金融投资策略设计报告

中药行业的量化投资策略
——基于 *XGBOOST* 的多因子选股模型

中药行业的量化投资策略——基于 XGBoost 的多因子选股模型

摘要：在面临全球经济下行压力的背景下，中药行业凭借新冠中药治疗、国家政策扶持、群众医疗卫生健康意识提高等机遇，成为近两年来的投资热点板块。目前，中药市场一片繁荣，行业正处于上升阶段，为了让投资者们抓住机遇，选择具有高成长价值的中药板块股票进行投资，我们设计了这份中药行业的量化投资策略。

本文首先结合中药行业的宏观经济环境，分析了当前所处的经济周期、行业政策及行业背景等多个基本面维度，分析了中药行业目前的发展状况。其次，我们利用 XGBoost 算法优化传统的多因子选股模型，利用中药行业 30 只上市公司股票在 2008 年-2019 年间的因子数据进行 XGBoost 算法的优化训练，最终得出因子重要性排序，并据此结果得出了 3 只具有高成长性、高投资价值的股票；我们还通过技术分析和财务分析两个微观层面对 3 只股票进行更加全面的评价。此外，我们利用 2019-2021 年两年的数据对模型的精度和选股效果进行回测，以沪深 300 指数为基准，通过收益率、风险度量等方式衡量我们的量化选股模型所具有的优越性。最后，我们介绍了 XGBoost 模型能够怎样减少投资风险，并为投资者提供了投资组合策略建议。

关键词：中药行业、XGBoost 算法、多因子选股、绩效评价

Abstract

Against the backdrop of the downward pressure of the global economy, the Chinese medicine industry has become a hot investment sector in the past two years with opportunities such as new crown Chinese medicine treatment, national policy support, and increased awareness of public health care and health. Currently, the Chinese medicine market is booming and the industry is on the rise. In order for investors to seize the opportunity and select stocks with high growth value in the Chinese medicine sector for investment, we have designed this quantitative investment strategy for the Chinese medicine industry.

In this paper, we first analyze the current development of the Chinese medicine industry by taking into account the macroeconomic environment of the industry and analyzing several fundamental dimensions such as the current economic cycle, industry policies and industry background. Secondly, we use the XGBoost algorithm to optimize the traditional multi-factor stock selection model, using the factor data of 30 listed stocks in the Chinese medicine industry during 2008-2019 for the optimization training of the XGBoost algorithm, and finally arrive at the factor importance ranking, and based on this result, we arrive at three stocks with high growth and high investment value; we also use technical analysis and financial We also evaluate the 3 stocks more comprehensively through two micro-levels: technical analysis and financial analysis. In addition, we use two years of data from 2019-2021 to back-test the accuracy and stock selection effectiveness of the model, using the CSI 300 index as a benchmark to measure the superiority of our quantitative stock selection model by means of return and risk metrics. Finally, we present how the XGBoost model can reduce investment risk and provide investors with portfolio strategy recommendations.

Keywords: Chinese medicine industry, XGBoost algorithm, multi-factor stock selection, performance evaluation

目录

一、项目背景 1
二、宏观经济分析 3
2.1 宏观经济背景分析 3
2.1.1 国内疫情 3
2.1.2 俄乌冲突 4
2.1.3 国内通胀与美联储加息 5
2.2 宏观经济周期分析 5
2.3 宏观经济政策分析 6
2.3.1 宏观经济发展政策 6
2.3.2 中药行业相关政策 7
三、行业分析 8
3.1 全球及中国大健康行业运行情况 8
3.2 中药行业现状 10
3.2.1 产业组织 10
3.2.2 新冠的中医治疗 10
3.3 中药产业结构 11
3.3.1 产业上下游 11
3.3.2 产业细分： 12
四、量化选股 8
4.1 多因子选股模型 13
4.1.1 多因子选股模型基本介绍 13
4.1.2 多因子选股模型流程图 14
4.2 XGBoost 算法 14
4.2.1 XGBoost 算法基本介绍 14
4.2.2 XGBoost 算法流程图 15
4.2.3 XGBoost 算法与传统多因子模型比较 15
五、基于 XGBoost 的多因子选股模型构建 16

5.1 方案设计框架..... 16
5.2 因子池构建与数据预处理..... 16
5.2.1.构建因子池..... 16
5.2.2 数据预处理..... 17
5.3 XGBoost 算法参数优化详解及因子筛选..... 19
5.4 股票筛选结果..... 22
六、技术分析 24
6.1BOLL 指标 24
6.2 形态理论 25
6.2.1 矩形整理 25
6.2.2 突破形态之双重顶 26
七、财务分析 28
7.1 盈利能力 28
7.1.1 每股盈利 28
7.1.2 营业利润率 29
7.1.3 净资产收益率 29
7.2 成长能力 30
7.2.1 营业收入增长率 30
7.2.2 净利润增长率 31
7.3 偿债能力 32
7.3.1 现金流量比率 32
7.3.2 速动比率 32
7.4 营运能力 33
7.4.1 总资产周转率 34
7.4.2 存货周转率 34
7.4.3 应收账款周转率 35
八、历史回测与绩效评价 37
8.1 模型准确检验..... 37

8.2 历史回测绩效评价..... 37
九、风险控制 39
9.1 定期修正策略来预防风险 39
9.2 模型自身存在的优势可以减少风险 39
9.2.1 更好的容忍异常值缺失值 39
9.2.2 更广的适用范围 39
9.2.3 更加的稳定 39
9.2.4 减少人为随意性 40
9.3 运用大量的因子减少风险 40
9.4 特定的筛选方法降低风险 40
十、主要结论和投资建议 41
参考文献 43
附录 44

图目录

图 1 中药行业营业收入变化（亿元）..... 1
图 2 中药行业销售毛利率变化..... 2
图 3 中药概念股近十年涨幅..... 2
图 4 全国疫情分布图（4 月 1741 日）以及全国本土确诊病例：当日新增 3
图 5 2022 年乘用车月销量同比变化趋势（万元） 4
图 6 每日原油和石油产品现货价格 5
图 7 2020-2021 季度产业增加值 6
图 8 中国大健康产业居民医疗支出数据 8
图 9 中国大健康产业居民医疗支出数据 9
图 10 2020 年中国新经济投资热度行业分布及交易情况 9
图 11 2017-2022 年中国中药行业市场规模 10
图 12 2021 年中国中药产业细分规模 12
图 13 多因子选股模型流程图..... 14
图 14 XGBoost 算法流程图 15
图 15 方案设计技术框架..... 16
图 16 因子重要性输出 1..... 20
图 17 因子重要性输出 2..... 21
图 18 因子重要性输出最终结果..... 22
图 19 云南白药 BOLL 图..... 24
图 20 矩形整理的 2 种形态..... 25
图 21 白云山矩形整理 26
图 22 片仔癀双重顶形态..... 27
图 23 每股收益对比图..... 28
图 24 营业利润率对比图..... 29
图 25 净资产利润率对比图..... 29
图 26 营业增长率对比图..... 30
图 27 净利润增长率对比图..... 31
图 28 偿债能力对比图..... 32
图 29 速动比例对比图..... 33
图 30 总资产周转率对比图..... 34
图 31 存货周转率对比图..... 35
图 32 应收账款周转率对比图..... 35
图 33 ROC 曲线模型图..... 37

图 34 累计收益率比较..... 38

表目录

表 1 中药行业相关政策 7
表 2 新冠中医治疗机理 11
表 3 中药上下游产业链..... 11
表 4 公司层面因子分类..... 13
表 5 XGBoost 算法与传统多因子模型比较 15
表 6 财务、技术等部分因子..... 17
表 7 宏观、债券、投资等因子数据..... 17
表 8 数据转换..... 19
表 9 gamma 参数优化结果..... 20
表 10 cv 参数优化结果 21
表 11 中期-最终得分表..... 22
表 12 中期股票投资组合构建..... 23
表 13 短期-最终得分表..... 23
表 14 短期股票投资组合构建..... 23
表 15 总资产周转率..... 38

一、项目背景

中医是一笔宝贵的财富，在与疾病斗争的漫长岁月中，对中国的繁荣昌盛以及对人类的身体健康起到了至关重要的作用，具有重大战略意义。如今，在这场突如其来的新冠肺炎疫情之中，中医更是发挥了巨大的作用。2020 年五月十六日，国家中医药管理局发布了有关加强中医药工作的通知。习近平在今年六月二日主持召开了关于医疗卫生领域的重大风险的专家学者座谈会。习近平主席指出，中西医结合，中西药并用，是中医药传承精华、守正创新的一个生动实践。

在 2015-2019 年间，中药行业的营业收入呈连续增长趋势。2020 年由于疫情的影响，限制医院治疗以及感冒药退烧药等处方药的购买，中药行业的营业收入下降了 3.8%。

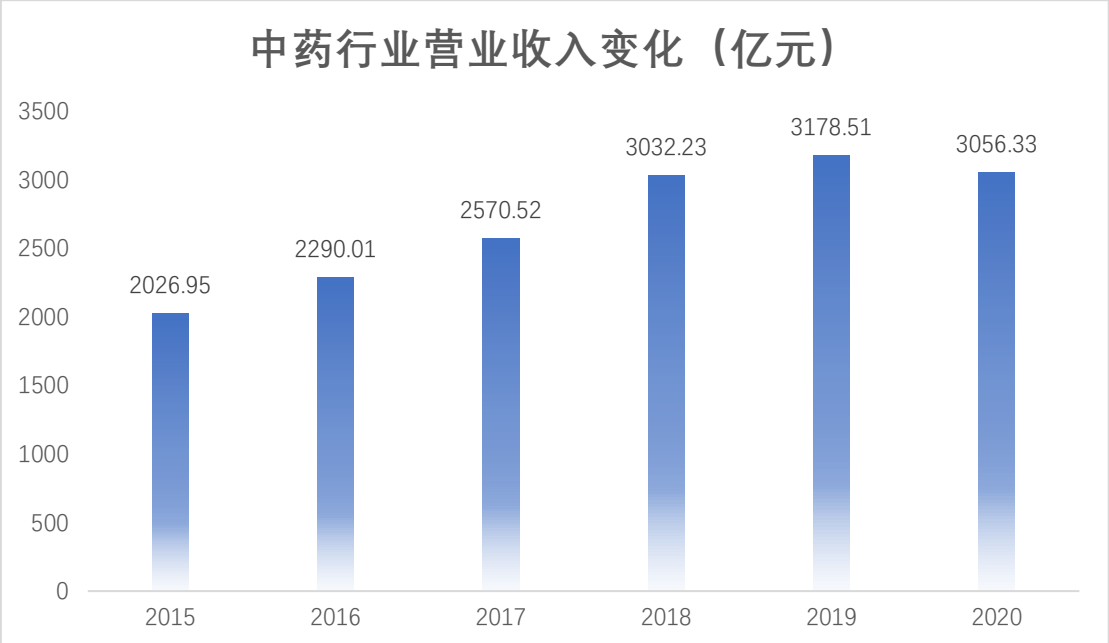


图 1 中药行业营业收入变化（亿元）

同时中药行业也是中国国内盈利能力较强的行业，产品利润率较高。2015-2021 年间，中药行业的毛利率达到了 40%以上，利润率高于我国大部分其他行业。

中药概念股从 2012 年年底的 881.41 元到 2022 年 4 月 15 日涨到了 2007.43 元，涨幅达到了 127.75%。尽管略有波动，但整体趋势利好。在 2020 年大盘指数跌跌不止的时候，中药行业总体仍呈上涨趋势。

图 2 中药行业销售毛利率变化

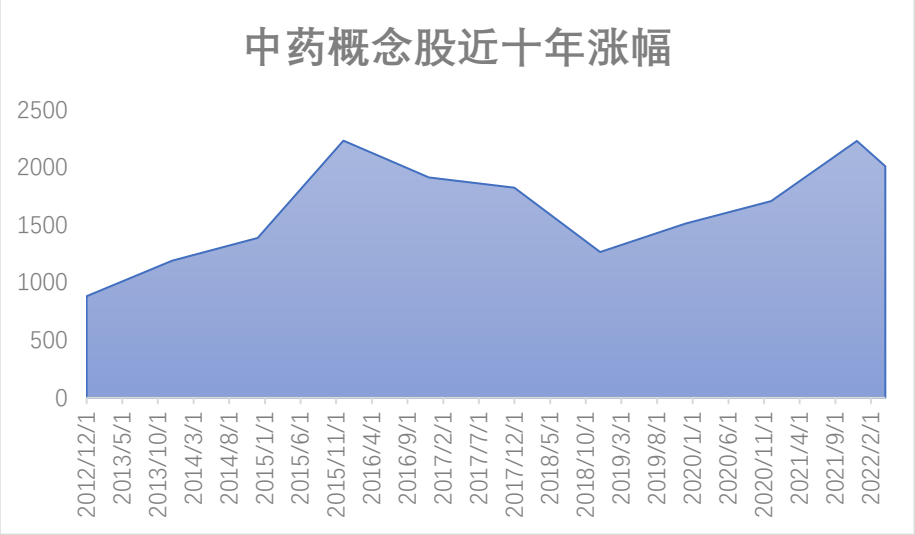


图 3 中药概念股近十年涨幅

虽然中药行业的利润规模在 2022 年第一季度出现了下降，但在利润下降的同时，行业的盈利能力仍处于高水平，另外，在疫情背景下，国家对中药行业扶持的政策也为这个行业的走向奠定了发展新格局。

