2024 3rd International Conference on
Big Data Economy and Digital Management
第三届大数据经济与数字化管理国际学术会议（BDEDM 2024）

基于面信息、历史走势与市场情绪相结合的股价涨跌关联因子探究

通讯单位:香港大学　作者:文鑫　指导老师:刘凌

2024 年 1 月

摘要

股票市场作为金融市场的重要组成部分，对其价格涨跌的研究是学术界的热点。随着信息化时代的到来，基本面分析、技术分析等传统方法在探究股价涨跌关联因子方面暴露出了一定的局限性。因此，本文结合股票基本面、历史走势与市场情绪数据，研究对于股价涨跌的影响，以挖掘股价涨跌的关联因子。本研究运用统计词频和深度学习文本情感分类的方法，加入市场情绪因子，并提出有监督二分类机器学习模型算法。以股价涨跌为因变量，并运用 **Python** 对模型进行建模。结果显示：历史股价最高价、股票交易额、股票所在行业因子、股价波动率与股价涨跌有着强的联系。根据研究结论，本文对投资者和企业经营者提出以下建议：对于股票投资者而言,在分析股票市场行情和做出投资决策时，可以重点关注上述指标变化情况；对于股票公司经营者而言，在日常企业经营过程中，需要监控上述指标变化情况并分析深层原因，以保持企业健康持续运营。

关键词：股价涨跌；关联因子；市场情绪；二分类模型

Abstract

As an important part of the financial market, the stock market is a hot spot in academia for the study of its price rise and fall. With the advent of the information age, traditional methods such as fundamental analysis and technical analysis have exposed certain limitations in exploring the correlation factors of stock prices. Therefore, this paper combines stock fundamentals, historical trends and market sentiment data to study the impact on stock price fluctuations to explore the correlation factors of stock price fluctuations. In this study, statistical word frequency and deep learning text sentiment classification methods are used to add market sentiment factors, and a supervised dichotomous machine learning model algorithm is proposed. Take the rise and fall of stock prices as the dependent variable, and use Python to model it. The results show that there is a strong correlation between the historical stock price high, the stock trading volume, the industry factor of the stock, and the stock price volatility and the rise and fall of the stock price. According to the research conclusion, this paper puts forward the following suggestions for investors and business operators: for stock investors, when analyzing the stock market and making investment decisions, they can focus on the changes of the above indicators; For stock company operators, in the daily business operation process, it is necessary to monitor the changes of the above indicators and analyze the underlying reasons to maintain the healthy and continuous operation of the enterprise.

Keywords: stock price rise and fall; correlation factors; market sentiment; dichotomous models

目录

1.绪论

1.1 研究背景

近年来，我国数字经济正以快速的态势持续发展着，许多科研机构也在持续推进数字产业化的进程。借助数字技术，国内各行业全面改革不断进行。在大数据，人工智能和云计算落地大环境中，传统产业科技化和数字化程度显着提升。其中金融产业的科技化和数字化越来越发达，一系列前沿科学技术被创新应用到金融领域中已经成为业界普遍关注的话题[1]。

数字经济与人工智能背景下，自然语言处理技术 NLP 受到学术界与工业界的普遍重视，文本数据情感分析量化是自然语言处理技术 NLP 的一个主要研究领域，有着非凡的研究价值与应用价值。金融领域债券，股票和基金的中文文本数据日益增多，其生成速度很快，所含信息数量级也较大，如何能够快速准确的挖掘出金融文本数据深层次的信息，这是当前学术界，工业界亟待解决的课题之一[2]。

寻找股票收益率的有效关联因子，并将其用于股价涨跌的分析与预测中，是资产定价和量化投资领域经久不衰的话题。传统的股票投资研究多基于经典的经济金融框架，包括马科维茨的投资组合理论、CAPM 模型、套利定价理论、Fama-French 因子模型等,国内外学者对此开展了大量且深入的研究,从中发掘出了市值、账面市值比、盈利能力等解释性强的因子。然而，随着经济高速发展和信息化时代的到来，基本面分析、技术分析等传统方法在探究股票收益率影响因子方面似乎暴露出了一定的局限性。例如，几乎只能从广义线性的角度来挖掘因子，并且初始因子的选取由于主要依赖于人工搜集而维度较少。在此背景下，引入机器学习和深度学习方法、借助海量数据资源和新兴的模型算法来进一步探究股价关联因子，逐渐成为了热点议题[3]。

众所周知，股票市场上的投资者会聚在微博，社区，论坛，贴吧这些网络空间里表达他们对于股票市场的观点[4]。这些文字信息不仅能够表达投资者目前的心情，它能否探讨股票市场量化指标数据如股价涨跌等也是一个值得关注的问题。财经网站

上的股民评论往往被看作是含有大量市场情绪信息的，这是因为这些情绪信息将成为影响股民投资决策的主要因素，同时对于股票价格涨跌都有影响。

从这一视角出发，研究者可以利用自然语言处理、文本分析、情感量化技术来衡量市场参与各方的情绪或态度、从而构建新的因子，可以建立高阶深入的分类或预测模型来提高因子评价的准确度和效率，还可以通过高维数据训练和模拟来提升方案的稳健性和适用性。

1.2 研究内容

本文以研究背景和研究意义为引导，从梳理国内外相关研究文献开始，对基于面信息、历史走势与市场情绪相结合的股价涨跌关联因子进行研究。建立文本情感分类与有监督机器学习分类模型来探讨股票面信息、历史走势与市场情绪对股价涨跌的影响，通过 Python3.0 软件来验证包括股票公司注册信息、业绩报告、股价开盘价、股价收盘价和市场情绪 2 因子对股价涨跌的作用。本文研究数据来源于 Python 中 Tushare 接口提供的数值型数据和网络爬虫的东方财富网股吧评论文本型数据，采用客观数据验证研究模型，为股价涨跌关联因子相关研究提供新思路。最后对研究结果进行总结，并反思研究的不足之处与改进方向。

1.3 研究意义

1.3.1 理论意义

本研究丰富了股价涨跌关联因子相关研究。本研究以东方财富网股吧评论作为股票市场情绪的来源，将股票市场情绪即股民评论文本型数据引入股价涨跌关联因子探究中。在原有股票公司基本面财务数据、历史走势股价数据的基础上，加入市场情绪数据，研究数据更加充分完整，研究范围更加广泛。因为这些文字信息通过反应当下投资者的情绪，对股票市场的量化指标数据中股价涨跌进行探索。财经网站股民评论常被认为蕴含着丰富的市场情绪信息，其中所包含的情感信息对股票价格涨跌具有一定的影响。股价涨跌关联因子引起了众多学者的关注，但多数文献仅仅采用股票公司截面财务数据和股价历史走势数据，少有文献加入市场情绪因子数据。

东方财富网在中国的访问量最大、影响力最大的财经证券门户网站。东方财富网

以树立专业，权威为己任、为了用户，财经媒体应运而生[5]。东方财富网一直秉承网站内容权威、专业，多方位地涉及各个财经领域，每天都有数以万计的数据和信息在不断更新。东方财富网的股吧评论在 2006 年正式推出，用户活跃量稳步提升，累积起来的广大投资者群体。东方财富网股吧评论，对于投资者群体的投资决策构成重要影响，故而从东方财富网平台获得股票公司股民评论进行相关研究具有一定的现实意义。因此，本文以东方财富网的股吧评论为依据，为研究股价涨跌关联因子提供了一种新思路，同时，它还是对已有相关研究的一种补充与扩展。

1.3.2 实践意义

金融科技的大潮之下，引入人工智能方法、借助于大量数据资源，结合新兴模型算法，对股价关联因子进行了进一步探索，渐成热点议题。在这样的背景下，通过机器学习和深度学习方法对金融场景中数据进行挖掘，具有充分的创新和实践意义。在本研究中，研究问题是有标签二分类问题，因此通过引入机器学习中有监督分类的逻辑回归和树型分类方法，可以有效解决该研究问题。在模型算法效果良好的情况下，通过输出机器学习分类算法特征重要性，历史内股价最高价、股票交易额、股票所在行业因子的重要性高，即与股价涨跌的关联性高。

综上所述，一方面，对于投资者来说，投资者在分析股票市场行情和做出投资决策时，可以重点关注这些指标变化情况；另一方面，对于股票公司经营者而言，在日常企业经营过程中，需要监控这些指标变化情况并分析深层原因，这对于企业健康持续运营有着关键的作用。

1.4 研究方法

1.4.1 文献研究法

本研究从行为金融学、心理学、人工智能等多学科理论入手，参阅了国内外研究主题相关的文献，通过对已有研究成果的总结归纳，从而明确该研究方向的现有成果和最新进展，为今后的研究寻找新的切入点。本文主要通过运用“关键词法”和“滚雪球法”的方法，对有关文献进行检索与梳理，涉及到的关键词包括“金融科技”、“文本分析”、“情感分类”、“股价涨跌”等。在文献研究过程中，基于已有文献定性

地分析市场情绪、股价涨跌等概念，为模型构建提供理论基础。基于文献回顾与梳理的结果,本文明确了东方财富网股吧评论中股价涨跌的市场情绪因子因素及其测量维度，从而为之后模型的构建奠定了坚实的理论基础。

1.4.2 模型分析法

本文通过 Python 的 Tushare 接口直接获取发行股票公司的基本面信息(财务数据)和历史走势(股价数据)数据，通过 Python 程序爬取从东方财富发行股票公司的股吧评论中获得文本型数据市场情绪(股吧评论)数据作为原始数据。采用描述性统计的方法对各个变量的分布情况进行分析，并且构建文本情感分类和机器学习分类模型，运用 Python 的 Jupyter 软件来检验沪深股票公司面数据（财务数据）、历史走势（股价数据）和市场情绪(股民评论)对股票股价涨跌的影响。

