数进行择时研究，通过回归构建多因子模型，然后设计择时方法来减少资产投资风险[37]。李琳洁（2018）基于隐马尔可夫模型对我国证券市场进行择时策略研究[38]。同样，王峰虎也将此方法用于了我国黄金期货的择时研究[39]。

本文尝试研究对我国大盘上证指数进行预测，来判断未来走势，从而做出是否执行买入操作。参考王燕和郭元凯通过改进的 XGBoost 对单只高换手率股票预测的表现，本文选用此模型来对大盘指数进行预测[40]。

5.3.2 XGBoost 大盘择时

随机森林支持对列进行采样，XGBoost 将该思想引入，并进一步减少了计算量，能有效的避免过拟合。XGBoost 根据结点分裂前后的增益差来决定是否分裂，当完全生成一棵树时，对其剪枝以防止过拟合，每一轮所生成的树都是拟合样本真实值与上一轮的预测值的“残差”，使得预测结果逐渐逼近真实值。XGBoost 是一种非常强大的集成机器学习算法，也经常用来处理股票中的时序数据。XGBoost 核心是通过损失函数展开到二阶导数来进一步逼近真实损失。其理论过程如下：

1) 每轮训练增加一个新的树模型；对于损失函数进行二阶泰勒展开，并定义一棵树及一棵树的复杂度。

2) 每轮训练开始，首先计算梯度统计：

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(s-1)})}{\partial \hat{y}_i^{(s-1)}}, \tag{5-1}$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(s-1)})}{\partial \hat{y}_i^{(s-1)}}. \tag{5-2}$$

3) 根据贪心算法及梯度统计信息生成一棵完整树 $f_s(x)$。节点分裂通过如下公式评估，选择最优切分点：

$$obj_{split} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda}\right) - \gamma, \tag{5-3}$$

得到最终树叶子节点的权重为

$$\omega_j^* = -\frac{G_i}{H_i+\lambda}. \tag{5-4}$$

4）将新生成的树模型 $f_s(x)$ 加入模型中

$$\hat{y}_i^s = \hat{y}_i^{(s-1)} + \eta f_s(x_i). \tag{5-5}$$

XGBoost 在代码中的实现分为四步。首先定义树的棵树，学习率，节点分裂的最小样本数，树的最大深度。然后定义损失函数，将损失函数回归定义为平方损失函数。再初始化分类树，以遍历构造每一棵决策树，拟合每一棵树后将结果累加。最后进行预测，导入数据集，划分数据集，创建 XGBoost 学习器，进行模型的拟合，预测输出评估指标和绘制

比较时序图。

本文在操作时的具体过程如下：

1) 数据在优矿网提取上证综合指数自 2015 年 1 月 1 日日行情数据，包括收盘价，开盘价，最高价，最低价，成交量等基础行情数据；

2) 特征工程：参考选取了三个特征，分别是，最高价和最低价每天的差值（先用一日差值，效果不好会增加多日差值作为特征），开盘价与收盘价的差值，以及成交量；

3) 划分数据集，由于样本量数据较多，我们以 9：1 划分训练集和测试集；

4) 网格搜索调整参数后训练模型并计算训练误差；

5) 进行性能评价：分别在训练集和测试集上计算 RMSE，MAPE；绘制时序图放大预测结果。

绘制预测数据与真实数据对比图 5-3，其中黄色是真实数值，蓝色是预测数值，它能够大概预测出上证综合指数的走势，但是对于极值的预测达不到。这符合金融数据的特点，因为政治社会因素等无法纳入技术因子的特征对于沪深指数的影响是突然而剧烈的。预测后的误差数据在表 5-2 中展示。