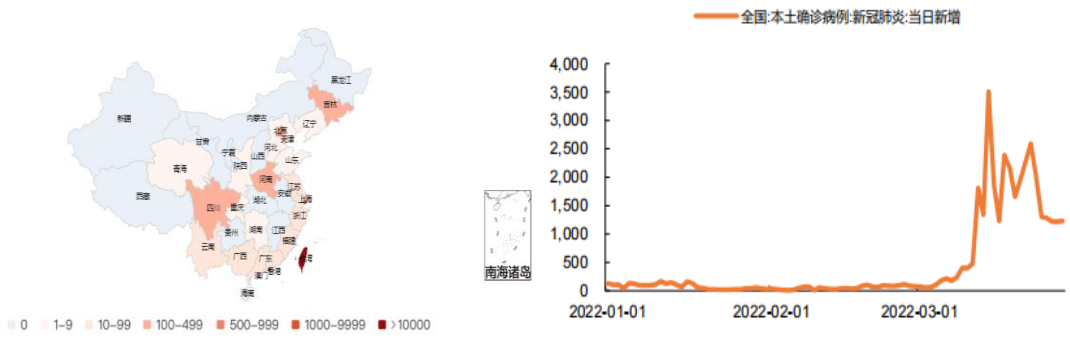
二、宏观经济分析

2.1 宏观经济背景分析

2.1.1 国内疫情

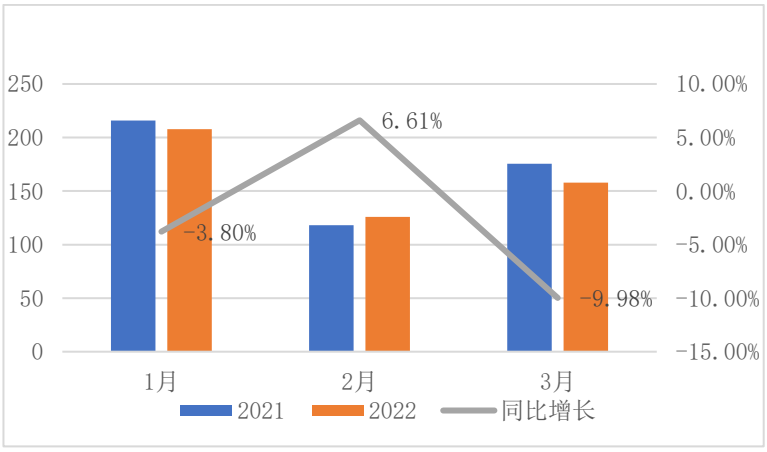
自 2022 年依始至今，国内及海外的疫情都处在一个较高风险的状态。尤其是对于国内，今年三月份以来陆续在吉林省、广东省、福建省、上海市等地疫情又迎来了小爆发，新增感染者的数量呈指数式增长。同时，对于其中重要城市的广州、上海的疫情防控形势恶化，使得疫情对经济影响范围蔓延至全国。面对“高危”的疫情冲击，政府采取的疫情防控措施等级随之上升。而更为严格的疫情防控措施也使得人员流动、物流等一系列活动显著受限，速度放缓。从总体上看，对于消费的影响压力最大，同时工业生产、投资等活动也一定程度受限。

于消费方面，疫情冲击严重影响消费需求，疫情高危地区乃至全国的消费需求呈断崖式下降。具体表现于：1）销售面积（商品房）同比存在较大幅度的下降。2）据乘联会报告披露， 1-2 月的乘用车月销量呈上涨趋势，而受疫情冲击最为严重的 3 月，该月乘用车零售同比下降 9.98% 。



资料来源：新浪新闻

图 4 全国疫情分布图（4 月 17 日）以及全国本土确诊病例：当日新增



资料来源：乘联会

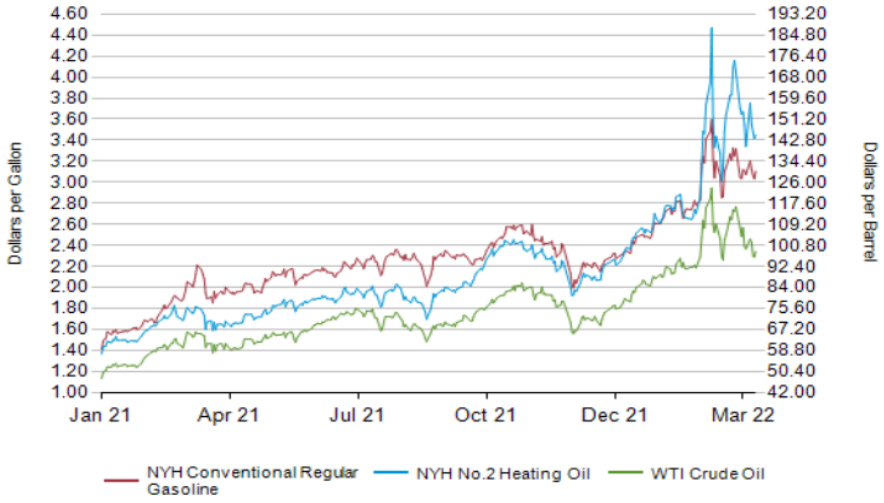
图 5 2022 年乘用车月销量同比变化趋势（万元）

于国际贸易方面，疫情趋势严峻化，防控升级，对我国进出口的影响呈逐渐增大的态势。对于当前中国的疫情现状，陆运等物流活动速度显著放缓。一方面，这使得原材料的运输速度减慢，导致生产因原材料供应不足而降低生产效率。另外一方面，多地的产品的供给由于物流放缓未能形成有效供给。于外贸港口方面，上海市的外贸港口吞吐量位中国第二，其外贸货物占比最高，占比高达 55% 。由于疫情防控措施的升级，且上海市当前的疫情处于一个较为不可控的状态， 4 月 12 号上海港港区实行全环闭管理，这在一定程度上减缓了货物的流动速度以及港口正常运转水平。

2.1.2 俄乌冲突

“俄乌冲突”爆发于年初二月末，面对来自美国为首的“北约派”的制裁，俄罗斯对本国商品出口发布了限制令予以回应。由于俄罗斯本身为出国大国，尤其于能源出口方面，这使得国际上大宗商品价格迅速上升。但这对我国经济层面的直接冲击较小，但对我国经济发展仍是有着显著间接影响。针对于俄罗斯限制能源出口方面，由于此前欧盟各国的能源供给大多数源于俄罗斯，所以这将很大概率上导致欧洲各国进入能源配给限制而进入衰退，将产生一股由欧盟各国波及至世界范围的经济冲击波。当前由于该军事冲突，国际油价已经出现了不合理的波动趋势。根据 EIA 所提供的国际原油价格变化态势图，国际油价自“俄乌冲突”以来呈现了反常

水平的上涨趋势。这对整个全球经济体的发展是不理性的，对于中国经济的国际贸易部分会受到较大冲击。



资料来源：EIA

图 6 每日原油和石油产品现货价格

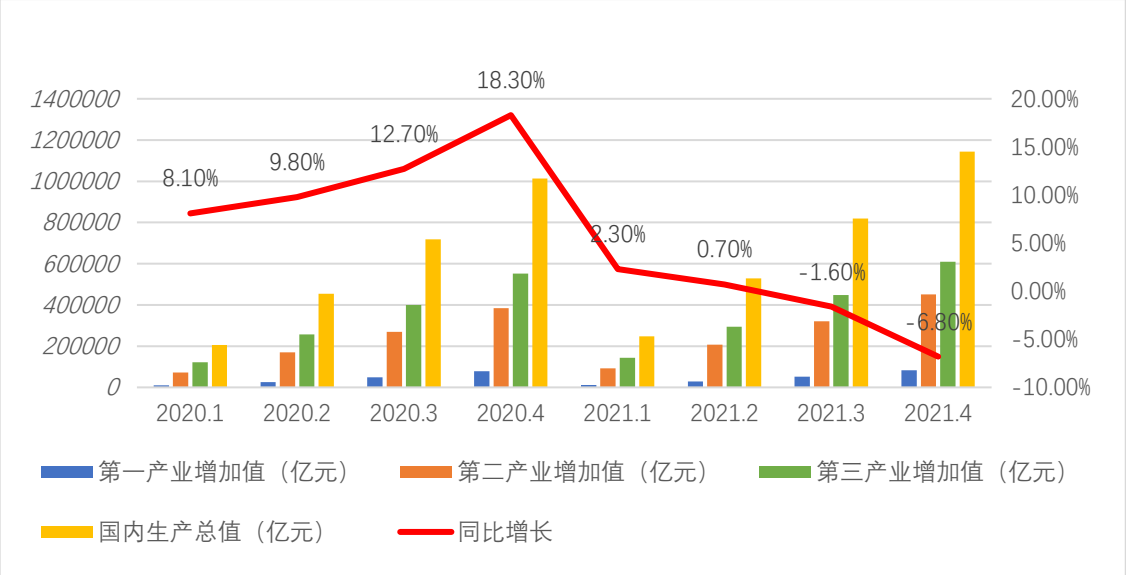
2.1.3 国内通胀与美联储加息

相比于美国当前经济发展出现了严重的通胀现象，当前我国的通胀压力不大。根据 4 月 11 日国家统计局所公布的报告披露：3 月 CPI 同比为 1.5%，较上月回升 0.6%；同时该月 PPI 同比上涨 8.3%，较 上月存在 0.5%的下降值。由此看，当前我国的通胀水平维持一个较为正常的状态。

放眼至国际局势，由于欧美国家自 2021 年以来通胀压力呈显性化态势。再加之“俄乌冲突”的冲击，美国向乌克兰实行援助，美国今年 2 月 CPI 同比增速 7.9%，突破 2008 年高位。因此，美联储也在今年年内预计将实行 6 次加息，这将会使得美元升值，全球美元资产回流至美国本国。于中国投资层面，这会使得外资回流，北向资金净流出。

2.2 宏观经济周期分析

宏观经济层面下，2021 年我国 GDP 增速为 8.1%，名义同比增速为 12.8%，GDP 平减指数为 4.38%。受疫情反复、政策性因素对经济支持边际弱化以及中长期结构性调整等措施对短期经济产生冲击等因素影响，GDP 当季同比呈现前高后低的走势。



资料来源：国家统计局

图 7 2020-2021 季度产业增加值

工业运行总体平稳，三大需求稳固增长，国内需求持续恢复，进出口仍维持较强的韧性，CPI 温和上涨，国内通胀压力并未随着经济发展而剧增， 2021 年我国通胀率为 0.9% ，相比于美国 4.7% 的通胀率水平，我国的经济通胀维持在一个较为合理的水平。

对于宏观经济周期分析，当前我国经济发展周期处在衰退后的恢复期，国家政府也相应地提出“稳发展”的政策，经济发展拟在小周期内出现波动，长期发展拟呈平稳增长态势。

2.3 宏观经济政策分析

2.3.1 宏观经济发展政策

今年三月初，国务院金融稳定发展委员会召开专题会议，会议释放出中央高度重视经济平稳健康发展的信号，说明了当前的政策中心是经济的稳增长。其次，于货币政策方面，会议提到：货币政策的调控应把握好时间差，应完善政策的主动性、积极性，及时运用降准降息等多种货币政策工具，针对经济动荡的局势，释放中央拟实行降准降息的信号。

3 月 30 日国常会再次强调了经济稳增长的重要性，在疫情冲击下，对于这点的预期会更高以及落实力度将会更大，政策的力度可能会加大，待疫情缓解之后，稳

增长主线仍然值得关注。

2.3.2 中药行业相关政策

表 1 中药行业相关政策

《“十四五”中医药发展规划》	国家中医药管理局 2022 年 3 月 29 日	就中药行业的服务体系与高质量、开放发展提出十项主要任务，旨在推进中药行业快速高质量发展，并且推进中药文化的繁荣。
《推进中医药高质量融入共建“一带一路”发展规划》	国家中医药管理局 2022 年 1 月	提出“十四五”时期中药行业的发展目标
《医保支持中医药传承创新发展的指导意见》	国家医疗保障局；国家中医药管理局 2021 年 12 月	《意见》指明了在医疗医保方面、中医药行业的前端销售方面对行业发展的重要性。

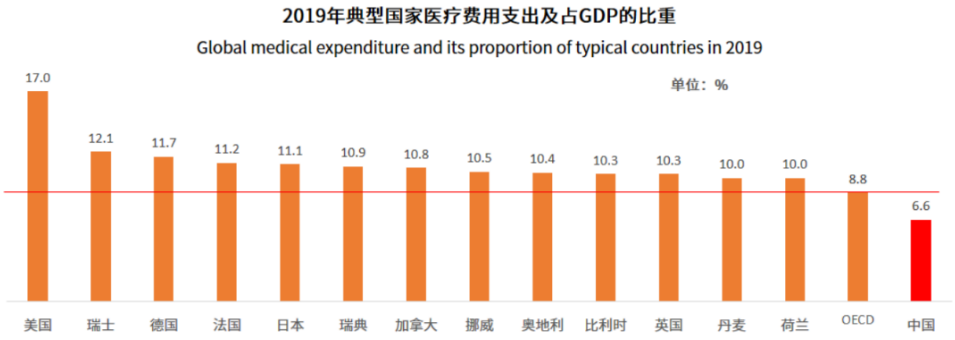
针对于中药行业的相关政策，聚焦于近期的《“十四五”中医药发展规划》，国家中医药管理局明确指出：进一步支持推动中药行业的创新、中药基药，鼓励建设相关的高素质人才；持续关注中药的推广化、国际化；从原材料上，关注中药材质量控制；社会层面，完善中医医疗服务、医保支付等多方面举措，促进行业稳健快速发展。