1.5 技术路线

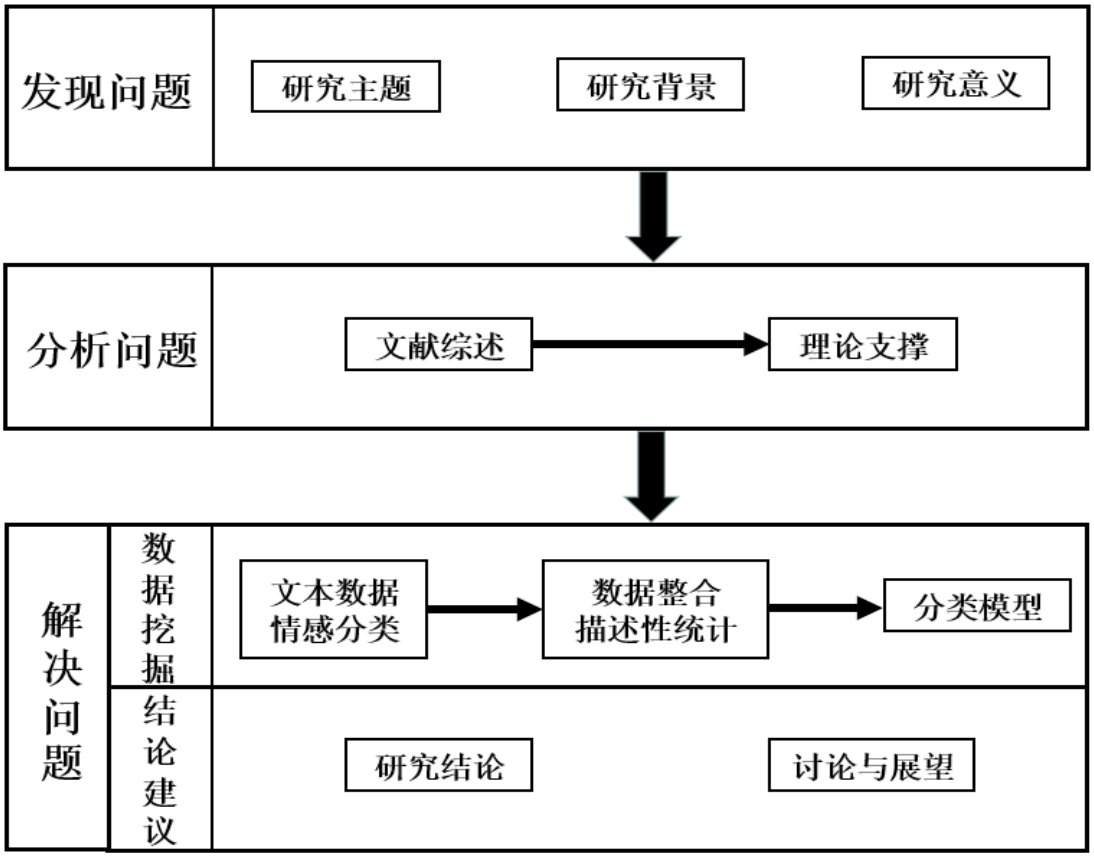

图 1-1 研究技术路线图

2. 文献综述与理论支撑

2.1 股价涨跌相关研究

在科技和金融不断融合的今天，在这一课题之下，研究亦越来越热，逐步成为工业界，学术界普遍探讨的焦点话题，很多学者关注金融科技在股票和债券，基金及其他方向运用[6]。股票市场行情已经被认为是反映整个社会经济状况的晴雨表[7]，对股票所产生的利益与风险进行合理控制，已成为许多机关部门的工作，和投资者为之奋斗的方向。

在与金融科技有关研究方面，股价预测是日益引起社会公众重视的发展方向。用数学方法在股票历史交易数据上寻找规律，是股票价格预测的首要手段。其中最常见的方法就是基于历史数据建立数学模型进行分析与判断，然后根据模型结果来预测出未来一段时间内股票价格趋势。由于股票历史交易数据是结构化数值型数据，便于统计与计算。它被广泛应用于股票预测。所以量价预测又是最普遍的量化方法之一[8]。而且股价预测通常是建立在有效市场假说的基础之上，也就是参与到市场中去，投资者是充分理性的，市场信息是透明的。在这一严格而又严苛的假定下，常常研究所得的股价预测结论并不符合事实。本文将采用一种新的视角来考察股价波动中的各种不确定性因素对价格走势的影响作用。股票价格的上涨或下跌，并非单纯地受到当前环境与过去股价之间的关系所左右，受公司财报，国家政策的影响较大、专家推测时事，以及其他诸多复杂因素作用。

为股票收益率找到有效关联因子，用它来分析股价涨跌，在量化投资领域是一个历久弥新的发展方向，传统股票投资研究大多是建立在经典经济金融框架之上。股票价格与公司信息密切相关，因此可以通过股票历史交易数据来进行股票价格的分析。国内外学者对这一问题已经进行了许多和深入的探讨，从中挖掘出市值、账面市值比，盈利能力和其他解释性更高的因素。包括 Markowitz（1952）的投资组合理论中以最小化标准差并最大化期望收益为目标来进行资产配置，利用不同证券收益的相关性

来分散风险[9];William Sharpe（1964）等人的CAPM资产定价模型研究证券市场中资产的预期收益率与风险资产之间的关系，以及均衡价格确定方式、Fama 和French(1992)年对美国股票市场决定不同股票回报率差异的因素的研究发现,股票的市场的beta值不能解释不同股票回报率的差异，而上市公司的市值、账面市值比、市盈率可以解释股票回报率的差异。

然而，随着经济高速发展和信息化时代的到来，基本面分析、技术分析等传统方法在探究股票收益率影响因子方面似乎暴露出了一定的局限性。例如，几乎只能从广义线性的角度来挖掘因子，并且初始因子的选取由于主要依赖于人工搜集而维度较少。

基于这种情况，本文引入了机器学习与深度学习的方法、借助于大量数据资源，结合新兴模型算法，对股价关联因子进行了进一步探索，渐成热点议题。随着金融市场的发展以及投资者对市场信息需求的增加，越来越多的学者开始关注股市与基本面之间的联系。刘新月（2022）提出了创业板股价短期上涨或下跌走势预测难题，提出一种运用财务指标分析股价涨跌的预测模型。通过选取不同时间段的上证指数收益率与同期股票开盘价作为输入变量构建训练集和测试集，采用支持向量机模型对其进行训练和测试。支持向量机模型的精度达到了0.812，说明机器学习分类模型对于股票涨跌趋势具有较好的适用性[10]。崔炎炎（2022）使用情感分类模型，在爬取到的 11 万多条微博文本中挖掘出了投资者情绪，最后，将长短时记忆神经网络模型（Long-Short-Term-Memory）组合起来，研究认为，网络舆情在金融科技个股收盘价的预测中起到了至关重要的作用[11]。徐攀（2021）利用公众社交平台“东方财富网”和百度指数，对投资者情绪与投资者关注度进行分析与测算，并且在股价涨跌预测模型的基础上，增加了上述的两个变量，所得预测准确率较单纯使用财务指标的预测模型提高了5.65%[12]。刘素辉(2021)基于深度学习卷积神经网络(Convolutional Neural Network）模型，考虑金融学中具有可解释性的有效市场假设、行为金融学理论等、适应性市场假说等。构建了基于新闻情感倾向性的股票走势分类算法，并通过实验对比分析不同特征在分类效果上存在差异。研究认为，新闻标题文本数据较公众情绪文本数据更能显着影响股价走势[13]。

由此可见，利用机器学习或是深度学习作为框架，将网络舆情、新闻标题、股吧评论等文本型数据加入到模型当中，得到对股价涨跌趋势的预测效果提升显著。在总

结上文研究方法后，本研究主题将结合发行股票公司的财务数据、股价走势数据与公众情绪文本数据，采用机器学习经典分类算法对股价涨跌趋势进行探究，从中挖掘出有效的股价关联因子。

2.2 文本情感分类相关研究

财经新闻与股票股吧评论往往被视为包含了大量信息,其蕴含的情感信息是投资者投资决策中重要因素之一，对于股票收益率亦有一定作用。其中，基于机器学习的股票涨跌预测模型是一个研究重点。情感分类模型也伴随着科技的发展，经历了由规则匹配，机器学习，深度学习三个发展阶段。

顾文涛（2020）从基于文本的大数据方法，构建了适用于金融投资领域的财经新闻情感词典来对财经新闻进行文本分析，对财经新闻文本中的情绪进行量化，得到情绪指数，并以此为基础，以模型的得分作为权重，将多种预测模型结合起来，构造了一个得分加权模型，对股票收益率进行了预测[14]。张姝（2022）是在人工构建的文本情感词库之下，利用朴素贝叶斯分类算法，由于朴素贝叶斯分类算法在数学逻辑上推导严密，经过不断地调整和优化，获得了电商评论之下良好的情感分类结果[15]。Jacob(2019)在以注意力机制下的 Transformer[24]下构建自然语言处理预训练模型 Bert,该模型在多个斯坦福公布的文本情感分类数据集上均表现优越[25]。王利利(2019）通过大量搜集学生发出的言论，利用批量梯度下降算法、随机梯度下降算法等、小批量梯度下降算法 3 种梯度下降算法将收集的学生所发言论分类文本情感，最后，以上 3 种深度学习分类算法的准确率达到预期效果[16]。从以上对文本情感分类的研究成果来看，已有以情感词典为基础、基于机器学习和深度学习的情感分类算法得到了广泛应用，并在财经新闻和金融领域其他情景情感量化方面有不俗表现。因此，本研究引入的东方财富网 2799 只沪深股票的股吧评论可以采用上述方法进行情感量化得到市场情感因子。

2.3 技术理论支撑

本论文研究主题为沪深 2799 支股票的基本面数据、历史走势与市场情绪相结合的股价涨跌关联因子探究。因此本文的算法模型框架为有监督二分类问题，根据数据

量级与目标任务，选择逻辑回归作为基础模型，机器学习中树型算法包括决策树、随机森林、XGBoost[26]、LightGBM 算法模型，最后使用 Stacking 将模型加权融合得到最优模型。

一种重要的适用于 0-1 型因变量的回归分析模型：逻辑回归。从理论上讲，逻辑回归是广义线性回归模型的一个特例。这个模型的本质就是某种线性模型,具有线性模型共有的优良性质[17]。非常简单，而且参数个数少，因此能够支持相对比较小的样本估计。同时，因为模型结构简单，参数估计结果也很好解读。但是逻辑回归不灵活，因为这是一个线性模型，因此它的函数形式单一，难以描述更加灵活的函数形态。

以决策树模型为基础的机器学习树型结构模型广泛适用于有监督的二分类问题中。树型结构模型的原理是：构造根节点，将所有训练数据集都放到根节点，选择一个最优特征，将训练数据集分割成两个或多个子集，使得训练集在当前条件下有最好的分类[18]。如果这些子集已经可以很好的分类，那么构建叶子节点，如果还不能很好的分类，继续选择最优特征并对其分割，构造相应的节点，如此递归进行，直至所有训练数据集的分类准确率达到预期，或者没有合适的特征进行分裂为止。此类模型具有适用于分类模型，且算法逐渐优化拓展，便于进行比较探究；多支树模型结构，能降低过拟合的风险，非常稳定；易用便捷，可解释性强，在同一主题其他场景下已取得良好结果[19]。基于上述机器学习树型结构模型使用场景和优势，与本研究内容良好的匹配，因此选用上述决策树、随机森林、XGBoost、LightGBM 算法模型作为研究的主要模型。

Stacking 模型融合是将多个分类模型进行融合，通过组装新的模型来学习每个分类器的特征权重。Stacking 模型融合具有使得模型更加稳定，具有更好的泛化性能,从而确保在大多数情况下测试用例具有良好的性能[20]。融合多个不同的模型结果，来获得更好的泛化性能。目前，具有多层次架构的深度学习模型，相较于只有浅层结构的模型来说，具有更好的性能表现。因此采用 Stacking 模型融合方法可以使得逻辑回归模型和其他树型结构模型更加全面，保证在截面数据，历史走势数据，以及文本型数据中都有着良好的分类性能。