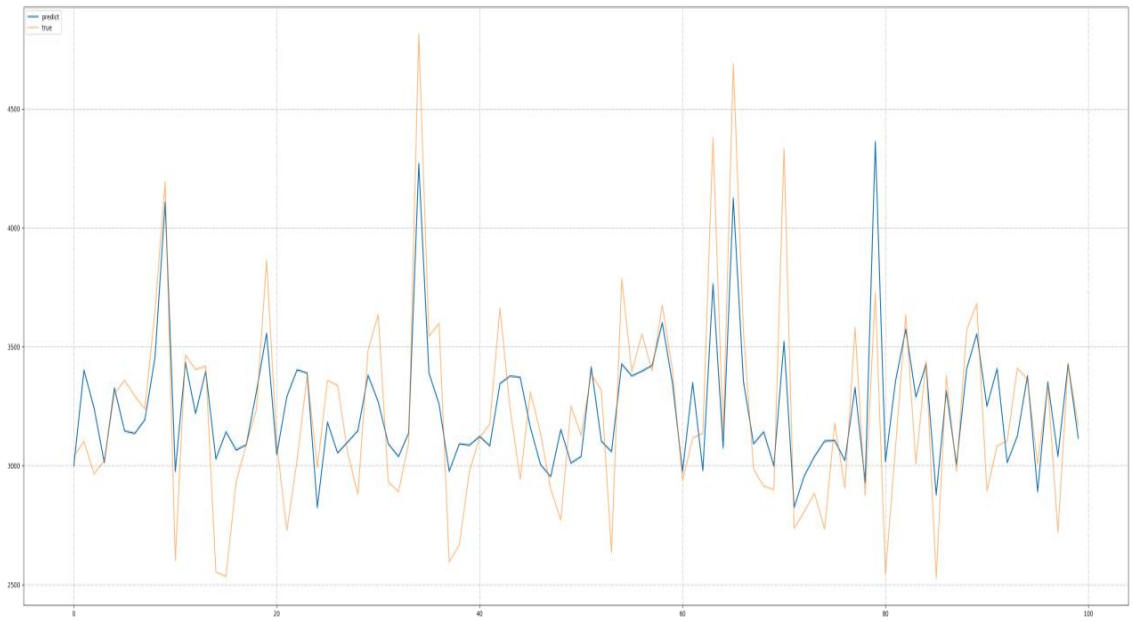


图 5-3　指数预测与真实值对比图

表 5-2　指数预测误差指标

评价指标	数值
训练集上的 RMSE	231.06354603834083
训练集上的 MAE	181.89992515882557
验证集上的 RMSE	260.02755166016857
验证集上的 MAE	206.03334456326849

下面是对代码实现中的模型参数进行设置，参数设置和其意义如下表所示。

表 5-3 XGBoost 参数设置及解释

模型参数	设置解释
max_depth=6，	每一棵树最大深度 6，每棵树的预测结果都要
learning_rate=0.1	乘以这个学习率 0.1
n_estimators=37	用多少棵树来拟合，也可以理解为多少次迭代
booster='gbtree'	gbtree 使用基于树的模型进行提升计算，默认为 gbtree
gamma=0	叶节点上进行进一步分裂所需的最小"损失减少"。默认 0
min_child_weight=5	可以理解为叶子节点最小样本数，设置为 5
subsample=0.6	训练集抽样比例，每次拟合一棵树之前，都会进行该抽样步骤。默认 1，取值范围(0, 1]
colsample_bytree=0.7	每次拟合一棵树之前，决定使用多少个特征，参数默认 1，取值范围(0, 1]
reg_alpha=0.05	控制模型复杂程度的权重值的 L1 正则项参数，参数值越大，模型越不容易过拟合
reg_lambda=0.1	控制模型复杂度的权重值的 L2 正则化项参数，参数越大，模型越不容易过拟合

策略不同，收益和风险也就不同，下面加入择时策略后，来看策略效果是否有明显的提升。其实选股和择时之间本就密切相关，选股可以在股票池中筛选出好的股票标的，择时可以帮助策略人判断是否是交易的好时机，何时是交易的最佳时间。二者结合起来可以筛选更为安全且有增长收益潜力的股票标的和确定交易期，最终都为获取更优的收益服务。

回测选股和前文相同，首先选出十只优质的股票标的，然后利用 XGBoost 模型来对大盘指数做出状态的预测识别，预测大盘次日上涨则买入操作照常，次日下跌则延后操作。回测依旧在聚宽进行，回测区间不变，回测结果如图 5-4 所示。其中蓝色为策略收益曲线，

较单一选股策略有了很大的提升，策略的基准收益由 32.00%提升到了 45.99%。

图 5-4　量化选股＋择时后的回测图

为便于比较，本文将加入择时策略后的回测指标与量化选股策略回测指标在表 5-4 中展示，并计算改进后的指标提升值。

表 5-4　量化选股＋择时后的改进对比

策略名称	CNN—LSTM	择时+CNN-LSTM	改进效果
策略收益	32%	45.99%	提高 13.99%
策略年化收益	32.91%	47.35%	提高 14.44%
超额收益	63.77%	81.12%	提高 17.35%
α	0.444	0.587	提高 0.134
β	0.65	0.646	差异不大
夏普比	0.964	1.475	提高 0.521
最大回撤	25.20%	23.31%	提高 1.89%
策略波动率	0.3	0.294	差异不大