同样地，国家在近几年来出台了一系列引导中医药行业积极高速发展的政策文件，反映了国家对中医药行业高度重视，加上行业本身的发展潜力，中医药将会迎来一个发展春天。

三、行业分析

3.1 全球及中国大健康行业运行情况

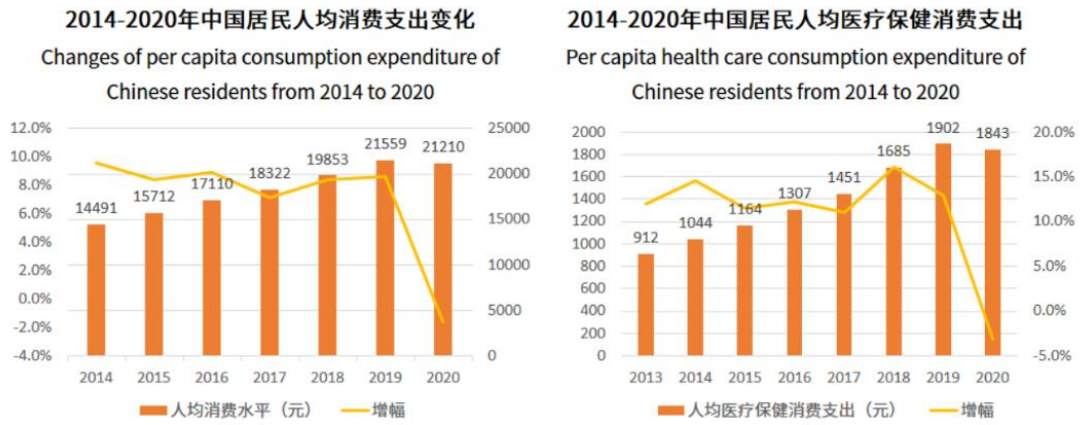
放眼全球，受到疫情、人口老龄化等因素影响，国家医疗费用支出不断增加，欧美发达国家的医疗卫生支出在国内生产总值中的比例一般都在 10%以上；近几年，随着医学技术的进步，人民对健康的需求也越来越高，可以预见，大健康产业会在相当长的时期内持续增长。



数据来源：WIND，艾媒咨询

图 8 中国大健康产业居民医疗支出数据

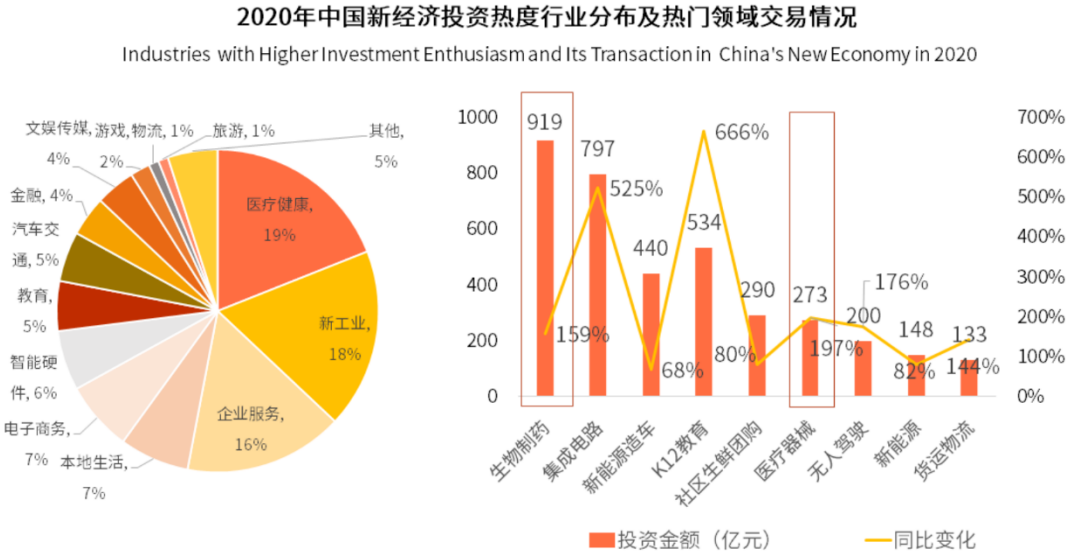
回到国内，近年来政府高度重视人民群众的健康问题，中药和其他卫生保健事业也取得了巨大的发展。2019 年 7 月，我国国家卫生健康委出台了《健康中国行动（2019——2030 年）》，体现了我们党对人民健康重要价值和作用的认识达到新高度；2022 年 5 月，国务院办公厅发布《“十四五”国民健康规划》，规划要求持续推进健康中国建设，满足人民日益增长的健康需求，同时指出“十三五”期间我国中医药服务体系持续完善，中药行业独特优势日益显现。



数据来源：WIND，艾媒咨询

图 9 中国大健康产业居民医疗支出数据

在消费端，中国 2014-2019 年人均消费支出与卫生保健支出保持了一定的同步增长态势，2020 年由于疫情的冲击，人均消费支出出现了小幅下降，医保支出也同步下降。未来，随着疫情的常态化和居民消费水平的不断提高，我国居民对卫生保健的需求将持续增加。



数据来源：WIND，艾媒咨询

图 10 2020 年中国新经济投资热度行业分布及交易情况

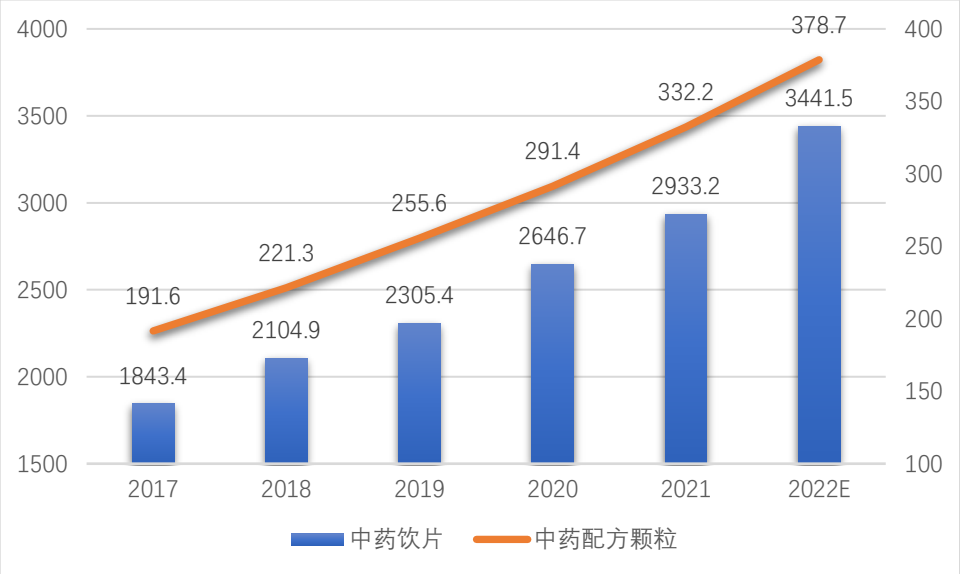
在投资端，近两年，医疗卫生是中国投资热度最高的新经济产业。其中，投资额最高的是生物制药，超过 900 亿元。足以说明，疫情下大健康产业的融资热情依然高涨。

3.2 中药行业现状

3.2.1 产业组织

中药配方颗粒市场目前呈“6+N”的竞争格局。截至 2021 年，全国各类中药配方颗粒试点企业共 60 余家，其中 6 家国家级试点企业占据了中药饮片市场 80%以上的市场份额。“中国中药”为行业龙头，据统计，中国中药集团 2020 年市场份额占比超过 52%，市场占有率排在二三位的是红日药业和华润三九，其他企业份额相对较小，市场集中度较高。

2017 年-2020 年我国中药饮片市场规模由 1843.4 亿元增至 2646.7 亿元，年均复合增长率为 8.9%，预计 2022 年我国中药饮片市场规模或达 3441.5 亿元。我国中药配方颗粒市场规模由 2017 年的 191.6 亿元增长至 2020 的 年的 291.4 亿元，年均复合增长率约 15%，占比中药饮片行业比例约为 11%。



数据来源：中商研究院

图 11 2017-2022 年中国中药行业市场规模

3.2.2 新冠的中医治疗

新冠肺炎属于中医“疫病”范畴，主要具有发病快、传播快和传染性强三个主要特点。在临床治疗方面，目前我国中医治疗主要以防病、治疗和康复为主，对患者进行治疗，以达到改善患者临床症状的目的。在我国卫健委发布的历版《新型冠状病毒肺炎诊疗方案》中，第三版开始将中药治疗方案正式纳入新冠治疗方案中，并在随后几版的《诊疗方案》中对中医治疗方案进行了不断的补充和完整。从治疗

机理上来看，中药对于新冠病毒的治疗主要在于免疫调节、炎症抑制和抗病毒三个层面。

表 2 新冠中医治疗机理

功能	作用机理
免疫调节	中药在新冠肺炎治疗中发挥双向调节作用，能够系统性地调节新冠肺炎所引发的免疫紊乱，抑制促炎细胞因子的表达水平，调节细胞因子之间的平衡，从而改善炎症对组织和器官的损害。
炎症抑制	研究表明，中药对于病毒感染引发的炎症反应有抑制作用。如：连花清瘟胶囊可抑制不同流感病毒的增殖，同时能够减少病毒诱导的炎性细胞因子基因表达；热毒宁注射液可以有效减少促炎细胞因子的表达、炎症介质的释放和生成等。
抗病毒	中药在抗病毒和症状减缓等方面具备良好的效果，包括：治疗病毒性肺炎、降低巨细胞病毒复制水平以及调节病毒感染的免疫应答等。

数据来源：国家卫健委

3.3 中药产业结构

3.3.1 产业上下游

中医药产业链包含上游中药原材料端、中游中药材加工制造端和下游中医类机构服务端，由此引申出的中医药市场可大致细分为中成药、传统中药饮片和中药保健品。

表 3 中药上下游产业链

产业链	子行业	概述
上游	中药材	中药材是指植物、动物和矿物去除非药用部位的商品药材，共有 12807 种。近年随着中药材需求幅度和市场价格提高，中药材整体市场规模呈上涨趋势，2020 年我国中药材市场规模有近 1800 亿元。
中游	中药饮片	指可直接用于调配或制剂，经过加工的中药材，

	中成药	按规定的处方和制剂工艺将其加工制成一定剂型的中药制品。
下游	销售终端	医药流通企业、医疗机构、药店和线上电商。

数据来源：国融证券

3.3.2 产业细分：

中国中药市场大致可分为中成药、中药饮片和中药保健品三类。中成药以中药为主要成分，产品有药丸、胶囊、片剂、粉剂、口服液、汤剂等多种形式。中成药广泛用于治疗和缓解心脑血管、消化系统、胃肠及妇科疾病的不适症状。传统中药通过加工动物组织中而成的，药物的有效成分是传统草药、植物提取物和动物组织。

根据现代中药集团招股说明书披露数据，按生产销售额算，2021 年中成药占中国传统中医药市场的比例为 63.4%，传统中药饮片占比为 14.7%，中药保健品占比为 21.3%。

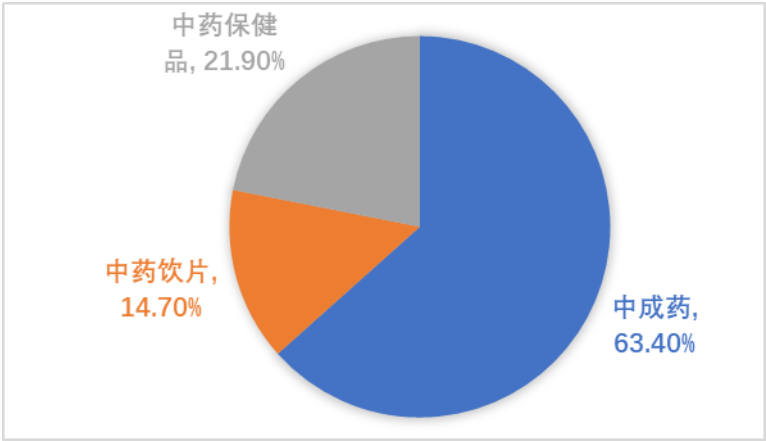


图 12 2021 年中国中药产业细分规模

四、量化选股

4.1 多因子选股模型

4.1.1 多因子选股模型基本介绍

1.定义

多因素选股模型是定量选股策略的重要组成部分。所谓的多因素选股是使用历史数据来选择一些因素可能影响股票价格波动的购买和持有股票在一段时间内，通过分析相关数据，并最终确定哪些因素可以显著影响股票价格。之后，我们可以利用这些因素进行投资。

2.因子种类

因子的来源主要有公司、外部环境、市场表现三个方面：

（1）公司层面因子：它来自于公司的微观结构，与公司的生产经营密切相关。它一般来源于公司的财务指标，反映了公司的盈利能力、经营状况、债务状况和成长状况。主要因子分类如下表 3：

表 4 公司层面因子分类

因子分类	说明
价值类因子	PE、PB
成长类因子	ROE、净利润增长率
规模类因子	净利润、营业收入
情绪类因子	预期利润增长率
质量类因子	速动比率、应收账款周转率

（2）外部环境因子：外部环境对一个行业或企业至关重要，如政治和法律、宏观经济、社会习俗和技术发展等。重要且容易量化的外部环境因素主要是宏观环境因素，如经济增长率、利率等;还包括中观行业层面数据，包括市场占有率等。

（3）市场表现因子：市场表现因素主要反映股票在交易过程中的价格和交易量。这些因素主要包括成交量、资金流、换手率等等。

4.1.2 多因子选股模型流程图

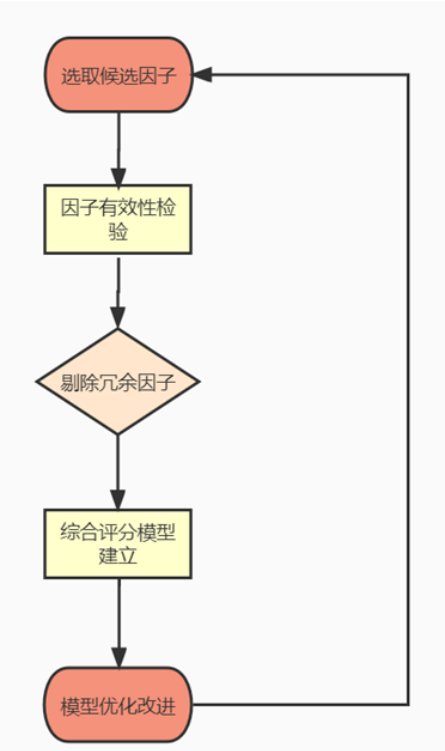


图 13 多因子选股模型流程图

4.2 XGBoost 算法

4.2.1 XGBoost 算法基本介绍

XGBoost 算法是一种提升算法。在传统的多因素选股模型的基础上，利用 XGBoost 算法的随机选择特性，在筛选股票因素的过程中判断股票因素的效果，得到因素分类结果。其原理是将原始数据集划分为多个的子数据集，每个子数据集随机分配预测，然后计算弱分类根据一定重量的结果来预测最终结果。