3. 数据收集与描述

3.1 数据收集

本论文数据来源主要分为2799家沪深股票的基本面信息（财务数据）、历史走势（股价数据）和市场情绪（股吧评论）三个层面，数据涵盖000001号平安银行股票到688981号中芯国际股票2799家沪深股票。

具体而言，基本面信息包括发行股票公司的代码信息、注册信息、业绩报告、盈利能力、营运能力等财务数据;历史走势(股价数据)包括发行股票公司的代码信息、开收盘价、交易量、交易额、波动率等价格相关指标在一定时间间隔3天、7天、10天、15天、21天、30天、60天的变化率;市场情绪（股吧评论）则是通过Python程序爬取2799家沪深股票公司在东方财富网股吧中股民评论后进行情感分类量化得到市场情绪因子。

本论文数据从数据形式类型上主要分为数值型数据和文本型数据,其中数值型数据包括发行股票公司的基本面信息(财务数据)和历史走势(股价数据)数据，来源是通过Python的Tushare接口直接获取。Tushare是一个金融大数据平台，数据内容包含股票、指数、基金、期货、债券、外汇、行业大数据，同时包括了数字货币行情等区块链数据，内容完善准确;文本型数据市场情绪(股吧评论)则是通过Python程序爬取东方财富发行股票公司的股吧评论中获得。股吧是一个股票交流平台，以股民为主要参与群体，分享投资经验，表达思想，的网络空间。东方财富是中国访问量最大，影响力最大的财经证券门户网站之一。

3.1.1 基本面信息（财务数据）

发行股票公司的基本面信息（财务数据）一方面全面系统地揭示企业一定时期的财务状况、经营果和现金流量，有利于经营管理人员者了解公司各项任务指标的完成情况。另一个方面对于投资者而言，财务数据是掌握企业的财务状况、经营成果和现

金流量情况，进而分析企业的盈利能力、偿债能力、投资收益、发展前景的重要来源，为他们投资、贷款和贸易提供决策依据。

基本面信息（财务数据）主要是指发行该沪深股票的公司财务数据，由于公司财务数据一般是按季度每三个月进行整理，因此从 Tushare 接口获取到相应沪深股票公司最新季度在 2022 年第 3 季度发的财务信息。具体维度和因子如下表 3-1 基本面信息因子所示。主要分为:注册信息、业绩报告、盈利能力、营运能力、成长能力、偿债能力、现金流量七个方面共 40 个特征。

表3-1 基本面信息因子

维度	变量名	变量含义（因子意义）
注册信息	AREA	公司所在地区
	INDUSTRY	公司所在行业
	MARKET	上市版块
	LIST_DATE	上市日期
业绩报告	EPS	每股收益
	EPS_YOY	每股收益同比（%）
	BVPS	每股净资产
	EPCF	每股现金流量(元)
	NET_PROFITS	净利润(万元)
	PROFITS_YOY	净利润同比(%)
	DISTRIB	分配方案
	REPORT_DATE	发布日期
盈利能力	ROE	净资产收益率(%)
	NET_PROFIT_RATIO	净利率(%)
	GROSS_PROFIT_RATE	毛利率(%)
	NET_PROFITS	净利润(万元)
	BUSINESS_INCOME	营业收入(百万元)
	BIPS	每股主营业务收入(元)

表3-1 基本面信息因子（补充）

营运能力	ARTURNOVER	应收账款周转率(次)
	ARTURNDAYS	应收账款周转天数(天)
	INVENTORY_TURNOVER	存货周转率(次)
	INVENTORY_DAYS	存货周转天数(天)
	CURRENTASSET_TURNOVER	流动资产周转率(次)
	CURRENTASSET_DAYS	流动资产周转天数(天)
成长能力	MBRG	主营业务收入增长率(%)
	NPRG	净利润增长率(%)
	NAV	净资产增长率
	TARG	总资产增长率
	EPSG	每股收益增长率
	SEG	股东权益增长率
偿债能力	CURRENTRATIO	流动比率
	QUICKRATIO	速动比率
	CASHRATIO	现金比率
	ICRATIO	利息支付倍数
	SHEQRATIO	股东权益比率
	ADRATIO	股东权益增长率
现金流量	CF_SALES	经营现金净流量对销售收入比率
	RATEOFRETURN	资产的经营现金流量回报率
	CF_NM	经营现金净流量与净利润的比率
	CF_LIABILITIES	经营现金净流量对负债比率
	CASHFLOWRATIO	现金流量比率

在通过Tushare接口获取发行股票公司七个基本面数据后，将每个维度按照唯一标识CODE（股票代码）进行拼接操作，并去除了部分重复特征，ROE（净资产收益率）出现在了业绩报告和盈利能力中，只保留其一。整理后的基本面信息数据如下图3-1基本面信息因子汇总所示，展示部分因子。

	code	name	area	industry	market	list_date	eps	eps_yoy	bvps	roe	epcf	net_profits	profits_yoy	distrib	report_date	net_profit_ratio
0	000001	平安银行	深圳	银行	主板	19910403	1.78	27.14	18.32	10.15	4.77	3665900.00	25.82	NaN	10-25	26.51
1	000002	万科A	深圳	全国地产	主板	19910129	1.47	2.35	NaN	7.12	NaN	1705042.18	2.17	NaN	10-29	NaN
2	000004	ST国华	深圳	软件服务	主板	19910114	-0.38	-281.88	NaN	-6.17	NaN	-5676.56	-268.26	NaN	10-28	-67.39
3	000005	ST星源	深圳	环境保护	主板	19901210	0.01	-94.82	NaN	0.73	NaN	1007.11	-94.82	NaN	10-31	5.77
4	000006	深振业A	深圳	区域地产	主板	19920427	0.15	-64.21	NaN	2.61	NaN	20108.34	-64.22	NaN	10-28	13.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2866	688789	宏华数科	浙江	专用机械	科创板	20210708	2.45	-7.89	NaN	11.86	NaN	18641.96	14.32	NaN	10-29	26.73
2867	688798	艾为电子	上海	半导体	科创板	20210816	0.33	-78.29	NaN	1.45	NaN	5449.46	-72.14	NaN	10-31	3.26
2868	688800	瑞可达	江苏	元器件	科创板	20210722	1.82	121.95	NaN	18.24	NaN	19709.18	175.66	NaN	10-24	16.92
2869	688819	天能股份	浙江	电气设备	科创板	20210118	1.47	32.43	NaN	11.14	NaN	142489.87	33.73	NaN	10-29	4.74
2870	688981	中芯国际	上海	半导体	科创板	20200716	1.19	27.96	NaN	7.80	NaN	938950.70	28.30	NaN	11-11	24.86

图3-1 基本面信息因子汇总

3.1.2 历史走势（股价数据）

参考发行股票公司股价历史数据，例如每日开盘价、收盘价和波动率可以总体评估市场对公司的基本印象，可以作为评价现在估值情况的参考依据。尤其是对运行比较稳定的，主营业务无明显变化，行业竞争格局未发生明显改变的企业等，历史数据更显得尤为重要。由于历史股价数据更好地告诉投资者，目前，公司股票价格总体上处于高估或低估的状态，利于研判大趋势，判断总体行情是否过热，或被严重低估[21]。

选择因子的具体方法为：以 2022 年 12 月 30 日为所研究的历史最后一天，获取该天若干天前（3 天前、7 天前、10 天前、15 天前、21 天前、30 天前、60 天前）当天的股价相关指标（开盘价、收盘价、最高价、最低价、交易额、交易量），并计算相应时间间隔内上述指标的变化率。其中还按照下列公式计算新增了股价波动率因子。

$$波动率 = \frac{最高价 - 最低价}{2 \times （开盘价 + 收盘价）}$$

整理出的部分股价因子展示如下图 3-2 股价数据因子汇总

	code	open_7	open_10	open_15	open_21	open_30	open_60	high_7	high_10	high_15	high_21	high_30	high_60
0	000925	-0.023392	0.002924	0.016082	0.064327	0.083333	0.010234	-0.021771	0.008708	0.055152	0.065312	0.079826	0.076923
1	000926	0.006993	0.055944	0.060606	0.097902	0.235431	-0.114219	0.013793	0.043678	0.059770	0.110345	0.218391	-0.126437
2	000927	0.000000	0.024561	0.031579	0.098246	0.042105	-0.094737	-0.003472	0.024306	0.055556	0.090278	0.069444	-0.104167
3	000928	-0.075908	0.021452	0.019802	0.041254	0.049505	-0.143564	0.037398	0.016260	0.034146	0.037398	0.034146	-0.138211
4	000929	-0.000813	-0.129268	-0.130081	-0.173984	-0.271545	-0.332520	-0.027338	-0.224460	-0.205036	-0.257554	-0.335252	-0.398561
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4941	003025	-0.043703	-0.061090	-0.059680	-0.032425	-0.097274	-0.051692	0.005169	0.004699	0.002820	-0.016917	-0.067199	0.000000
4942	603316	-0.012195	0.015679	0.052265	0.054007	0.074913	-0.067944	-0.059016	-0.029508	0.004918	-0.003279	0.013115	-0.106557
4943	002965	-0.085850	-0.079063	-0.196980	-0.187648	-0.161860	-0.117747	-0.121054	-0.073056	-0.139766	-0.214123	-0.177351	-0.111617
4944	300595	-0.034853	-0.046768	0.003277	-0.014596	-0.055704	-0.080429	-0.038355	-0.111479	-0.059603	-0.080574	-0.112859	-0.140728
4945	300610	-0.038014	-0.001552	0.031808	0.121024	0.057409	-0.006982	-0.036980	-0.003082	0.039291	0.114022	0.052388	0.026194

图 3-2 股价数据因子汇总

基于历史走势(股价数据)计算有监督学习二分类模型的标签，还获取了 3 天后即 2022 年 12 月 30 日的股价数据，并计算了这未来三天间的股票收益率，并据此给股票打上标签，收益率为正记为 1，否则记为 0。

3.1.3 市场情绪（股吧评论）

情感分析就是发掘文本信息中的情感倾向，主要应用于舆情监测、商品评论的分析和信息检索。在社交媒体迅猛发展的今天，文本数据量呈爆炸性上升趋势，文本情感分析已经成为自然语言处理领域中最主要的热点问题之一。

由于市场信息不对称的存在，投资者很难有效评判上市公司真实价值，转向企业最近发生的事情、评论观点和其他资料，经过个体分析，情绪倾向得以形成，继而影响投资决策。比如在市场行情节节攀升时，社会关注度增加、投资者的参与积极性得到了调动，在意见交换与传染之后，这类情绪会产生“盲目乐观”倾向[22]，继而吸引了新型非理性投资者，推动股市较快增长。

使用 Python 程序在东方财富网爬取了沪深 2799 支股票对应股吧在 2022 年 12 月份发布的所有评论，条数分别从几十余条到 40000 余条不等，然后依次对每只股票的评论序列进行情感分析，具体算法如下，处理示例可见下。