由上面的回测结果图以及该对比表知，选股策略有一定的效果，至少稳定地跑赢了基准收益。但是加入择时之后构成的组合策略效果更好，年化收益提升 14.44%，夏普比率提高，收益波动率和最大回撤值均有所降低，可见量化择时配合选股可以带来更好的收益率。

5.4 本章小结

本章利用 2021 年至 2022 年数据进行交易回测来验证策略的效果。回测结果说明本文所构建的多因子选股能够稳健的获取不错的超额收益，并在各回测指标评价下突出 CNN-LSTM 选股模型表现更为优秀。另外，本章还在选股模型的基础上加入了 XGBoost 大盘的择时，结果表明其各项回测指标都有进一步的改进。

第六章 总结与反思

目前，量化研究在我国学术领域和金融行业头部已经展开多年，也聚集了国内外顶尖的研究人才。一套成熟的量化策略是一步步结合实际市场不断完善起来的，如何研究其内部数据的规律性，去实现其“历史周期”的复刻，是量化研究初期的重点。而近几年机器学习的兴起和发展，让量化研究领域的人员不再拘泥于去研究数据本身，机器学习的暗箱模式让数据研究更加的工具化，实际化。由此，如何克服原始金融数据与理论模型的适配，获得一个在统计理论和金融实践中都理想的量化策略转而成为了现在量化研究者的目标。本文从金融数据的处理，选股模型的选择，策略的回测三个角度对中证养老指数股票池进行了选股和择时策略的结合。意在过程中寻找适合于目标股票池的量化策略，从而得到一个高收益并稳健的多因子选股策略。研究内容如下：

1) 首先，由于只选取了编制进中证养老指数的 80 只成分股数据，能够编制进指数数据中的股票，其相较于沪深 300 的股票更为优质。所以，对于原始数据的处理工作，重点在于因子数据的处理。由于因子众多，在对缺失值，极端值进行处理后，就要对因子做一个筛选。这里是用随机森林模型对因子重要性程度进行计算并作出取舍，选取因子重要性程度>0.005 的因子。然后计算因子之间的相关程度，剔除相关程度高的因子，形成最终因子池。

2) 接下来，本文遵循机器学习做模型预测的基本流程，做数据的处理，模型的训练，模型的评估和最终的预测。并建立了模型预测的评价指标，并从中筛选了最好的单一模型支持向量机，准确率为 0.829。随后本文做了集成模型硬投票的尝试，并发现逻辑分类—随机森林—支持向量机的集成硬投票效果更优，较单一的模型预测，准确率提升到了 0.86793。于是在此选股模型上建立轮动选股策略筛选股票。

3) 然后，本文引入深度学习算法，意在与上述选股策略进行比较来选择最终模型。首先，由于数据结构不匹配深度学习算法，本文采用滑动窗口采样来调整数据结构。然后调整数据参数，通过 CNN，RNN，LSTM，GRU 模型来对 2021 年 12 月份开盘日股票数据进行预测。本文中 2021 年 12 月份及以后数据在所有模型处理中都未被放入模型训练，其结果有参考价值。结果表明，LSTM 算法的 MAE,MAPE 指标都较优。本文又引入 CNN-LSTM 来对单一的算法进行改进，结果表明，其误差数据得到了改进，于是选择

CNN-LSTM 来作为深度学习选股模型。

4) 本文最后在 2021 年 12 月至 2022 年 12 月区间内进行了策略的回测。由于大环境动荡，在此时间区间内大盘走势低迷，呈现负增长，而本文的选股＋择时策略实现了低迷市场下 45.99%的策略收益，验证了策略的有效性。

本文在创作期间需要克服很多问题，也留下很多不足。首先，由于数据的特殊性，必须借助于特定的量化软件来进行回测的操作，而免费的开源软件在数据语言上做了专业化的改进，于是要进行平台的学习。然后，在因子的选取上，没有解决维度的单一问题，没有研究到加入宏观因子，情绪因子后是否会有模型上的改进。可供改进策略设想有扩大特征数目，加入更多行情因子，扩大滞后期因子，参数优化等。

参考文献

[1]苏炜杰.我国实施智能养老战略的现状,经验与措施[D].郑州大学,2023.

[2]B. G. Malkiel, E. F. Fama. Efficient Capital Market, A Review of Theory and Empirical Work[J]. The Journal of Finance,1970,25(2):383-417.

[3]Markowitz Harry. Portfolio Selection[J]. The Journal of Finance, 1952,7(1): 77-91.

[4]William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk[J]. The Journal of Finance, 1964,19(3):425-442.

[5]Ross Stephen A. The arbitrage theory of capital asset pricing[J]. Journal of Economic Theory,1976,13(3):341-360.