XGBoost 算法目标函数如下：

$$\begin{aligned} \text{Obj}=\sum_{i=1}^{n} l\left(\hat{y}_i,\ y_i\right)+\sum_{k=1}^{K} \Omega\left(f_k\right) \\ \Omega(f)=\gamma T+\frac{1}{2}\lambda \parallel w \parallel^2 \end{aligned} \tag{1}$$

4.2.2 XGBoost 算法流程图

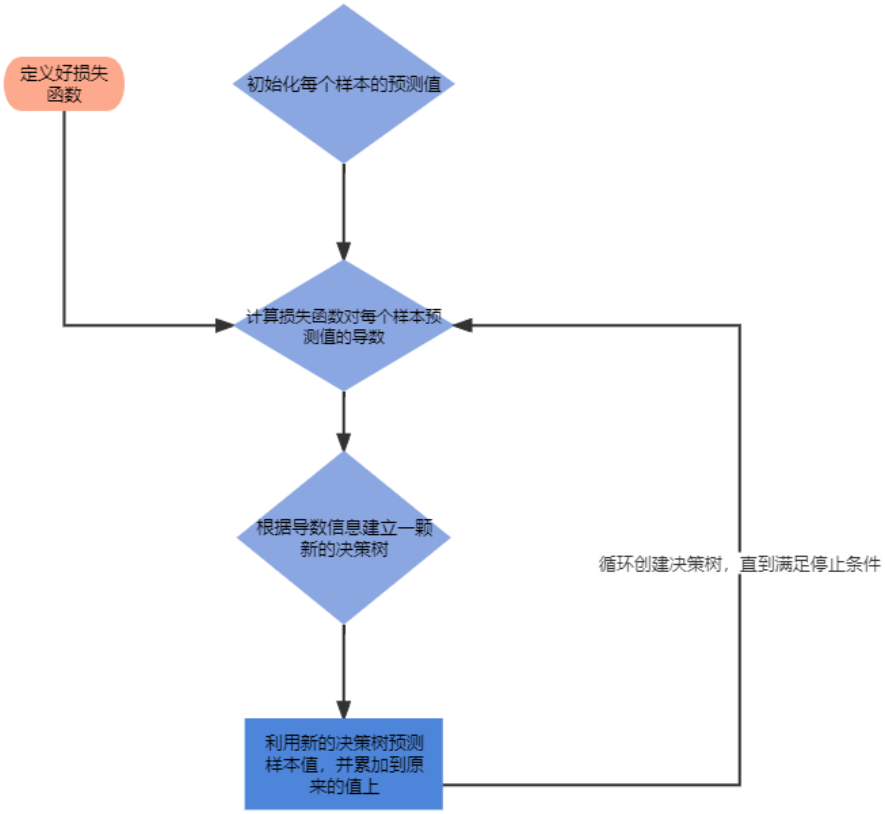


图 14 XGBoost 算法流程图

4.2.3 XGBoost 算法与传统多因子模型比较

基于 XGBoost 算法的多因子选股模型与传统多因子选股模型的比较如下：

表 5 XGBoost 算法与传统多因子模型比较

	基于 XGBoost 算法的多因子选股模型	传统多因子选股模型
优点	1. 采用量化方法，从市场中筛选出具有高成长性、高投资价值的股票 2. 有效的提高多因子表现，收益比等权多，回撤更小 3. 随机选取因子，不用进行因子筛选和检验	1. 综合考虑多层面因素，得出投资组合 2. 收益较稳定，灵活性较强
缺点	1. 初始代码在 Python、R 等机器学习语言上安装异常麻烦 2. 算法较为新颖，可供研究的材料较少	1. 并不是因子池中的每一个因子都会对股票的收益有影响，需要检测因子有效性 2. 部分因子的影响效果一样（冗余因子），需要剔除

五、基于 XGBoost 的多因子选股模型构建

5.1 方案设计框架

本文利用 XGBoost 算法优化传统的多因子选股模型，改善传统多因子选股模型中对因子重要性的判别方法。方案设计框架如下：首先从三个方面构建因子池，并进行数据预处理，接着对 XGBoost 多因子选股模型进行算法优化，得出因子重要性排序，以此选出具有高成长性、高投资价值的股票投资组合；最后，进行模型和投资组合的数据回测效果评估。具体如下图所示：

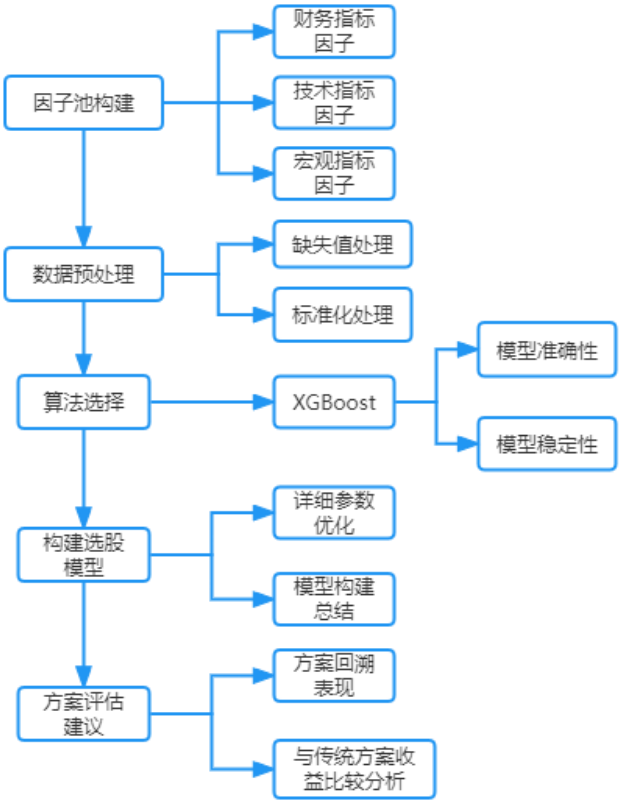


图 15 方案设计技术框架

5.2 因子池构建与数据预处理

5.2.1. 构建因子池

从世界各国的股市发展来看，股市在某种程度上也是反映了中国经济状况好坏的一个重要指标，尽管这一客观规律在中国 A 股市场并没有太大的体现，但市场环境、基本面、政策、人才等四个方面确实会对股市产生影响。在股票市场中，价格是由供求两个因素所决定的，即资本流动与股票数量的匹配。商品本身的价值决定

了商品的价格，而股票的价格又是由其所代表的公司的经营状况来决定的。由于国家的法规、法规、行业的禁止、扶持等都会对公司的经营产生一定的影响，所以国家的政策也会对股票价格产生一定的影响。

针对这几方面对股票价格有重大影响的因素，在当前条件下找到如下可以用到模板中的因子。

表 6 财务、技术等部分因子

证监会行业大类代码_Csrciccd2	每股未分配利润_UndivprfPS	销售毛利率_Gincmrt
稀释每股收益_DilutEPS	每股留存收益_RetearPS	销售成本率_Salcostrt
每股收益_EPS	净资产收益率_ROE	销售期间费用率_Pdcostrt
每股净资产_NAPS	资产报酬率_ROAEBIT	产权比率_Dbequrt
每股营业收入_MincmPS	资产净利率_ROA	三个月动量
每股营业利润_OpeprfPS	投入资本回报率_ROIC	1 个月换手率
每股息税前利润_EBITPS	销售净利率_Netprfrt	上影线长度
每股资本公积金_CapsurfdPS	股息率	机构投资评级情况
每股盈余公积金_SurrefdPS	下影线长度	近三个月资金流入流出占比
每股公积金_AccumfdPS	3 个月换手率	近一个月资金流入流出占比
每股经营活动现金流量_OpeCFPS		

来源：国泰安数据库、wind

表 7 宏观、债券、投资等因子数据

10 年中债国债到期收益率	M1	M2
5 年中债国债到期收益率	三个月存款利率	宏观经济景气指数:先行指数
1 年中债国债到期收益率	投资者信心指数:总指数	投资者信心指数:买入
投资者信心指数:股票估值	6 个月短期贷款利率	6 月定期存款利率
住宅商品房待售面积	定期存款利率:1 年(整存整取)	一年期贷款利率
商品房销售面积累计值	现房商品房销售面积累计值	住宅房屋竣工面积累计值

来源：国泰安数据库、wind

5.2.2 数据预处理

本文按市盈率排序选取了中药行业的 30 只股票，时间跨度为 2008 -2019 年，

从财务、技术、宏观、债券、投资层面的因子数据，每个层面选取 10 个因子训练模型。本文所采用的数据主要来源于锐思金融经济数据库、国泰安数据库、WIND 金融资讯数据库等。

（1）缺失值处理

本文中的因子数据范围广泛且数量众多。因此，在某些情况下，数据中可能缺少某些因素。如果缺失值过多，则放弃该股票，如果缺失值较少，则使用插值方法完成。

首先，从原始数据集中确定因变量和自变量，将缺失值前后的 5 个数据取出，并将 10 个数据编译成一组。然后取拉格朗日多项式公式：

$$ln(x) = \sum_{i=0}^{n} l_i(x) y_i \tag{2}$$

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_i}{x_i - x_j} \tag{3}$$

在公式中，x 表示缺失值的位置，另一个 x 表示组中的数据，L(X)是缺失值的近似值。

（2）异常值处理

一个异常值被定义为一个数据集中的四分之三或四分之三的位置。在处理异常值时，不能盲目地将其删除，否则会对数据的整体结构产生影响，从而对以后的建模效果产生一定的影响。通常要结合实际问题，进行具体的分析。

（3）数据标准化

通过数据标准化统一数据量纲，常用方法主要有以下三种：

1. 最小-最大规范化

$$x^* = \frac{x - min}{max - min} \tag{5}$$

其中 max 为所选取的样本中的最大值，min 为所选取的样本中的最小值。

2.零-均值标准化

$$x^* = \frac{x - \mu}{\sigma} \tag{6}$$

其中σ为原始数据的标准差，μ是原始数据的均值，这是当前用的最多的数据标准化方法。

3.小数定标规范化

通过移动数值的小数位数，将值映射到[-1,1]之间。转化公式：

$$x^* = \frac{x}{10^k} \tag{7}$$

（4）数据转换

通过以上处理，对数据丢失进行了插补，并对其进行了处理，使其它可以很好地构建一个通用的统计学习模型。但本文使用的模型是 XGBoost 算法因其本身特点，需要将数据转换为分类数据，包括等距分类和等量分类两种方法，具体举例如下：

表 8 数据转换

每股收益	等距分类	等量分类
0.1428	（0.1,0.2）	（0.1,0.18）
0.2466	（0.2,0.3）	（0.18,0.4）
0.3742	（0.3,0.4）	（0.18,0.4）
0.4256	（0.4,0.5）	(0.4,0.46)

5.3 XGBoost 算法参数优化详解及因子筛选

为了提高 XGBoost 模型的泛化能力，优化模型参数十分重要，本文主要从以下方面进行模型优化：

首先，为了更为精细化的优化参数，我们可以使用较高的学习速率，在一定的区间内穷举参数，选择模型泛化的参数组合；其次，在上一步给定的学习速率基础下，优化特定决策树参数；最后，降低学习速率，确定理想参数。

① 定义 learning rate 和参数优化的估计器的数量

max_depth = 5 :起始值在 4-6 之间。

min_child_weight = 1:根据股票数据的平衡性进行调整。

gamma = 0:开始时，选择较小的数值，后续继续进行调整的。

subsample,colsample_bytree = 0.8: 起始值在 0.5-0.9 之间。

scale_pos_weight = 1。

输出结果如下：

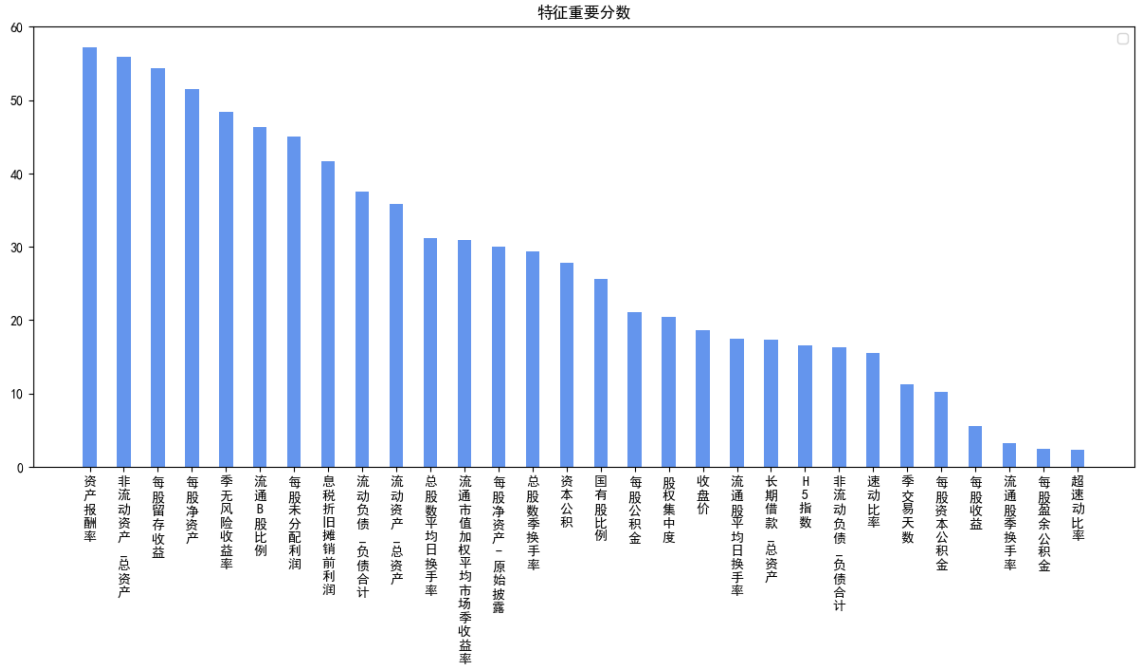


图 16 因子重要性输出 1

② 参数调优

对 max_depth 和 min_weight 的优化，选择多组数值，最后得出 max_depth 值等于 5，min_child_weight 值等于 5 时，模型 cv 值实现提高，达到了 0.9154，同时模型均值和方差较为理想。