Step 1: 定义网页爬虫函数，爬取评论（正则匹配仅在 12 月发布的所有评论）

Step 2: 对于特定股票的评论序列，将每条评论使用 Python 的 Jieba 分词器分词

Step 3: 利用正负情感词汇库，统计每条评论的情感词频，并打上情感标签（若正向词频多于负向则为 1，否则为-1）

Step 4: 分别用预训练好的微调后的 Bert-Chinese 模型和多项式朴素贝叶斯模型（Naive Bayes model）来预测情感标签

Step 5: 将三种方法的情感标签值相加，通过投票思想确定最终的情感标签（三者之和大于 0 则为 1，否则为-1）

Step 6: 切换下一只股票，重复 *Step 2* ~ *Step 5*，直至完成对所有股票的分析。

	comment	cutted_comment	positive_words_count	negative_words_count	sentiment_label	sentiment_sgd_clf	sentiment_MultinomialNB_clf
0	长安汽车(SZ000625)我又回来啦哈哈哈	长安汽车(SZ000625)我 又 回来 啦 哈哈哈	0	0	-1	-1	-1
1	长安汽车(SZ000625)新年愿望是你倒闭	长安汽车(SZ000625)新年 愿望 是 你 倒闭	0	1	-1	-1	-1
2	长安汽车(SZ000625)家人们，醋哥提钱祝大家新年快乐，财源广进，过完节记得减仓[梭...	长安汽车(SZ000625)家人 们 ， 醋哥 提钱 祝 大家 新年快乐 ， ...	0	1	-1	-1	-1
3	长安汽车(SZ000625)大盘要到2300。现在说啥也得跑了。	长安汽车(SZ000625)大盘 要 到 2300 。 现在 说 啥 也 得...	0	1	-1	-1	-1
4	长安汽车(SZ000625)年底总结，今年都是亏在，望明年能回本	长安汽车(SZ000625)年底 总结 ， 今年 都 是 亏 在 ， 望 明...	0	1	-1	-1	-1
...	...	...	...	...	...	...	...
16215	曾经说过雨虹今天说说长安	曾经 说 过雨虹 今天 说 说 长安	0	0	-1	-1	-1
16216	收盘没有大问	收盘 没有 大问	0	0	-1	-1	-1
16217	长安汽车(SZ000625)明天抢筹	长安汽车(SZ000625)明天 抢筹	0	0	-1	-1	-1
16218	对沪指形成的	对 沪指 形成 的	0	0	-1	-1	-1
16219	长安汽车(SZ000625)今天估计买不到	长安汽车(SZ000625)今天 估计 买不到	0	0	-1	-1	-1

图 3-3 情感处理示例

在完成对每只股票的每条评论添加情感标签后，构建市场情绪因子。包括市场指数 SEN_INDEX：2022 年 12 月份关于该股票发布的总评论数，反映了投资者的关注程度;情绪总分 SEN_SCORE：该股票 2022 年 12 月份所有评论的情感标签的加总，反映了投资者情绪的总体水平;情绪均值 SEN_MEAN:该股票 2022 年 12 月份所有评论的情感标签的加总与评论数之商，反映了投资者情绪的平均程度;情绪标准差 SEN_STD:该股票 2022 年 12 月份所有评论的情感标签的标准差，反映了投资者情绪的波动性;从市场情绪总量，平均值，差异度三方面构建市场情绪因子，可以将股票市场情绪完整充分的通过数值表达出来。处理结果如上图 3-3 情感处理示例，处理得到的市场情绪指标如下图 3-4 所示

	code	name	sen_index	sen_score	sen_mean	sen_std
0	1	平安银行	1793	-167	-0.093140	0.995653
1	2	万科A	3098	-942	-0.304067	0.952651
2	4	ST国华	416	-208	-0.500000	0.866025
3	5	ST星源	535	-247	-0.461682	0.887045
4	6	深振业A	10407	-6087	-0.584895	0.811109
...	...	...	...	...	...	...
2794	688789	宏华数科	49	31	0.632653	0.774435
2795	688798	艾为电子	134	2	0.014925	0.999889
2796	688800	瑞可达	107	59	0.551402	0.834240
2797	688819	天能股份	419	-149	-0.355609	0.934635
2798	688981	中芯国际	2146	-650	-0.302889	0.953026

2799 rows × 6 columns

图 3-4 情感处理接结果

3.2 数据描述性分析

3.2.1 数据整合

首先基本面信息（财务数据）、历史走势（股价数据）、市场情绪（股吧评论）三个方面的因子数据按照共有的变量 CODE（股票代码）进行拼接到一张总数据表图。如图 3-5 总数据表图所示

	code	eps	eps_yoy	roe	net_profits	profits_yoy	net_profit_ratio	gross_profit_rate	business_income	bips	...	sen_score
0	1	1.780273	27.140625	10.148438	3.665900e+06	25.812500	26.515625	72.625000	138265.000000	7.125000	...	-167
1	2	1.469727	2.349609	7.121094	1.705042e+06	2.169922	-4.980469	29.796875	9000.432617	6.492188	...	-942
2	4	-0.379883	-282.000000	-6.171875	-5.676560e+03	-268.250000	-67.375000	48.093750	84.224998	0.633789	...	-208
3	5	0.010002	-94.812500	0.729980	1.007110e+03	-94.812500	5.769531	13.148438	174.269699	0.164551	...	-247
4	6	0.150024	-64.187500	2.609375	2.010834e+04	-64.250000	13.500000	40.031250	1488.893066	1.102539	...	-6087
...	...	...	...	...	...	...	...	...	...	...	...	...
2794	688789	2.449219	-7.890625	11.859375	1.864196e+04	14.320312	26.734375	46.187500	697.163086	9.171875	...	31
2795	688798	0.330078	-78.312500	1.450195	5.449460e+03	-72.125000	3.259766	41.875000	1670.090332	10.062500	...	2
2796	688800	1.820312	121.937500	18.234375	1.970918e+04	175.625000	16.921875	27.125000	1164.326782	10.289062	...	59
2797	688819	1.469727	32.437500	11.140625	1.424899e+05	33.718750	4.738281	17.500000	30041.896484	30.906250	...	-149
2798	688981	1.190430	27.953125	7.800781	9.389507e+05	28.296875	24.859375	39.906250	37763.558594	4.765625	...	-650

2799 rows × 86 columns

图 3-5 总数据表图

而后进行数据清洗工作，分别对数据进行数据重复值，缺失值进行检验和处理。变量的缺失值统计情况如图所示。关于缺失值处理，对于缺失比例极高即超过 75%的变量，包括EPCF、DISTRIB、BVPS，由于有效信息极少而被直接剔除，除此之外，采用常见的处理方式，即用均值填充数值型变量、用众数填充类别型变量。考虑到特征缺失值不多，这里便采用了最简洁普适的方案。而对于总数据，发现不存在重复数据，故已完成数据清洗中数据重复值和缺失值的处理。

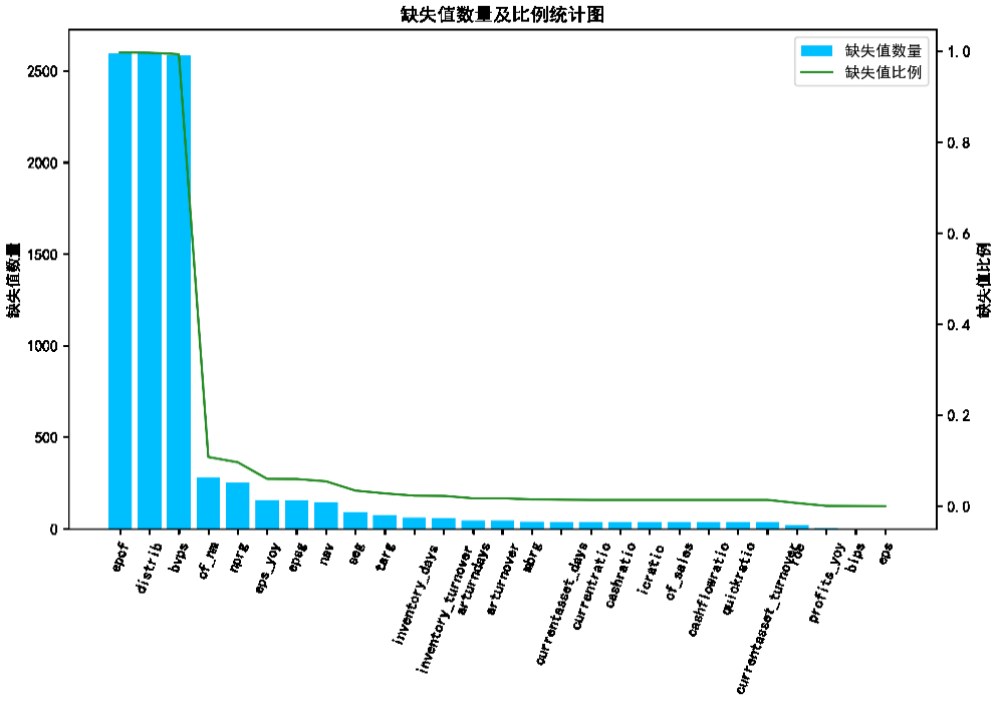

图 3-6 缺失值统计图

3.2.2 数据总体探索

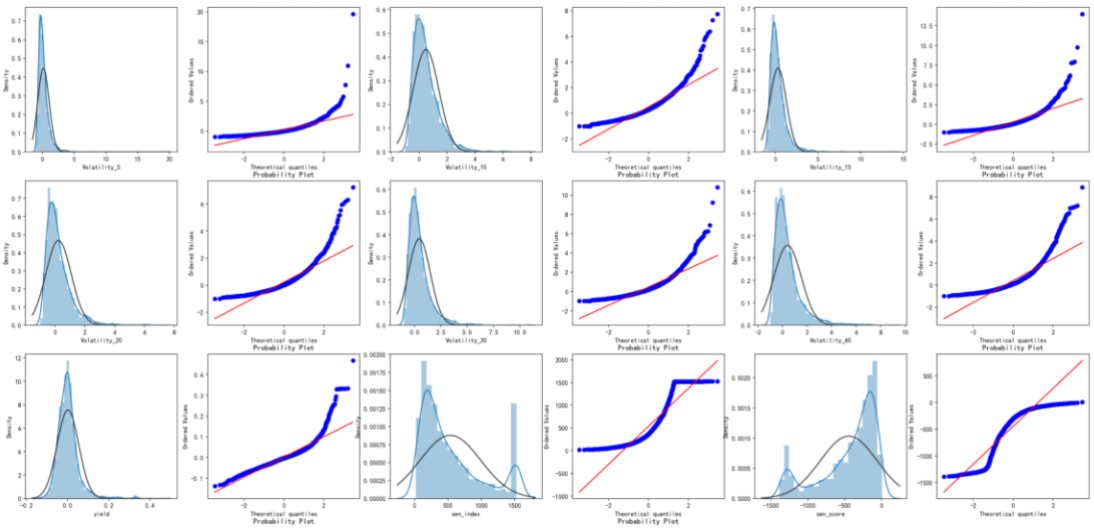

图 3-7 直方图和 QQ 图

直方图和 QQ 图是反映特征变量分布的重要可视化图表，从直方图和 QQ 图可以看出，部分变量（波动率 VOLATILITY）呈现偏正态分布，这不满足后续算法模型的要求，后续将