[6]E. F. Fama, K. R. French. The Cross-Section of Expected Stock Returns[J]. The Journal of Finance,1992,47(2):427-465.

[7]Merton R. C.An Intertemporal Capital Asset Pricing Model[J]. Econometrica,1973,41(5):867-887.

[8]于志军,杨善林.基于误差校正的GARCH股票价格预测模型[J].中国管理科学,2013,21(1)：341-345.

[9]杨琦,曹显兵.基于 ARMA-GARCH 模型的股票价格分析与预测[J].数学的实践与认识,2016,46(6):80-86.

[10]李想.基于 XGBoost 算法的多因子量化选股方案策划[D].上海师范大学,2017.

[11]H.InceT.B.Trafalis. Short Term Forecasting with Support Vector Machines andnApplication to Stock Price Prediction[J]. International Journal of General Systems,2008.37(6):677-687.

[12]Cao L. Financial forecasting using support vector machines[J]. Neural Computing & Applications, 2001,10(2): 184-192.

[13]杨新斌,黄晓娟.于支持向量机的股票价格预测研究[J].计算机仿真,2010.27(09):302-305.

[14]Huang C F. Hybrid Stock Selection Model Using Genetic Algorithms and Support Vector Regression[J]. Applied Soft Computing Journal, 2012, 12(2):807-818.

[15]吴微,陈维强,刘波.用 BP 神经网络预测股票市场涨跌[J].大连理工大学学报,2001(01):9-15.

[16]Peter G, Zhang. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model.

[17]张萧,韦增.随机森林在股票趋势预测中的应用[J].中国管理信息化,2018,21(03):120-123.

[18]李志杰.基于神经网络的证券市场预测[J].计算机应用,2002(05):31-33.

[19]谢琪,程耕国,徐旭.基于神经网络集成学习股票预测模型的研究[J].计算机工程与应用,2019,55(08):238-243.

[20]曾安,聂文俊.基于深度双向 LSTM 的股票推荐系统[J].计算机科学,2019,46(10):84-89.

[21]黄媛.基于 LSTM 神经网络的多因子选股模型实证研究[D].湘潭大学,2019.

[22]欧阳明哲.基于 GRU 的多因子量化选股策略[D].中南财经政法大学, 2020.

[23]张虎,沈寒蕾,刘晔诚.基于自注意力神经网络的多因子量化选股问题研究[J].数理统计与管理,2020,39(03):556-570.

[24]朱博雅.一种基于数据挖掘的量化投资系统的设计与实现[D].复旦大学,2012.

[25]Ross Stephen A. The arbitrage theory of capital asset pricing[J]. Journal of EconomicTheory,1976,13(3):341-360.

[26]HULTEN C R,HAO X.The Role of Intangible Capital in the Transformation and Growth of the Chinese Economy[R].National Bureau of Economic Research,2012:18405.

[27]RHODES-KROPE.Valuation Waves and Merger Activity:the Empirical Evidence[J].Journal of Financial Economics,2005,77(3):561-603.

[28]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.

[29]吴文茜.中国股票市场低波动率效应实证研究及选股策略应用[J].中国市场,2020.

[30]童克用，姚余栋等.中国养老金融 50 人论坛[R].中国养老金融发展报告.2021:31.

[31]李郅号,杜建强,聂斌等.特征选择方法[J],计算机工程与应用.2019,55(24).

[32]殷洪才,赵春燕.基于神经网络股票预测的研究[J].哈尔滨师范大学自然科学学报.2007(03):47-49.

[33]肖晨,谢真珍,唐宇等.基于卷积神经网络和门控循环单元网络的霾浓度检测[J].中国环境科学学会.2022(08).

[34]J Kim K J. Financial time series forecasting using support vector machines[J].Neurocomputing,2003,55(12):307-319.

[35]Huang W,Nakamori Y,Wang S. Forecasting stock market movement direction with support vector machine[J]. Computers & Operations Research,2005,32(10):13-22.

[36]Deng Y, Bao F, Kong Y, et al. Deep Direct Reinforcement Learning for Financial Signal Representation and Trading[J].IEEE Transactions on Neural Networks and Learning Systems,2016,28(3):1-12.

[37]王春丽,刘光,王齐.多因子量化选股模型与择时策略[J].东北财经大学报,2018(05):81-87.

[38]李琳洁.基于隐马尔可夫模型 HMM 的证券市场择时策略研究[J].时代金融,2018(21):158-159.

[39]王峰虎,尹朝鹏,贺毅岳.一种基于 HMM 预测模型的黄金期货择时策略[J].西安邮电大学学报,2020,25(05)：104-110.

[40]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.