接下来进行 gamma 的参数寻优，和上面相似，取步长为 0.5，在 0 到 5 的范围内继续优化 gamma 参数。

表 9 gamma 参数优化结果

Mean	Std	Gamma
0.91498	0.00438	0
0.91537	0.00436	0.1
0.91642	0.00435	0.2
0.91537	0.00439	0.3
0.91534	0.00465	0.4

从上表结果可知 gamma 的值取 0.2 最为合适，优化后因子选取结果如下：

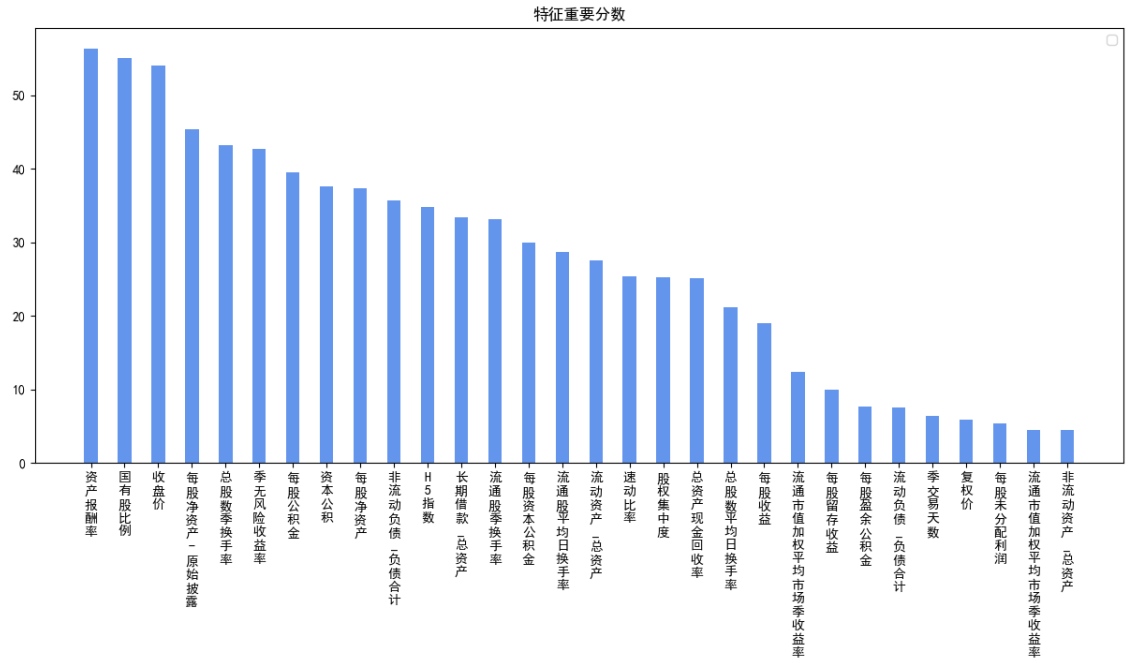


图 17 因子重要性输出 2

③ 降低学习速率

最后，我们使用较低的学习速率，并选择更多的决策树。我们可以用 XGBoost 中的理想的树数量和分值（cv）函数来进行这一步工作。

表 10 cv 参数优化结果

Tree_number	Train_AUC	Test_AUC
483	0.965726	0.925595
484	0.976021	0.9553
Accuracy : 0.8771 AUC Score : 0.96689		

最后，得出优化后的基于 XGBoost 算法选股模型得出的因子重要性排序：

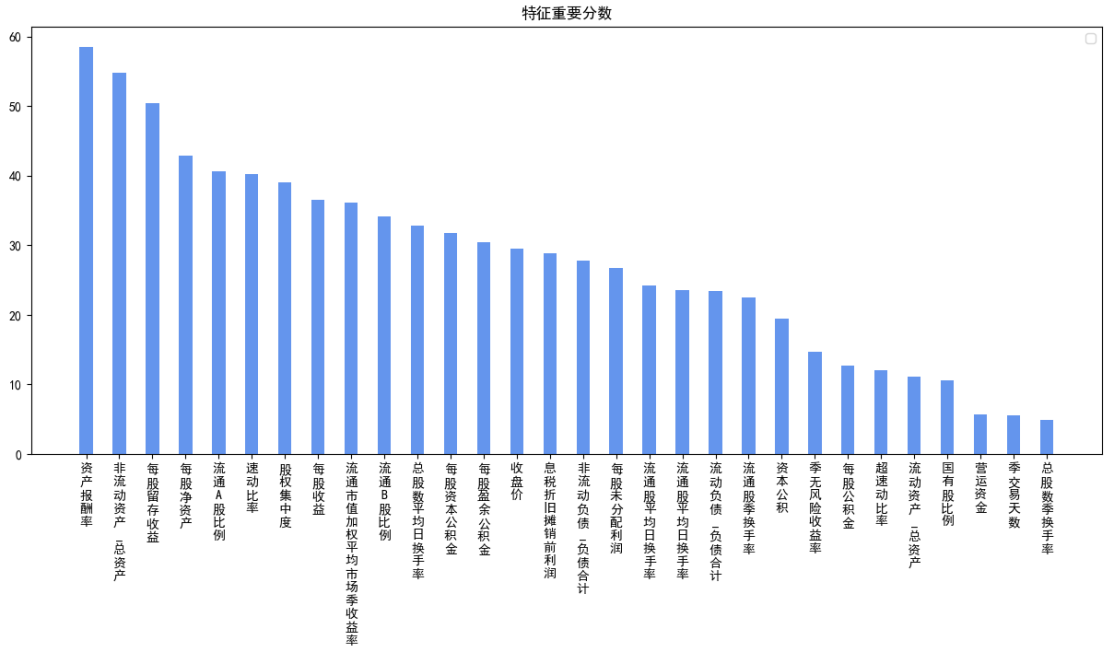


图 18 因子重要性输出最终结果

5.4 股票筛选结果

5.4.1 中期投资组合 - 两年期筛选结果

在得出优化 XGBoost 模型得出的因子重要性输出最终结果后，本文采用 2020 年 1 月 1 日到 2022 年 1 月 1 日的两年期中药行业公司披露的相关资料和有关机构对公司的评分，带入优化 XGBoost 模型进行中期股票池的筛选，最终得出我们的股票评分表如下：

表 11 中期-最终得分表

排名	股票代码	股票名称	综合得分
1	600436	片仔癀	86.2342
2	000538	云南白药	83.5420
3	600332	白云山	75.3800
4	600479	千金药业	69.3513
5	300039	上海凯宝	69.1347
6	002166	莱茵生物	62.8025
7	600085	同仁堂	60.0200
8	002566	益盛药业	58.6025
9	603858	步长制药	55.8165

10	603998	方盛制药	53.8646

根据 XGBoost 模型的输出结果，选取最终得分排名前 3 的股票，即为我们筛选出的中期股票投资组合，按照得分大小进行股票池配置，结果如下：

表 12 中期股票投资组合构建

排名	股票代码	股票名称	股票仓位
1	600436	片仔癀	35.18%
2	000538	云南白药	34.08%
3	600332	白云山	30.75%

5.4.1 短期投资组合 - 一年期筛选结果

在得出中期投资组合建议同时，继续采用采用 2020 年 4 月 1 日到 2022 年 4 月 1 日的一年期中药行业上市公司股票相关因子数据，带入优化 XGBoost 模型进行股票池的筛选，得出短期投资组合评分结果如下：

表 13 短期-最终得分表

排名	股票代码	股票名称	综合得分
1	600479	千金药业	85.1690
2	600436	片仔癀	82.8958
3	000538	云南白药	71.3169
4	002603	以岭药业	69.0380
5	600750	江中药业	67.1151

选取最终得分排名前 3 的股票，即为我们筛选出短期的股票投资组合。

表 14 短期股票投资组合构建

排名	股票代码	股票名称	股票仓位
1	600479	千金药业	35.58%
2	600436	片仔癀	34.63%
3	000538	云南白药	29.79%

六、技术分析

6.1BOLL 指标

图 19 云南白药 BOLL 图

BOLL 指标称为布林带（Bollinger Bands），作为一个研究价格趋势的分析方法而被人们广泛采纳。发明者约翰布林从统计学中的标准差原理出发，从而设计出这种实用又简单的技术指标。BOLL 指标可以研究股票股价在市场波动变化情况，从而推测股价的未来走势的一种重要技术分析工具。

本图显示的是云南白药 2021 年 11 月 22 日到 2022 年 5 月 13 日日 K 线的 16 BOLL 指标分析图，云南白药股票在 1 月 14 日前较大时间都是处于中轨之上，这同时说明该股票是强势股，当大部分处于中轨线以下时就处于弱势股，不建议选取操作。在 3 月 4 日时，该股票上下轨变窄，由原先扁平趋势，端口逐渐收缩变窄，这意味后续会有突变，最终在后续出现骤降的变化，此阶段随大盘持续下跌，是一个持币观望的时点。在 3 月 3 日价格经过一波下跌之后，上下轨分别向上向下运动，形成一个“双开”的喇叭口，预示此时股票行情下行趋势会停止，随后云南白药的股价变动趋于平缓。

当 BOLL 的上下轨开口急剧缩小时，即出现在 4 月 6 日的变化，这就预示此时行情目前在短期内会有较大下跌趋势，因此需要注意在行情保持小幅度上涨之后一

定会出现回落的趋势，在 4 月 10 日通过上图可以看到股票整体开始回落，价格在之后的一段时间内出现了持续下滑，而当在股价跌至 BOLL 中轨线时，盘面趋向于盘整状态，之后股价最终跌破了中轨，这一系列表现说明行情即将持续为下降通道，而且价格如果一直下跌，必定形成一波下跌趋势。如果价格跌破中轨，并且进入了下降通道，即中轨与下轨之间，这就表示市场为下降状态，此时价格沿着下轨道运行表示下跌还会延续。该股在经过近 50%的大幅下跌后，于 4 月 25 日跌至最低点 49.54 元。在图中已经出现了股价跌破布林线下轨的走势，在行情非常低迷的时候，股价持续大幅下跌，在 4 月 25 日跌至最低点，跌破布林线的下轨。但在 4 月 26 日，BOLL 的上下轨的开口增大，此时预示该股随后有上涨趋势，后面连续 3 日股价向 BOLL 线中线靠拢，这个时候就可以在低价时买入并当它升到一定高价时抛出。但是要注意短线买入的方法只能在使用跌破下轨快速回升买入时使用。

6.2 形态理论

形态分析是一种预测股票价格未来走势的方法，它主要依据价格图表中过去一段时间价格轨迹。它在 K 线基础上所组成，形态图中股价会在多空双方取得均衡的位置上下来回波动。

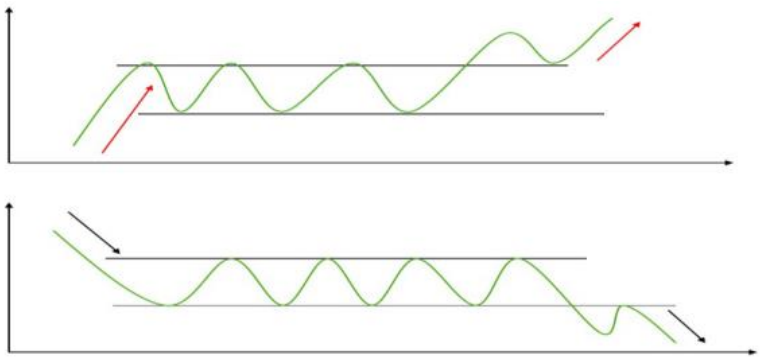


图 20 矩形整理的 2 种形态

6.2.1 矩形整理

矩形整理体现为股价在两条平行而且横向的直线之间来回上下摆动，在价格上升到上半部分位置时，又回落回到下半部分，这样来回摆动。若这笔投资的期限为中长期，股票的上升趋势保持稳定为发生改变的时候，且有着矩形整理形态存在时，

较多的持有股票是此时的建议投资策略;反之，需要等待更低的买入点。

数据来源：同花顺

图 21 白云山矩形整理

图中展示的是白云山 2020 年 6 月到 2022 年 5 月 15 日的周 K 线图，其中在 2022 年 6 月 15 日到 2022 年 5 月 15 日呈现出矩形整理的形态，该股票价格在 2 月中旬开始出现小幅度上升的趋势，并在其后一段时间内一直保持较小的涨跌幅度，此后在该股的上升趋势中，股票价格将会停滞在某个高点，随之的是遇到较大的空头力量使其向下，但当股价下落至一定程度的低点时候，又会上升至之前一次的高点，并会再一次遭遇阻力而回落，此后将会得到一条在低点处的支撑线。价格的波动均在上下两条平行线内，最高值不超过上阻力线，也不会下落突破下支撑线，在上下段直线之间摆动，但当到达 4 月底之后一直呈现下降趋势，并且这种趋势一直延伸下去，最终跌破下线端，矩形突破下边线为卖出信号，当突破下支撑线时，同时市场都发出卖出的信号，卖出股票，情绪低迷，导致股价进一步下跌，在 4 月 29 日最终跌到最低点，后续根据矩形整理形态，会出现一定程度回升。