采取取对数转换等操作；并且基本面因子、股价因子、市场情绪因子的数据分布呈现出明显的差异化，后续在特征工程上可根据这三个层面来进行聚类。

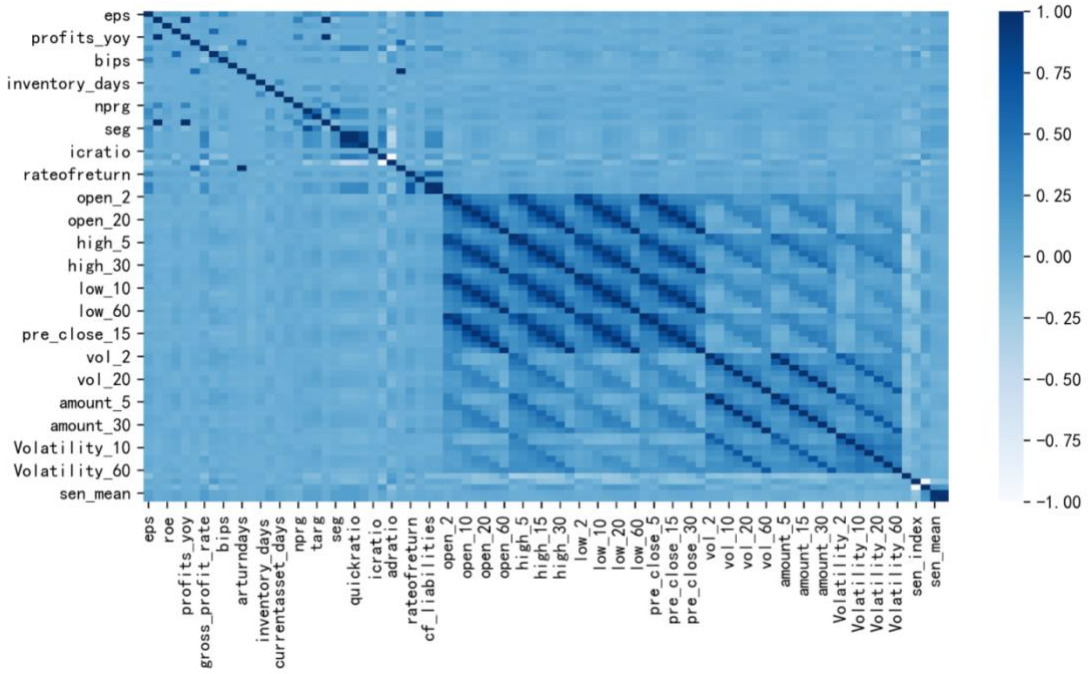

图 3-8 变量热力图

热力图又名相关系数图。根据热力图中不同方块颜色对应的相关系数的大小，可以判断出变量之间相关性的大小。热力图通过计算相关系数，相关系数度量出变量之间的线性相关关系；也就是说，相关系数越高，则变量间的线性相关程度越高。热力图不仅有助于发现数据间的关系、找出极值，也常用于刻画数据的整体样貌，方便在不同数据之间进行比较。从热力图可以看出，总数据的变量之间的相关性差异大。SEG（股东权益增长率）和 ICRATIO（利息支付倍数）相关性极高或 SHEQRATIO（股东权益比率） 和 BIPS（每股主营业务收入）极低的变量组。

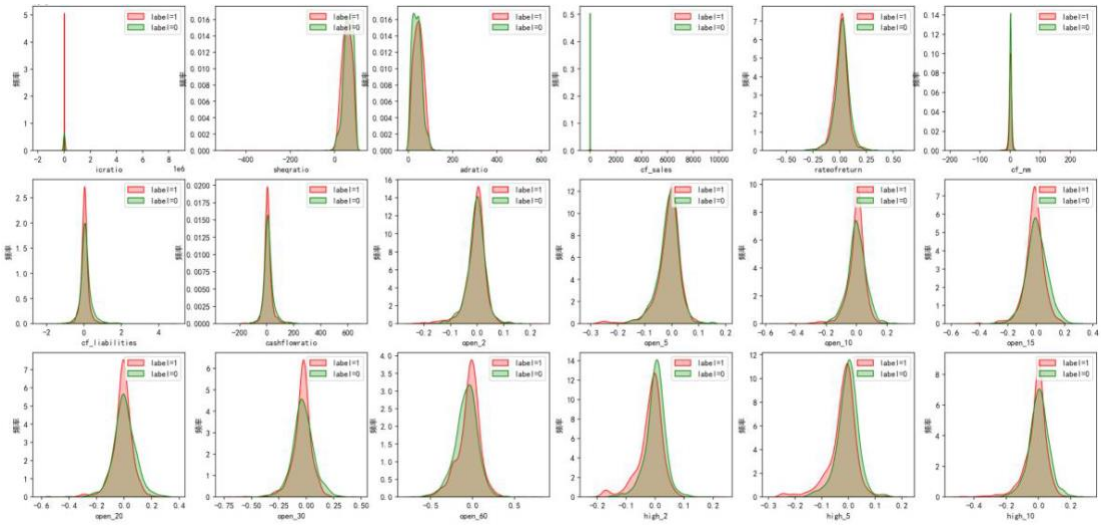

图 3-9关联标签下的变量分布

不仅通过机器学习模型训练数据后发现目标标签与训练特征之间的相关性在数据描述

性统计阶段考虑到不同标签下变量分布存在差异，从图 3-9关联标签下的变量分布可以看出，OPEN_60 等变量在不同标签下分布差异大，说明该因子能良好地区分股价涨跌，值得在实务中受到投资者以及股票公司重点关注。

3.2.3 代表变量探索

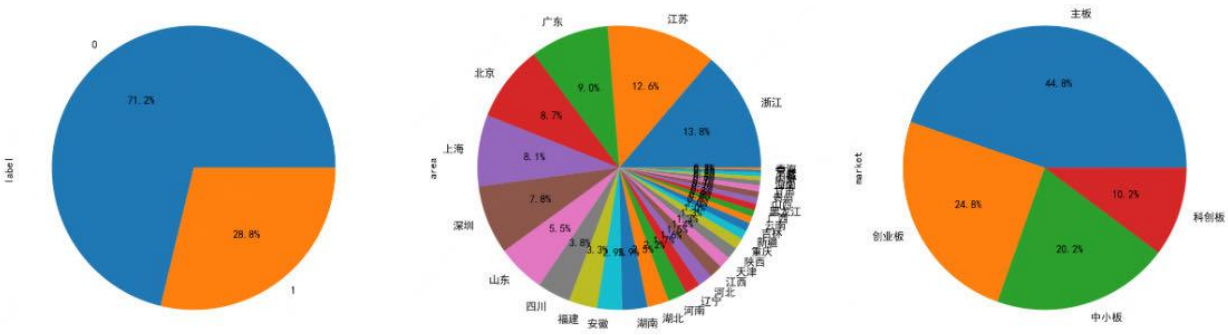

图3-10股价涨跌标签、公司地域、市场板块饼图

为更加深度挖掘能从整体角度理解股票市场，从众多因子中选取股价涨跌标签、公司地域、市场板块三大宏观特征进行描述性统计。从图中三张股价涨跌标签、公司地域、市场板块饼图可以看出，2799家发行股票公司有28.8%即806家是股价上涨的，其余71.2%即1993家发行股票公司是股价下跌的;发行股票所在公司主要分布在浙江、江苏、广东、北京、上海、深圳省市，这6个区域的发行股票公司数量占总公司数量超过半成;市场板块中主板占据近一半，而后是创业板、中小版和科创版。

为探究因子间的内在联系和相互作用，选取净资产回报率ROE、每股收益EPS、股东权益增长率SEG这三个财务状态代表性的变量，绘制了其在不同标签下与市场板块关联的小提琴图 3-11 代表性财务因子与市场板块小提琴图所示。其中可以看出明显的规律，例如主板股票所在公司的净资产回报率ROE相对更高，而科创板股票所在公司的每股收益EPS、股东权益增长率SEG相对更高，从这个角度可以得出不同板块之间的同一特征变量依然存在差异，因此可以由描述性统计得出发行股票公司所在版块是一个重要的变量特征。另一个惊奇的发现在于，在本次汇总的数据中，科创板股票的净资产回报率ROE、每股收益EPS这两个财务因子呈现出与股价的负相关关系，即净资产回报率和每股收益越高，未来股价反而倾向于下跌。这对于志在科创板股票投资的投资者和企业经营者来说都具有一定的参考意义，两方都可以重点监控这净资产回报率ROE、每股收益EPS来分析股价的走势。

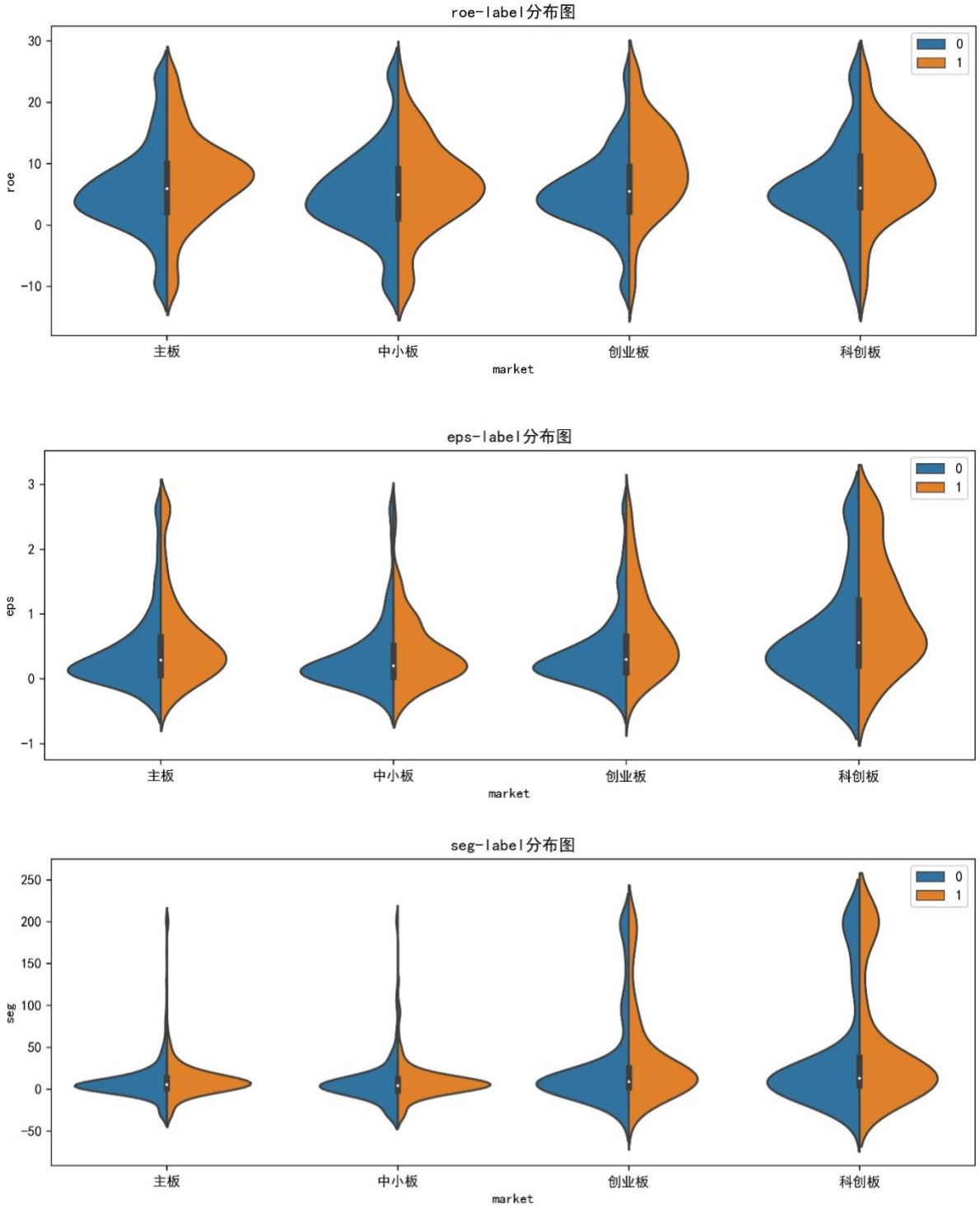

图 3-11 代表性财务因子与市场板块小提琴图

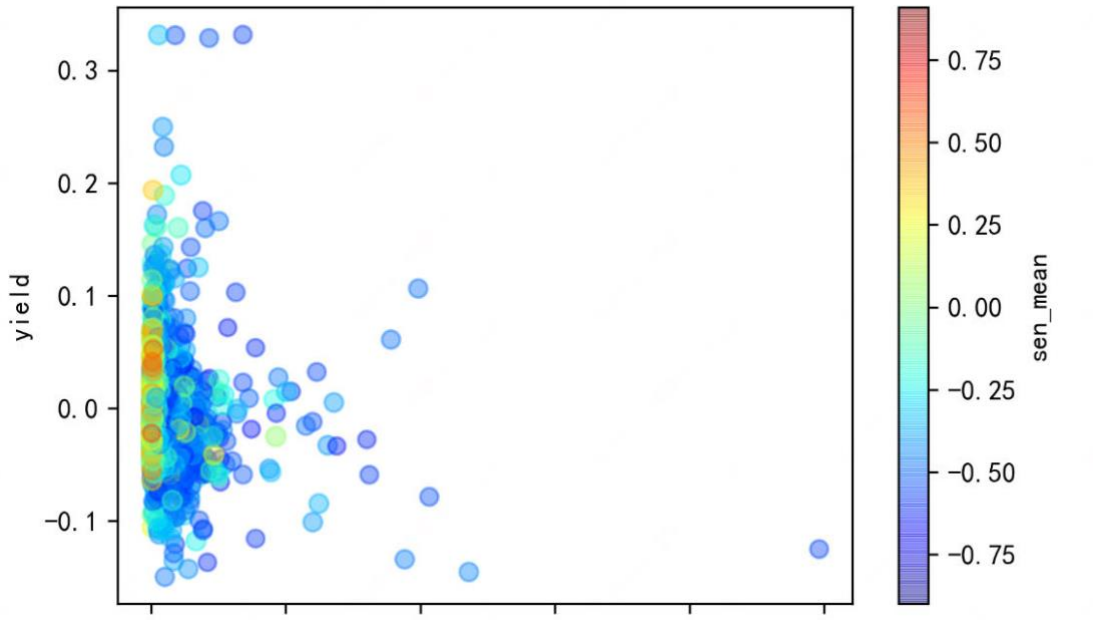

图3-12市场情绪因子气泡图

作为本次论文中重要引入的股票市场情绪因子。为探究市场情绪因子和股票收益率的关系，绘制可视化气泡图图3-12市场情绪因子气泡图所示，其中横轴表示市场指数、纵轴表示收益率，气泡颜色代表市场情绪均值、气泡大小代表市场情绪的标准差。可以看出，收益率越低的股票的评论数越多，且情绪越低，这反映了股民们往往倾向于对下跌情况发表言论且为负面态度，而对股价上涨少有积极评论。这从宏观的角度得出股民投资者倾向于在股吧中发表负面评论，带有激动的情绪进行分析股票。

3.3 特征工程

特征工程的目的是最大限度从原始数据中提取挖掘特征，以供算法和模型使用[23]。在本项目，根据特征的要求和特征之间的关系尝试了以下几种特征工程的手段，以期提升模型效果。

3.3.1 类别型变量转换：WOE 编码

对于公司地域 AREA、所在行业 INDUSTRY、市场板块 MARKET 这三个类别型的变量，像 XGBoost 这样的算法只能接受连续变量的输入，因此需要对类别变量进行映射，转换成数值型。常用的两种变量编码方法分别为 one-hot 编码和WOE 编码，而认为前者有两个明显缺点。

one-hot 编码可能会大大增加特征矩阵的维度，很多常用的模型都难以学习这种高维稀疏数据

one-hot 矩阵相当与只是简单地给编了个号，不同分类属性之间的关系不能很好地呈现出来

基于以上两点，在本次项目中采用后一种编码方式，即 WOE 编码。WOE（证据权重）是一种对原始类别型自变量进行编码的形式，它的实现定义为：

$$\text{woe} = \ln\left(\frac{p_{y_i}}{p_{n_i}}\right)$$

其中，p_{y_i}是分组中某一类别占样本中所有这一类别的比例，p_{n_i}是分组中另一类别占样本中所有这一类别的比例。WOE 实际上反映了上述两个比例之间的差异。WOE 越大，这种差异就越大，分组里样本响应的可能性越大。

需要说明的是，一般为了弱化极值影响、增加模型鲁棒性，也可以对数值型变量进行 WOE 映射（分箱），但在本次比赛中主要使用的树模型对极值和变量分布波动并不敏感，因此只对类别型变量做了WOE编码。

3.3.2 特征衍生之聚类分箱

根据之前的数据分析可知，所有股票在基本面信息、历史走势和市场情绪这三个方面的因子数据分布呈现出同一方面趋同化、不同方面差异化的特点。因此，分别按这三个方面以及三个类别型变量转换后的 WOE 编码，对股票进行了四次聚类分箱聚类数为 50，衍生出 CLUSTER_BASIC、CLUSTER_STOCK、CLUSTER_MARKET、CLUSTER_WOE 四个新的可能因子。

4. 模型应用

4.1 建模思路

在已经获取、清洗、整理好所有特征和数据的情况下。建模部分的基本逻辑是，将其看成一个有监督的二分类机器学习问题，特征为各个方面所有可能的因子，全部来自历史数据即2022年12月27日及之前，标签为未来3天后，即2022年12月30日股价是否涨跌，选取3天为观测期是一种平衡，在于既要尽量避免短时间内即一两天股价涨跌的偶然性，又要充分利用好因子数据的时效性。

按照充分训练，合理检验的规则，将所有的样本共2799个观测按照7:3的比例随机划分作为训练集和测试集，并选取常见的AUC即ROC曲线下方的面积大小作为模型的评价指标。应用多个机器学习二分类模型，并采用网格调优、模型融合优化方式，最大程度提升模型训练效果。

4.2 模型应用及优化

基于本文的数据量级是2799行数据，80余个变量，以及本文是有监督的二分类问题。基于上述数据量级以及分类问题，决定选用经典应用场景下的逻辑回归和机器学习树型结构模型，选取模型及选取理由如下表4-1所用模型及选取理由所示，可以看出，从基础的逻辑回归到高阶的XGBoost都进行了尝试，并将在后文中比较不同模型的分类效果，增强方案的置信度和稳健性。

表4-1所用模型及选取理由

分类模型	选用理由
逻辑回归	最基础的分类模型，可作为Baseline参考对标
决策树	1.树模型，适用于分类模型，且算法逐渐优化拓展，便于进行比较探究 2.多支树模型结构，能降低过拟合的风险，非常稳定 3.易用便捷，可解释性强，在同一主题其他场景下已取得良好结果
随机森林	
XGBoost	
LightGBM	

表4-2模型优化手段

模型改进方法	说明
反馈建模前处理	根据不同模型的特点和AUC值高低，返回调整数据处理和特征工程的方式。用特征重要性高的因子再次构造组合衍生新因子
调参	采用网格搜索方法，对模型的重要参数进行调试优化，主要针对的参数简要说明如下： ①learning_rate：学习率，过大可能无法收敛，过小则收敛慢 ②n_estimators：树的个数，理论上越大越好，但占用内存和训练时间也会相应增加，且边际效益递减 ③max_depth：树的深度，控制一定深度以防止过拟合
剪枝	对于树模型，可防止过拟合。
交叉验证	考量模型的泛化能力。
模型融合	尝试多种模型的堆叠效应。

下面以决策树模型的优化过程为例展示说明，其他模型的调参过程同理如下。

Step1: 调用决策树分类模型（Decision Tree Classifier），按照初始默认参数训练，在测试集上预测，得到 AUC ≈0.7，并画出决策树如下图 4-1所示

Step 2: 利用网格搜索方法（GridSearchCV），并采用五折交叉验证，调参选取最优剪枝系数 ccp_alpha，并绘制叶子不纯度的变化情况如图 4-2所示

Step 3: 代入最优剪枝系数0.005左右，从图4-2看出，重新训练模型，并再次在测试集上预测，得到AUC≈0.7，可见模型效果得到了一定提升，画出剪枝后的决策树图4-3也可以看出，决策树有了一定简化，说明之前的模型出现了一定程度的过拟合。