6.2.2 突破形态之双重顶

双重顶即人们所熟知的 M 头，双重顶顾名思义是共出现两个顶，也就是存在两

个的“顶点”，并且两点高度相当。双重顶作为反转形态，其形成的原因是股票上涨的趋势存在与其相当的空头力量。当股价上升形成第一个顶部，此时向下的阻力较大，使得股价下落，在落在颈线位置的时候存在着一个支撑的力量，股价随后呈上升走势。同样的，但股价呈上升走势继续上涨至形成第二个顶部，此时又有空头的力量使股价上升受阻，并且相对于原先的颈线位置，股价未能得到支撑，继而股价跌破颈线，保持一定的下滑趋势，理论上下跌幅度的大小不小于顶部至颈线位的高度。

数据来源：同花顺

图 22 片仔癀双重顶形态

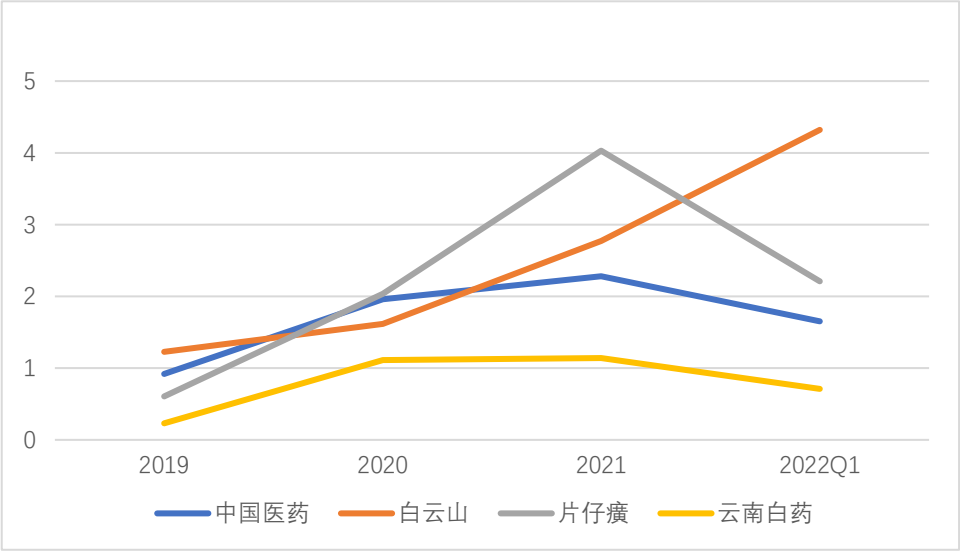
这是片仔癀(600436) 2020 年 6 月至 2022 年 5 月 13 日期间周 K 线图，如图中所示，该股在 2022 年 3 月中旬之前的这段时间内，出现了一个标准的双重顶形态即走出 M 型的形态。此股近期的走势，于 3 月初至今是呈平缓走态，拟形成第二个 M 型的形态。从此股随后的走势中可以看到，双重顶形态上升是非常可信的“空转多”信号，一旦出现此形态可考虑进场，以获取收益。

七、财务分析

针对前文选择的三只股票，为了更全面的检验这些股票的投资价值和未来的成长能力，需要进行财务分析。本文我们选择了从盈利能力、成长能力、偿债能力以及营运能力四个维度对三只股票进行分析，选取中国医药股票为行业均值基准，与三只股票进行对比，为后续的投资决策提供更加可靠全面的信息。

7.1 盈利能力

7.1.1 每股盈利

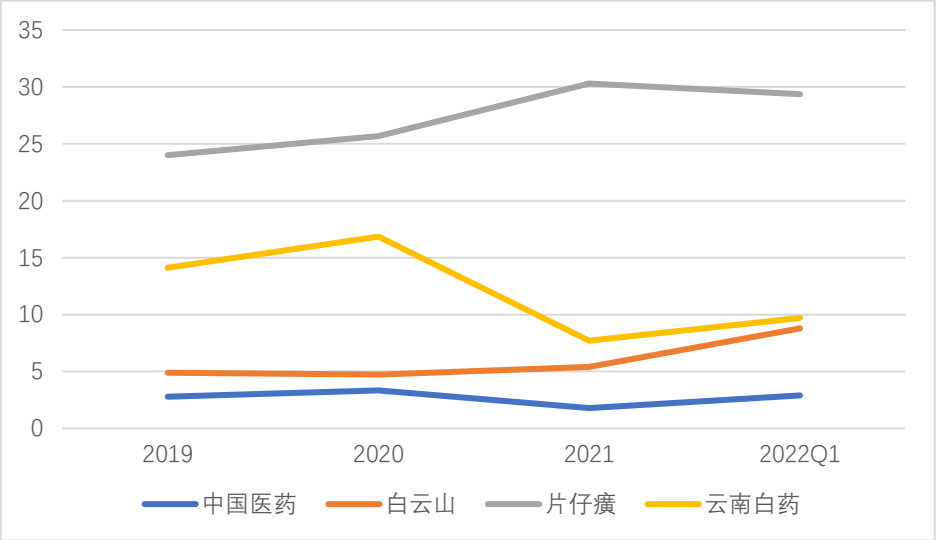


数据来源：东方财富网

图 23 每股收益对比图

由上图可以看出，白云山和片仔癀的每股收益在中国医药的每股收益值上下浮动。同时在 2020 和 2021 年，每一支股票的每股收益都大于 1，这表明当时股票的盈利能力较强，主要原因应该是受到疫情影响，国民对中药的需求迅猛增长。2022 年一季度，除了白云山，其余都呈下降状态，主要是受到外界大环境影响。但总体来看，白云山、片仔癀和云南白药都有较大的发展潜能，具有相对较高的投资价值。

7.1.2 营业利润率

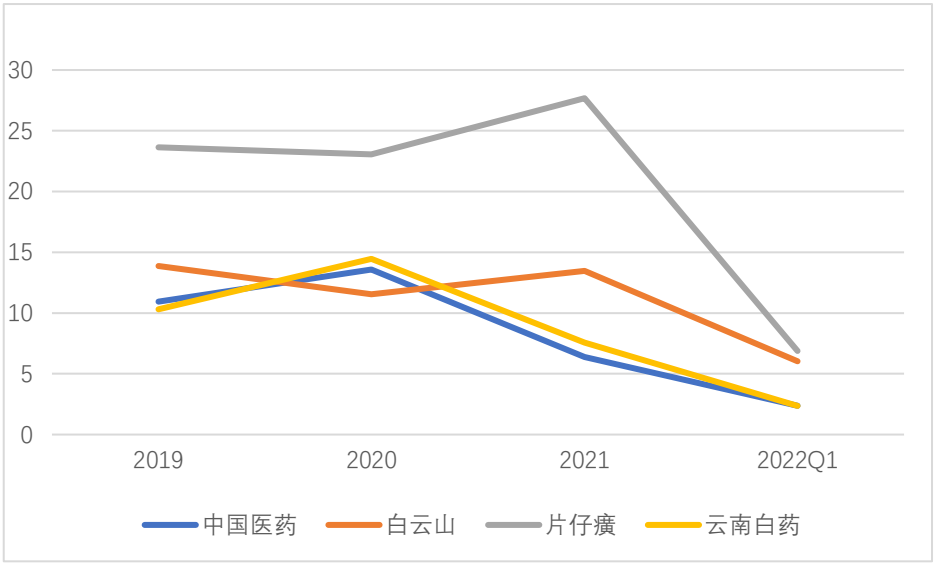


数据来源：东方财富网

图 24 营业利润率对比图

由上图可以看出，3 支股票的营业利润率都高于中国医药，其中片仔癀的营业利润率最高，说明片仔癀的业务能力很强的同时创造利润的能力也很强，经营风险相对较小。另外两支股票的营业利润率虽低于片仔癀，但高于行业平均值，有不错的盈利能力。

7.1.3 净资产收益率



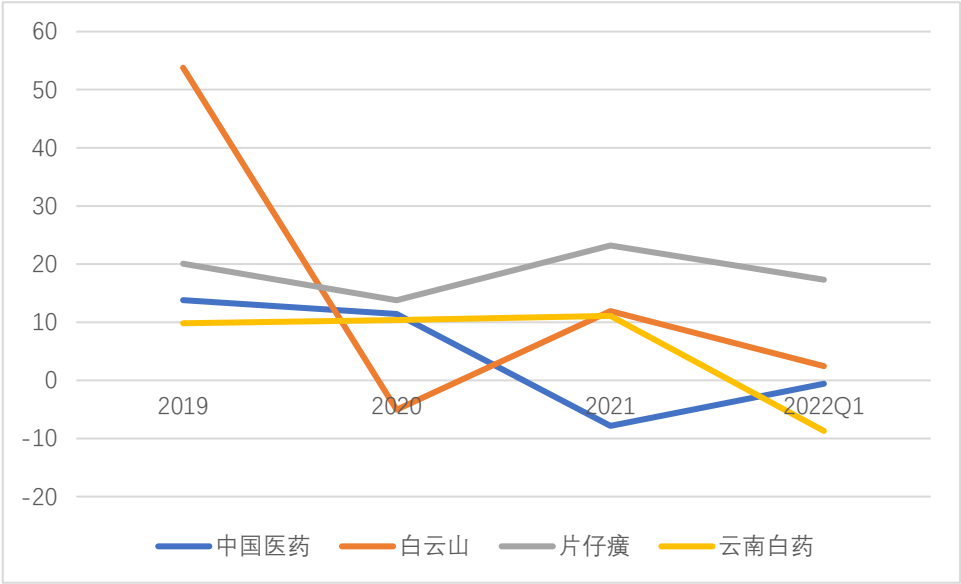
数据来源：东方财富网

图 25 净资产利润率对比图

由上图可以看出，三支股票中片仔癀的净资产收益率最高，尤其是 2021 年，净资产收益率高达 27.68%，说明片仔癀对股东投入资本的利用效率较高，即运用公司自有资本的效率较高。2022 年，片仔癀加大研发投入，预计到年底利润率大幅上升。而云南白药和白云山在运用自有资本效率方面相较于片仔癀较低。

7.2 成长能力

7.2.1 营业收入增长率

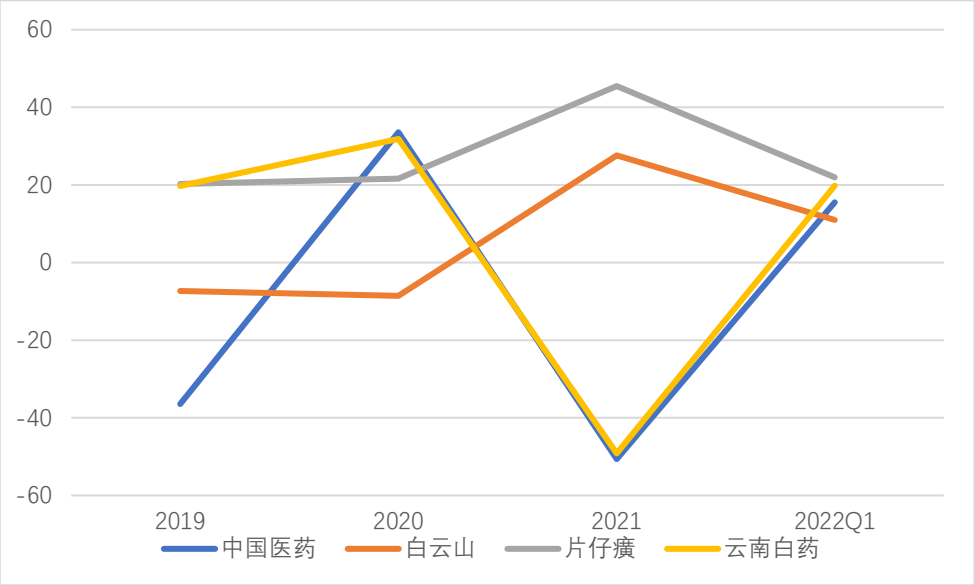


数据来源：东方财富网

图 26 营业增长率对比图

由上图可以看出，三年间白云山的营业增长率均有较大波动，2020 年受到疫情影响，无法正常销售，所以出现大幅下降的情形，但总的来说，三只股票的营业收入增长率都随着疫情的好转恢复增长。说明这些公司的产品在市场中都有一定的份额。其中，2021 年片仔癀的营业收入增长率超过了 20%，说明公司正处于一个快速发展期，未来将会继续保持较好的增长势头，市场前景较好。而云南白药和白云山的营业收入增长率在 10%左右，说明两家公司已经进入稳定期。

7.2.2 净利润增长率



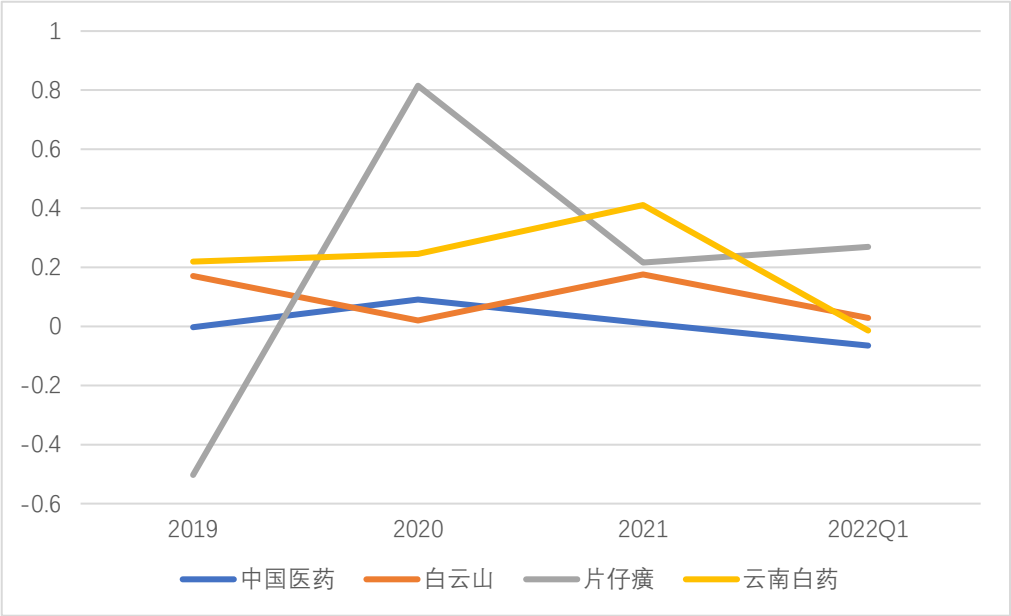
数据来源：东方财富网

图 27 净利润增长率对比图

由上图可以看出，2021 年片仔癀和白云山都有一个较高的净利润增长率，说明企业发展情况良好，产品销售增长快，市场竞争力强。而云南白药的净利润增长率为负，主要系股票投资等公允价值变动损益影响，随着疫情情况好转及公司减持股票，云南白药后续经营状况有望恢复增长。虽然三只股票的净利润增长率波动较大，但目前基本稳定在 20%上下。