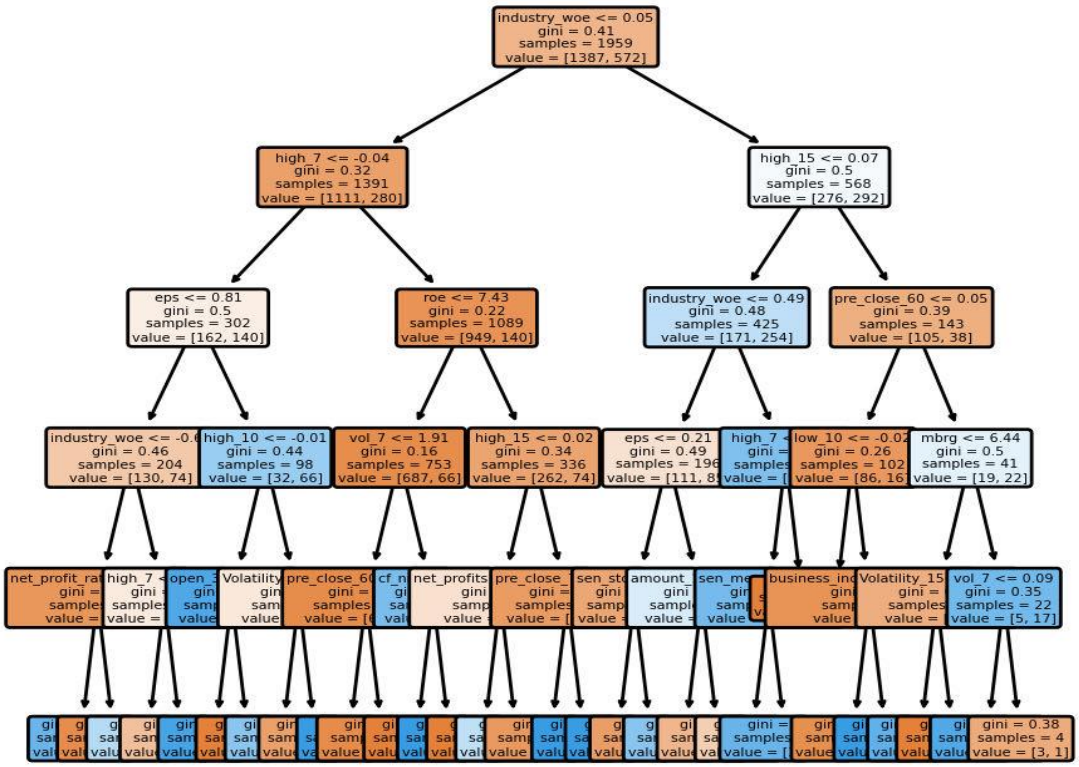

图4-1初始决策树

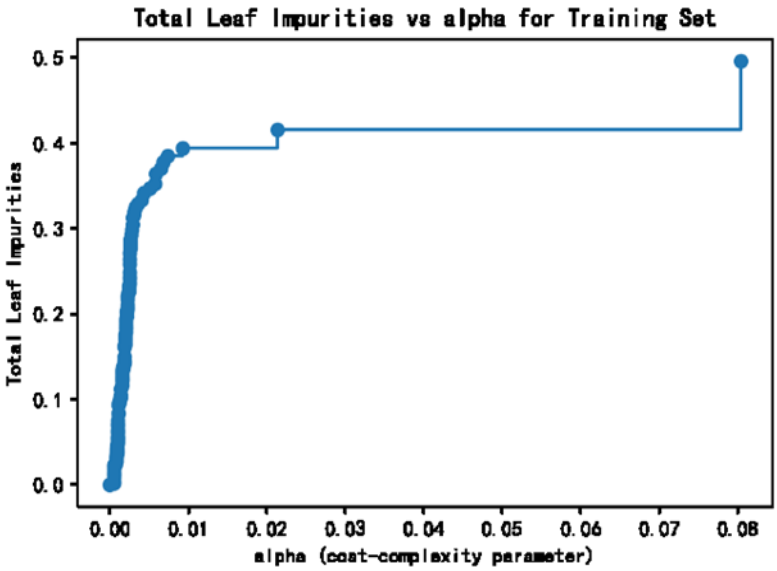

图 4-2 剪枝系数和叶子不纯度关系

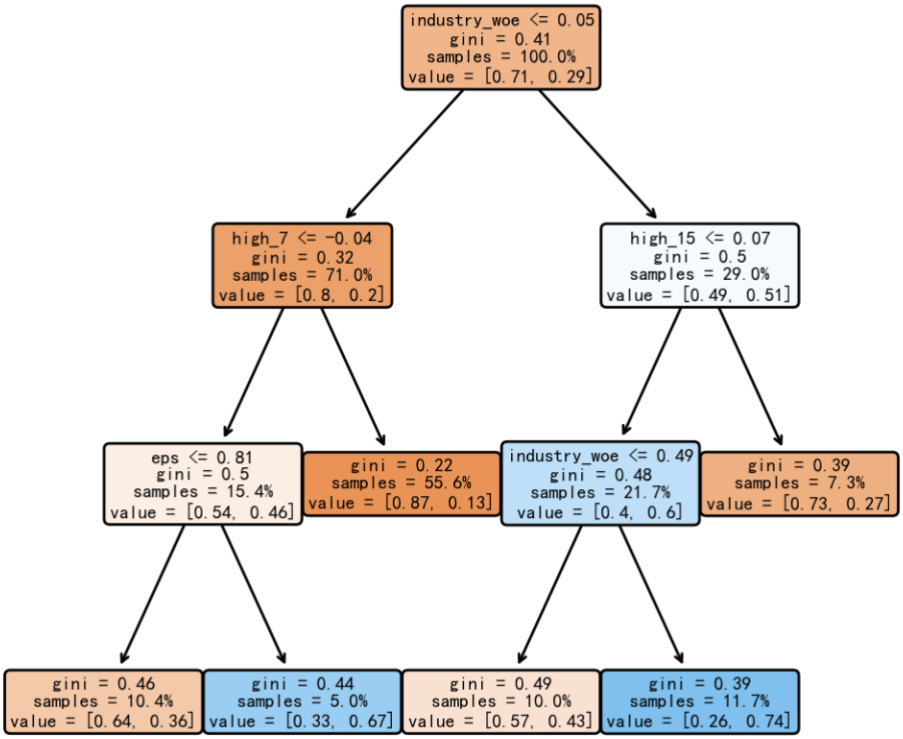

图4-3枝后的决策树

4.3 模型结果

表4-3各模型及模型融合的稳定AUC值

逻辑回归	决策树	随机森林	XGBoost	LightGBM	模型融合
0.4357	0.7199	0.7959	0.8139	0.8064	0.8100

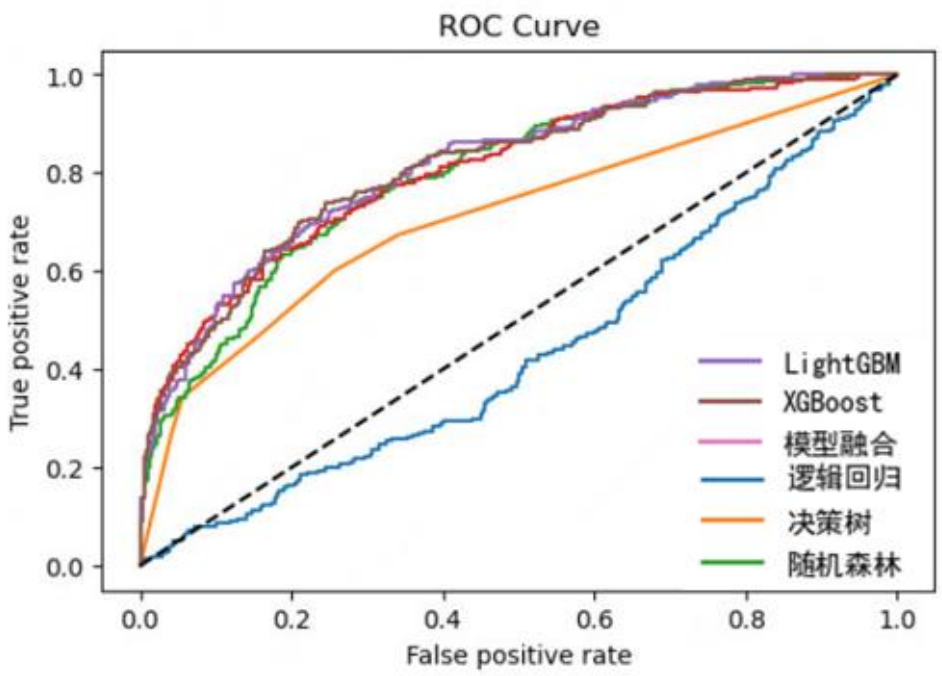

图4-4各模型 ROC 曲线

经过大量训练和不断优化，各模型及模型融合的稳定AUC值整理在表4-3中，相应的ROC曲线如图4-4各模型ROC曲线所示。可以看出，最基础的逻辑回归模型表现明显差，而表现相对最好的模型为XGBoost，AUC值达到了0.8139，证明了本项目整体方案具有良好的科学性和先进性。进一步，选择最优的XGBoost模型，统计其训练出的特征重要性高的因子排行榜如图4-5因子的特征重要性排行所示

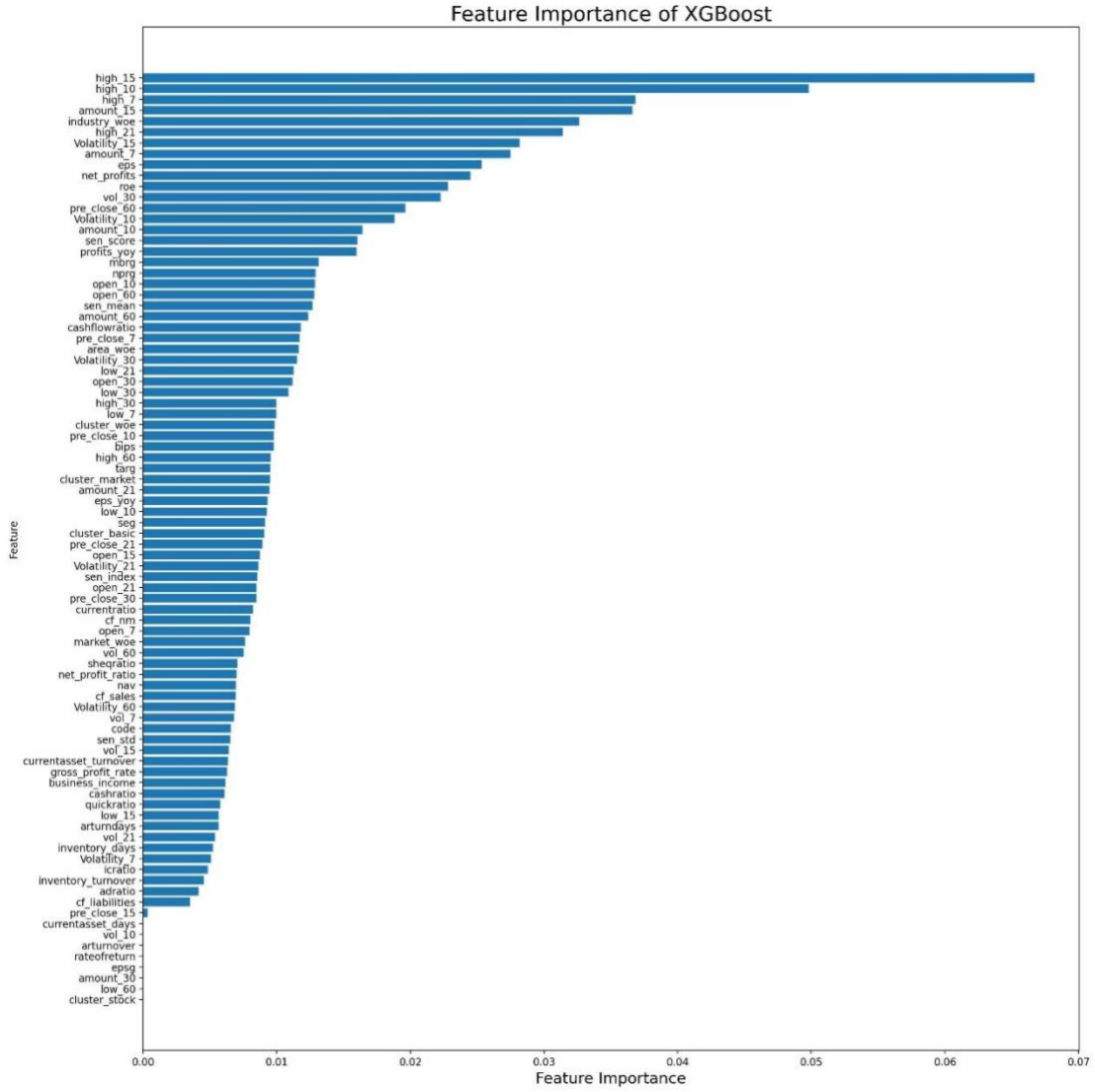

图4-5因子的特征重要性排行

基于最优的XGBoost模型，统计其训练出的特征重要性因子排行榜发现，发现沪深股票历史15日，10日，7日内股价最高价、股票交易额、股票所在行业因子的重要性高，即与股价涨跌的关联性高。其中股价最高价更是独占鳌头，说明沪深股票股价最高价与股价涨跌之间有着强的联系；

另外，相较于基本面因子和股价因子，市场情绪因子的重要性和关联度并不高，最高的是情绪总分因子重要性得分约为0.02，重要性排名为第16。说明东方财富网股吧每只股票评论的情感可能无法很好地反映未来该股票股价的涨跌，在结合数据描述性统计中的负向评论居多的情况下，原因为在东方财富网股吧发表

评论的大多为不太专业的散户，且言论常带有浓厚的情绪化口吻和激进式意味，缺乏专业分析和理性判断。

5. 结论与启示

5.1 研究结论

本文回顾股价涨跌关联因子和文本情感分类国内外相关研究文献，在吸取国内外专家学者研究成果的基础上引入市场情绪来研究对股价涨跌的影响，对基于面信息、历史走势与市场情绪相结合的股价涨跌关联因子进行研究。以沪深股票公司2022年第三季度的面数据（财务数据），2022年10月30日至2022年12月30日分日的历史走势（股价数据），2022年12月份在东方财富网股吧评论数据为依据。建立文本情感分类和有监督机器学习分类模型来探讨股票面信息、历史走势与市场情绪对股价涨跌的影响，得出以下结论：

首先，根据数据整合后描述性统计得出以下结论：沪深股票两个月60日前股票开盘价OPEN_60 变量在股票上涨和下跌时分布差异大，说明该因子能良好地区分股价涨跌。因此，从OPEN_60的曲线图分布状态可以在一定程度上预测股价上涨或下跌。当该变量分布集中和陡峭时，股价趋于上涨，当该变量分布分散和扁平时，股价趋于下跌。值得在实务中受到投资者以及股票公司重点关注；另一个惊奇的发现在于，在本次汇总的数据中，科创板股票的净资产回报率ROE、每股收益EPS这两个财务因子呈现出与股价的负相关关系，即净资产回报率和每股收益越高，未来股价反而倾向于下跌。这对于志在科创板股票的投资者和企业经营者来说都具有一定的参考意义，两方都可以重点监控这净资产回报率ROE、每股收益EPS来分析股价的走势。

其次，在算法模型拟合数据表现最优的XGBoost模型，统计其训练出的特征重要性因子排行榜发现，发现沪深股票历史15日，10日，7日内股价最高价、股票交易额、股票所在行业因子为重要性前5高的因子，即与股价涨跌的关联性高。其中股价最高价更是独占鳌头，说明沪深股票历史股价最高价与股价涨跌之间有着强的联系。对于投资者和企业经营者而言，两方都可以重点监控沪深股票历史股价最高价，因为在本研究中历史15日，10日，7日内股价最高价的特征重要性

加总超过15%。

最后，相较于基本面因子和股价因子，市场情绪因子的重要性和关联度并不高。最高的是情绪总分因子重要性得分约为0.02，其余市场情绪因子均在0.01左右，5个市场情绪因子特征重要性加总5%左右。说明东方财富网股吧每只股票评论的情感可能无法很好地反映未来该股票股价的涨跌。基于数据描述性统计阶段市场情绪数据方面，收益率越低的股票的评论数越多，且情绪越低，这反映了股民们往往倾向于对下跌情况发表言论且为负面态度，而对股价上涨少有积极评论。这从宏观的角度得出股民投资者倾向于在股吧中发表负面评论，带有激动的情绪进行分析股票缺乏专业分析和理性判断。

5.2 研究启示

本研究丰富了股价涨跌关联因子相关研究。以东方财富网股吧评论作为股票市场情绪的来源，将股票市场情绪即股民评论文本型数据引入股价涨跌关联因子探究中。通过引入机器学习中有监督分类的逻辑回归和树型分类方法，可以有效解决该二分类股价涨跌问题。基于本研究结果，可以对投资者和企业经营者带来以下启示。

首先，对于投资者而言，在金融股票市场，无时无刻不涌入海量的信息，这往往会使得投资者无法辨别有效的市场中的有效信息。在本研究中经过机器学习算法模型筛选，当投资者在分析沪深股票市场行情和做出投资决策时，可以重点关注历史股价最高价、股票交易额、股票所在行业因子、股价波动率等指标。沪深股票板块投资者可以在原有投资习惯下，关注上述指标的动态变化，以此对股价涨跌有更明晰的认识。

其次，对于股票公司经营者而言，股价涨跌一方面是反应企业经营状况的‘晴雨表’，另一方面对股票权益人和投资者的信心有十分重要的作用。因此企业经营者需要在日常企业经营过程中，需要重点监控股价最高价、股票交易额、股票所在行业因子、股价波动变化情况并从企业维度挖掘动态变化的深层原因，这对于企业健康持续运营有着关键的作用。

5.3 研究展望

本文的研究需要进一步的研究和改进，主要体现在以下两点：

第一，研究样本存在局限性。本研究由于时间限制，仅采集以沪深股票2799家公司2022年第三季度的面数据（财务数据），2022年10月30日至2022年12月30日分日的历史走势（股价数据），2022年12月份在东方财富网股吧评论数据为依据。共2799条数据，80余个特征因子，虽然达到了分类模型算法的标准，但还是存在样本量偏少、代表性不强等问题。后续研究可以扩大研究范围与数量，即拓展股票板块和数据周期，让样本更具有代表性。且扩大样本量可以提高样本的多样性，增强研究的可靠性。

第二，研究模型存在局限。本研究采用客观数据，建立文本情感分类算法和机器学习有监督二分类算法，模型相对简单。在今后的研究中，可以引入可尝试其他机器学习算法深度学习算法或是计量模型，进一步探索基于财务数据、股价走势和股民评论相结合的股价涨跌关联因子研究。

参考文献

[1] 闫璐. 网络舆情对企业价值的影响[D]. 吉林大学,2022.DOI:10.27162/d.cnki.gjlin.2022.007653.

[2] 金雪敏. 基于深度学习的财经文本情感分析技术研究[D].电子科技大学,2022.DOI:10.27005/d.cnki.gdzku.2022.000386.

[3] 王俊. 基于大数据视角的数字化互动媒体对股票市场影响性研究[D].西南财经大学,2019.DOI:10.27412/d.cnki.gxncu.2019.000906.

[4] 胡丹.金融学文本大数据挖掘方法分析[J].互联网周刊,2022(09):12-14.

[5] 黄亮.投资者情绪会影响股市收益率吗[D]. 湘潭大学,2019.DOI:10.27426/d.cnki.gxtdu.2019.000030.

[6] 熊健,张晔,董晓林.金融科技对商业银行经营绩效的影响：挤出效应还是技术溢出效应?[J]. 经济评论,2021(03):89-104.DOI:10.19361/j.er.2021.03.06.

[7] 曾鸿,苏越.我国经济增长与股票市场运行关系实证研究[J].特区经济,2019(10):60-64.

[8] 于赐龙,史振宇,谢允昊,黄军宏.基于自然语言处理的舆情分析和股价涨跌预测系统[J].系统工程,2021,39(05):114-123.

[9] 张明旭.投资组合选择理论与中国证券投资基金实务[J].商业文化,2021(17):20-21.

[10]刘新月,程希明.基于财务指标的股价涨跌预测模型[J].北京信息科技大学学报(自然科学版),2022,37(01):96-100.DOI:10.16508/j.cnki.11-5866/n.2022.01.016.

[11]崔炎炎,刘立新.网络舆情赋能金融科技股票收盘价预测研究[J].统计研究,2022,39(06):148-160.DOI:10.19343/j.cnki.11-1302/c.2022.06.010.

[12]徐攀.基于投资者情绪和关注度的股价涨跌预测研究[D].南京财经大

学,2021.DOI:10.27705/d.cnki.gnjcj.2021.000158.

[13]刘素辉.基于多源异构数据的股价趋势预测研究[D].北京科技大学,2021.DOI:10.26945/d.cnki.gbjku.2021.000375.

[14]顾文涛,王儒,郑肃豪,杨永伟.金融市场收益率方向预测模型研究——基于文本大数据方法[J].统计研究,2020,37(11):68-79.DOI:10.19343/j.cnki.11-1302/c.2020.11.006.

[15]张姝.基于情感分析的在线评论文本分类研究[D].江南大学,2022.DOI:10.27169/d.cnki.gwqgu.2022.001772.

[16]王利利. 基于深度学习的中文文本情感分类研究及应用[D].中国矿业大学,2019.

[17]刘飞.基于逻辑回归方法的股票价值分析——以数字货币板块为例[J].投资与合作,2022(12):32-34.

[18]袁欣宇. 基于机器学习分类算法的借贷风险评估解决方案[D].黑龙江大学,2022.DOI:10.27123/d.cnki.ghlju.2022.001176.

[19]尹儒.模型决策树方法研究[D].山西大学,2019.DOI:10.27284/d.cnki.gsxiu.2019.000218.

[20]张颖. 基于 Stacking 模型的网络车贷违约客户识别研究[D].重庆工商大学,2022.DOI:10.27713/d.cnki.gcqgs.2022.000183.

[21]顾祎芸. 不同行情下融资融券对股价波动的影响研究[D].华东师范大学,2022.DOI:10.27149/d.cnki.ghdsu.2022.001394.

[22]李文秀.投资者情绪对股票市场收益率影响的实证研究[J].老字号品牌营销,2022(24):55-57.

[23]满成剑. 基于特征自编码和时间卷积网络的股价预测研究[D].山东大学,2019.

[24]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[25]Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[26] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.