7.3 偿债能力

7.3.1 现金流量比率

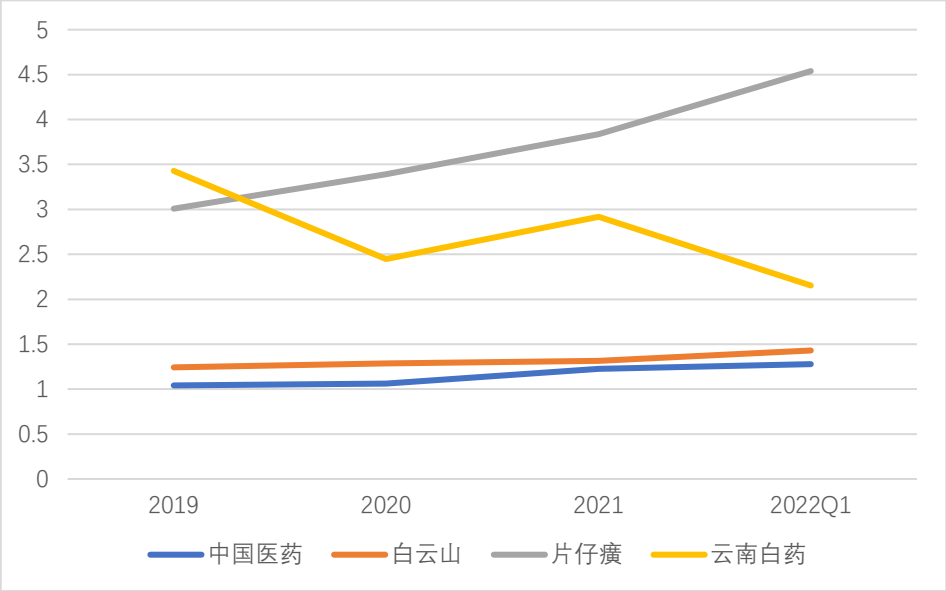


数据来源：东方财富网

图 28 偿债能力对比图

由上图可以看出，从 2019-2021 三年以及 2022 年第一季度所披露的数据上来分析，三支股票的现金流量比率都小于 1，这也说明了企业应该需要采取一定的筹资措施从而满足企业日常的基本需要，增强其还债能力，预计这将会使市场上的投资者更加充满信心。

7.3.2 速动比率



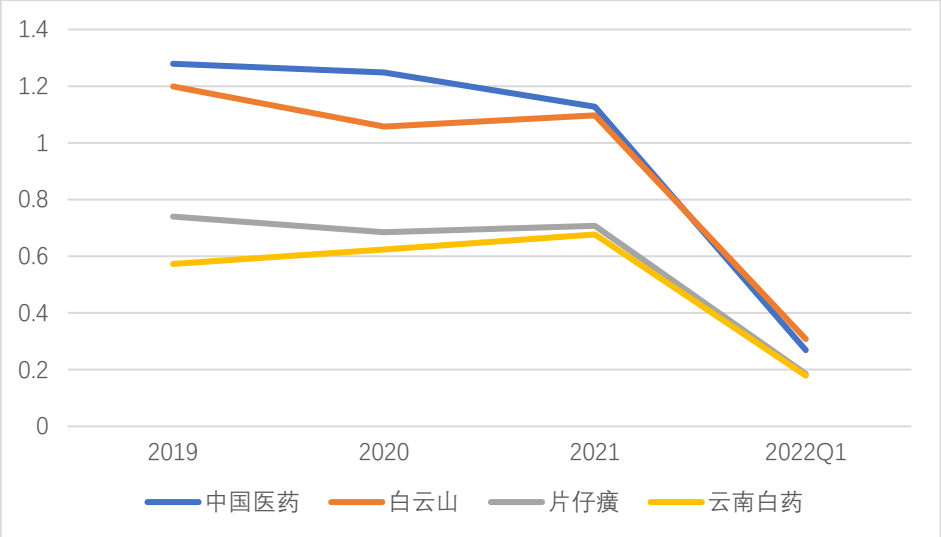
数据来源：东方财富网

图 29 速动比例对比图

通过上图可以看出来，在速动比率这项指标中，白云山是最接近于 1 的，是较为接近行业均值，尤其看来，虽然白云山企业自身的偿债能力并不强大，但这个数值也在一定程度说明了企业本身是能适宜与行业一同发展的节奏的。这表明企业在承担每一份流动的负债计入资产负债表的同时，此时的企业相对于拥有了接近一份额的流动资产，并且该份流动资产是具备易于变现的特点，该份资产作为储备补偿，这也使得在短期内，企业的偿债能力可值得较高的信任，且并不会给企业增加太多的投资机会成本。相比之下，片仔癀和云南白药的速动比率远大于 1，企业在速动储备资产占用太多的资金，大大减少了企业获取收益的能力，很大程度上会成为企业快速发展的一个阻碍。

7.4 营运能力

7.4.1 总资产周转率

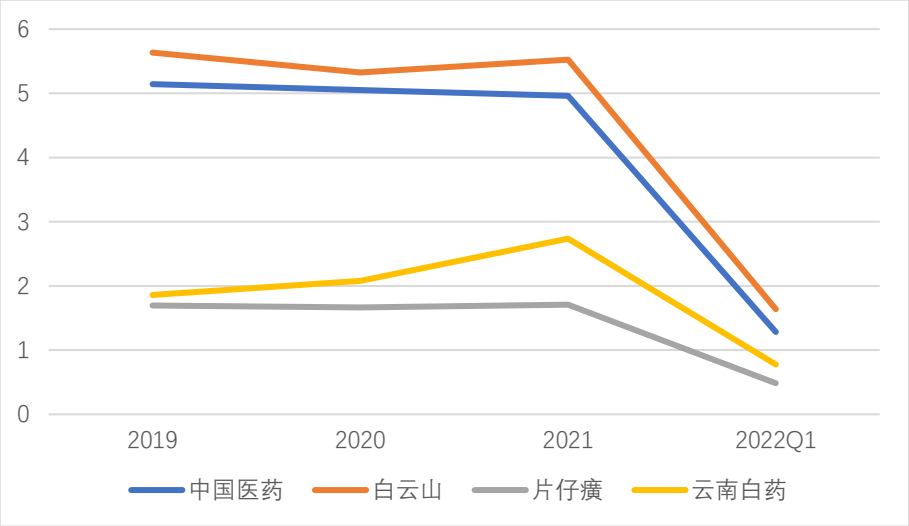


数据来源：东方财富网

图 30 总资产周转率对比图

由上表可以看出，2022 年前三支股票的总资产周转率都大于 50%，说明在这三家企业的生产经营期间，从前期的资产投入环节，再到进行产品的产出，其经历的时长较短，企业的投入-产出速度是较为快速的，以及对于产品链的前端分析，在前端当中，产品的销售表现是较为良好，这也进一步地说明资源的配置是合理的，资产的利用效率、增值能力是较高的，企业也拥有着较好的管理层级，最大限度地发挥着优秀的企业家才能。白云山的总资产周转率与中国医药最为接近，但对于 2022 年第一季度的表现，中药行业整体的总资产周转率是下降的态势演变，预期后期会有一定上涨的表现。

7.4.2 存货周转率

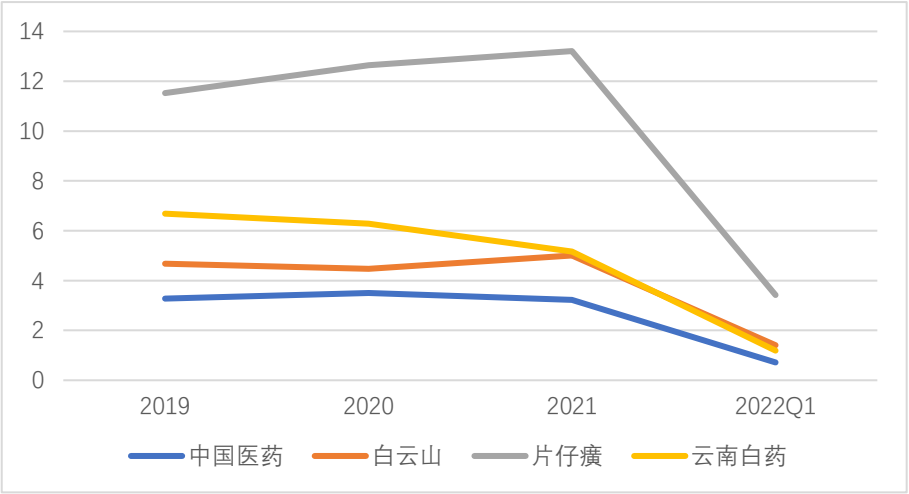


数据来源：东方财富网

图 31 存货周转率对比图

2014-2021 年中国中药行业的存货周转率一直处于震荡形态，区间在 2.4-4.4。由上图可以看出，三支股票的存货周转率都在这区间左右，围绕着中国医药平均水平上下波动。说明企业整体的存货变现能力较为良好。

7.4.3 应收账款周转率



数据来源：东方财富网

图 32 应收账款周转率对比图

由上图可以得出，三支股票的应收账款周转率都高于中国医药，表明企业在回收账款和资金管理上处于较高水平，拥有较好的发展前景。

八、历史回测与绩效评价

8.1 模型准确检验

对于进行算法正确性我们常用的指标在上文对算法进行选择时候已经进行过相应的介绍，因此在对于模型的准确率进行评价时，我们采用和上文相同的指标，即结合 ROC 曲线和 AUC 值来评估模型的优劣。

我们前期进行了大量的参数调整以及优化，最终挑选出最合适的参数，构建了相应的训练和测试模型，从 ROC 曲线图可以看出，训练模型准确度为 86.97%，AUC 分数为 0.79，ROC 的曲线比 y=x 直线要高出不少可以说明我们的模型结果好于随机的猜测，在本模型中我们只考虑正例在什么情况下被预测为正例的准确性是最大的，因此在 ROC 曲线图上，越接近 1 的地方，就说明整体的准确度越高，所以我们在每次回测时都会采用不用的阈值。

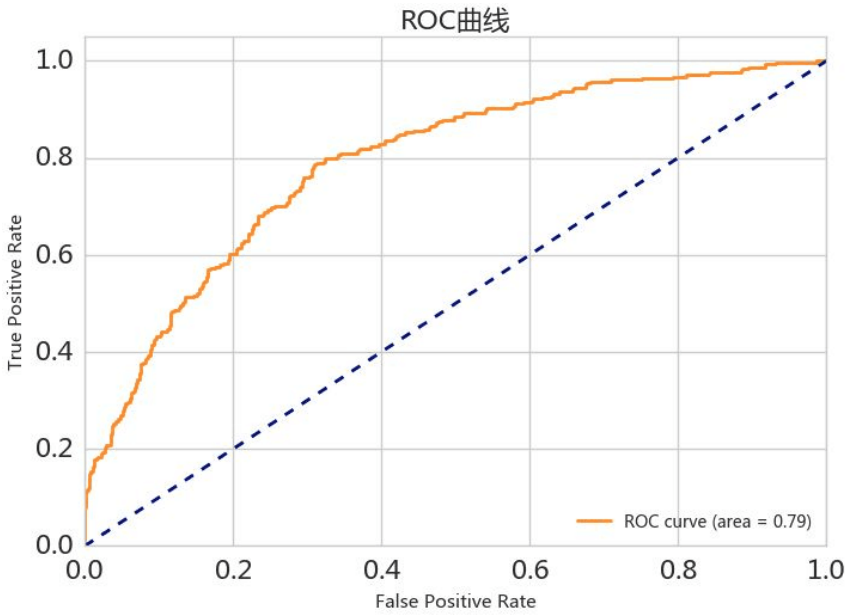


图 33 ROC 曲线模型图

8.2 历史回测绩效评价

为了检验本模型是否能以较低的风险获取较高的收益率，将对模型进行回测，所选取的历史数据时期为 2016-2018 两年期的 A 股市场数据，并且将选取沪深 300 指数为基准数据，通过进行比较的方法，进而展示我们模型的可靠性。因此，基于风险和收益两个方面，所选取的指标为收益率指标与风险度量指标。

收益率指标我们采用的是年化复合收益率，年化收益率可较为直观地比较两种产品直接收益的高低。

风险度量指标主要有贝塔系数、夏普比率、最大回撤、信息比率等。

图 34 累计收益率比较

基于以上评价准则，文采用 2016 年 1 月 1 日一季度到 2018 年 1 月 1 日的两年期历史数据回测验证该模型的有效性。选取沪深 300 为基准收益率进行比较，得出结果如下图所示：

由下表可知，在回测中，基于 XGBoost 的多因子选股模型选出的股票组合的年化复合收益率高达 20.09%，所选取的基准指标“沪深 300”年收益率为 4.04%，本模型高出 16.05%。从图 30 可以看出，该模型所选出的股票在 88% 的时间所取得的收益是高于沪深 300 指数，并且有一半的投资天数具有正的收益率，最后年化达到 42.95%，该收益率水平是远超基准沪深 300 指数收益率。综上所述，该模型是能在较低的风险下保持较为高水平的收益率。

表 13 总资产周转率

年化收益率	20.09%	基准年化收益率	4.04%
α	0.1614	β	0.83
夏普比率	0.68	波动率	0.2448
信息比率	0.79	最大回撤	0.2055
年化换手率	0.02		

九、 风险控制

9.1 定期修正策略来预防风险

“基于 XGBoost 的多因子选股模型”在多因子策略的基础上进行了一定的完善。它可以根据给定的历史数据，对其进行分析，输出“因子特征重要性”，并根据“因子特征重要性”来筛选股票。他是一个动态的过程。

本方法按照一个月的周期，根据市场的变动情况，重新筛选因子的“特征重要性”，并对其进行重新排序，根据新的排序进行投资组合的重新配置，来进行股票的重新选择，以此来解决投资策略中存在的一些风险，例如政策的改变的风险，企业的资金流失、人才变动等风险。

例如本方案选择 2020 年 1 月 1 日到 2022 年 1 月 1 日的两年期中药行业公司披露的相关资料和有关机构对公司的评分，选出的是片仔癀、云南白药、白云山三只股票；而根据 2021 年 4 月 1 日到 2022 年 4 月 1 日一年期中药行业公司披露的相关资料和有关机构对公司的评分的数据，选出的股票为千金药业、片仔癀、云南白药。

9.2 模型自身存在的优势可以减少风险

“基于 XGBoost 的多因子选股模型”相较于其他方法来说拥有很多的优势，这些优势也让其可以更好的控制风险。

9.2.1 更好的容忍异常值缺失值

对于异常值，由于每个特征的数值都只用在大小比较，所以 XGBoost 算法可以更好的容忍异常值的存在。相比较于其他投资方法，运用了 XGBoost 算法，可以很好的解决缺失值、异常值的存在，降低一定的非系统风险。

9.2.2 更广的适用范围

相比于其他算法，XGBoost 算法加入了复杂度函数，因此，泛化能力更上一层楼。拥有了很好的泛化能力的 XGBoost 算法可以更好的适应新鲜样本，避免出现差错。有很好的降低风险的能力。

9.2.3 更加的稳定

相比于其他的选股方法，本文中的方法更加稳定。因为本模型综合考虑了公司层面、外部环境、市场层面三方面的 302 个因子，几乎包含了所有的因子，可以使

得本模型更加的稳定，很好的降低相应的风险。

9.2.4 减少人为随意性

基于 XGBoost 算法的多因子选股策略减少了人为定义权重的随意性，结果客观合理。因为每个因素的权重是按照自己的方差贡献率得到的，所以变量的方差越大，权重越大；反之，变量方差越小，权重越小。减少了人为的随意性就会使得该方法的风险更小。可以很好的解决非系统风险。

9.3 运用大量的因子减少风险

本方法收集了相对全面的因子数据，共使用了 307 个因子。除了大部分投资者使用的财务、红利、动量等因子，还添加了规模、估值、宏观、债券和楼市相关因子。因子越多，对市场的相关信息的利用就越完善，大量的因子的运用可以有效地降低本策略的风险。

9.4 特定的筛选方法降低风险

本方法改变了以往的因子筛选方式以及建模流程，使用边训练边筛选的方式，筛选的方法更为科学合理。

本方法利用中药行业 30 只上市公司股票在 2008 年-2019 年间的因子数据进行 XGBoost 算法的优化训练，在优化算法的同时进行因子的选取，一边优化一边筛选可以有效降低风险。

十、 主要结论与投资建议

10.1 主要结论

本文从宏观经济环境和行业环境两个基本面出发，通过量化选股、投资组合构建、风险管理以及绩效评估等角度，采用财务指标分析、技术指标分析、XGBoost 算法等方法分析得到中药行业中具有高成长能力、高投资价值的股票，并求得最优的投资组合。综合以上分析，得出以下的结论与相应建议。

结论一：中药市场主要包括中成药、传统中药饮片和中药保健品三大部分，我国中药配方颗粒和中药饮片市场规模逐年攀升，行业发展势态良好。

结论二：受全球大健康行业良性发展、国民卫生健康要求提高、国内外疫情反复等因素影响，中药行业将迎来上升期；此外，近年国家出台了一系列鼓励、支持中医药行业发展的政策文件，推动中药企业发展创新，投资政策优势明显。

结论三：通过量化筛选，我们得到了片仔癀、云南白药、白云山、千金药业等具有良好投资价值的股票，投资者可以对其进行投资组合。

结论四：投资者需要充分考虑投资中可能存在的风险因素，利用我们构建的 XGBoost 多因子选股模型，可以减少人为投资的任意性，更加稳定地实现收益和风险控制。

10.2 投资建议

我们分别利用一年期（2021.4-2022.4）和两年期（2020.1-2022.1）两段历史数据对 XGBoost 多因子选股模型进行优化，按照模型输出的因子重要性对股票因子数据进行评分，最终得出投资方案：

短期：千金药业（35.58%）、片仔癀（34.63%）、云南白药（29.79%）；

中期：片仔癀（35.18%）、云南白药（34.08%）、白云山（30.75%）。

在更长的投资周期当内，投资者需要持续更新股票因子数据，以保证多因子评分能够包含最新市场信息；投资者可以利用动态更新的股票因子数据训练 XGBoost 多因子选股模型，获取最新股票特征因子重要性评分，进而得出下一周期的投资组合配置建议。长期动态投资组合配置示意图如下。

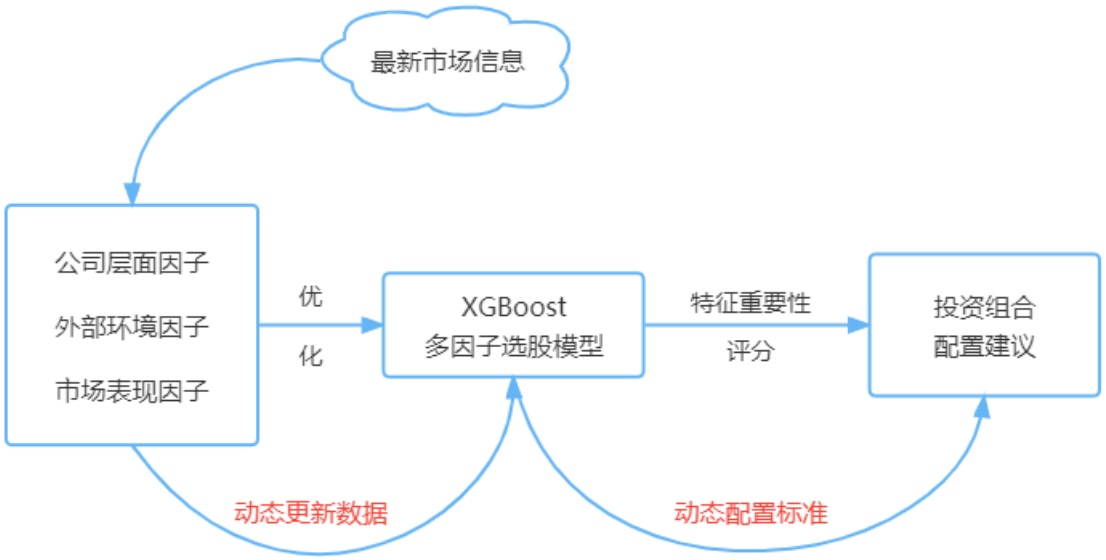


图 35 长期动态投资组合配置示意图

回顾我们量化投资策略设计的过程，我们依据投资组合的基本原理，利用编程软件实现了数据模拟与优化，模型新颖实用。同时，结合了宏观、行业、公司的多视角分析以及财务基本面分析，结构较为严谨，数据完整，得到了简明的投资策略。同时由于时间与知识水平的限制，仍存在许多欠缺。需要指出的是，本报告数据源于公开历史数据，用于模型及分析方法的展示，不构成最终操作建议。

参考文献

[1] 石磊. X 企业中药配方颗粒营销策略研究[D].云南大学,2017.

[2] 李君子. 生命周期视角下的制药企业内部控制评价与优化研究[D].浙江工商大学,2020.DOI:10.27462/d.

[3] 董婵. 基于我国 A 股市场的动态多因子选股研究[D].天津财经大学,2017.

[4] 张玉科. 量化投资分析模型的研究与应用[D].电子科技大学,2020.DOI:10.27005/d.

[5] 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.

[6] 田浩. 基于 XGBoost 的沪深 300 量化投资策略研究[D].上海师范大学,2018.

[7] 祝养豹. 基于 XGBoost 和 LightGBM 算法的多因子选股方案设计[D].南京大学,2020.DOI:10.27235/d.

附录

一、XGBoost 算法参数优化及因子筛选

```
train_data = pd.concat(train_data_b,ignore_index=True)
test_data = pd.concat(test_data_b,ignore_index=True)
xgb1 = XGBClassifier(
learning_rate =0.01,  #学习速率
n_estimators=5000,
max_depth=4,          #决策树参数
min_child_weight=1,   #决策树参数
gamma=0,              #决策树参数
subsample=0.85,       #决策树参数
colsample_bytree=0.85, #决策树参数
reg_alpha=0.0001,     #正则化参数调优
objective= 'binary:logistic',
nthread=4,
scale_pos_weight=1,
seed=27)
dtrain = train_data
xgb1.fit(dtrain[f_l], dtrain['label'],eval_metric='auc')
dtrain_predictions = xgb1.predict(test_data[f_l])
dtrain_predprob = xgb1.predict_proba(test_data[f_l])[:,1]
test_data['pred'] = list(dtrain_predprob)
test_data['pred1'] = list(dtrain_predictions)
select_sec = test_data[test_data['pred']>np.percentile(dtrain_predprob,98)]
test_df.append(select_sec)
ret_box.append(np.sum(list(select_sec[u'季收益率_Qtrret1']))/float(len(select_sec)))
sel_df = pd.merge(df1,select_sec[[u'上市公司代码_Comcd','pred']],how='inner',on=u'上市公司代码_Comcd')
date2 = pd.tslib.Timestamp(date1[q-3])
date3 = pd.tslib.Timestamp(date1[q-2])
sel_df1 = sel_df[(sel_df.日期_Date>date2) & (sel_df.日期_Date<date3)]
ret_list = list((sel_df1[u'日收益率_Dret']/len(select_sec)).groupby(sel_df1[u'日期_Date']).sum())
ret_b1 = []
for i in range(len(ret_list)-1):
    ret_b = []
    for j in range(i+1,len(ret_list)):
        s = 1
        for x in ret_list[i:j+1]:
            s *= (x+1)
        ret_b.append(s-1)
    max_recall_i = np.min(ret_b)
```

```
        ret_b1.append(max_recall_i)
    max_recall = np.min(ret_b1)
    max_recall_b.append(max_recall)
    print(q)
    sharp = (np.mean(ret_box)-0.0262)/np.std(ret_box)
    Max_Recall = np.min(max_recall_b)

    sharp = (np.mean(ret_box)-0.0262)/np.std(ret_box)
    Max_Recall = np.min(max_recall_b)
    print(sharp,Max_Recall,np.sum(ret_box))

    import matplotlib
    plt.figure()
    hs.plot()
    zhfont1 = matplotlib.font_manager.FontProperties(fname='imhei.ttf')
    plt.rcParams['font.sans-serif'] = ['SimHei']
    plt.rcParams['font.serif'] = ['SimHei']
    plt.xticks(fontproperties=zhfont1)
    plt.show()
```

二、特征重要性输出

```
    def modelfit(alg, dtrain, predictors,useTrainCV=True, cv_folds=5, early_stopping_rounds=50):
        target = 'label'
        if useTrainCV:
            xgb_param = alg.get_xgb_params()
            xgtrain = xgb.DMatrix(dtrain[predictors].values, label=dtrain[target].values)
            cvresult = xgb.cv(xgb_param, xgtrain,num_boost_round=alg.get_params()['n_estimators'],
nfold=cv_folds,metrics='auc', early_stopping_rounds=early_stopping_rounds,verbose_eval=True)
            alg.set_params(n_estimators=cvresult.shape[0])
    #Fit the algorithm on the data
        alg.fit(dtrain[predictors], dtrain['label'],eval_metric='auc')
        #Predict training set:
        dtrain_predictions = alg.predict(dtrain[predictors])
        dtrain_predprob = alg.predict_proba(dtrain[predictors])[:,1]
    #Print model report:
        print("\nModel Report")
        print("Accuracy : %.4g" % metrics.accuracy_score(dtrain['label'].values, dtrain_predictions))
        print("AUC Score (Train): %f" % metrics.roc_auc_score(dtrain['label'], dtrain_predprob))
        feat_imp = pd.Series(alg.booster().get_fscore()).sort_values(ascending=False)[:200]
        import_f = feat_imp.index
        feat_imp.plot(kind='bar', title='Feature Importances')
```

```
# plt.ylabel('Feature Importance Score')
    import matplotlib
    zhfont1 = matplotlib.font_manager.FontProperties(fname='D:\Desktop\simhei.ttf')
    plt.ylabel('特征重要性分数',fontproperties=zhfont1)
    plt.title('特征重要性',fontproperties=zhfont1)
    plt.rcParams['font.sans-serif'] = ['SimHei']
    plt.rcParams['font.serif'] = ['SimHei']
    plt.xticks(fontproperties=zhfont1)
    plt.show()
    return import_f
```

三、XGBoost 模型的 ROC 曲线图绘制

```
from sklearn.metrics import roc_curve
fpr,tpr,thresholds = roc_curve(test_data['label'],dtrain_predprob,pos_label=1)
plt.plot(fpr,tpr,linewidth=2,label='ROC of xgboost')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.show()
```