基于风格轮动和集成学习的多因子选股投资策略研究

摘要

我国金融市场日渐成熟，在大数据和信息化的时代特征下，国内市场的量化产品比重逐渐上升，量化投资的方式能够获得超额收益，量化投资的理念也获得了更多投资者认可，因此越来越多地被使用。量化选股是量化投资的主要内容之一，而多因子模型又是量化选股模型中应用最广泛的一种，经过国内外多年的研究发展，已经成为量化投资中比较成熟的、系统的理论，近年来在我国资本市场有着良好的发挥。风格轮动也是量化选股的重要组成部分，其中风格现象又主要包括成长、价值风格轮动和大盘、小盘风格轮动两种，本文针对大小盘风格轮动进行建模分析。就目前而言，我国的量化选股存在着策略单一、业绩分化等缺点。因此，探寻与挖掘新的量化选股方式，推动量化投资的发展显得尤为重要。

从以往的研究来看，国内的大小盘轮动现象比较明显，风格收益差较大，如果能够在合适的时间进行选择恰当的股票池，则能够获得较高的投资收益。基于此，本文就传统的量化选股方法进行了改进，并针对“泰迪杯”比赛涉及的问题进行相应解决，本文的解决思路如下：

针对问题一，首先通过夏普比率与手动筛选相结合筛选因子，计算单因子选股策略对应的夏普比率，筛选出使得夏普比率大于 1 的因子，并根据经济意义和金融理论手动筛选，将夏普比率小于 1 但是经济意义较为重要的变量也保存在候选因子中；再通过 Adaptive Lasso 进行因子筛选；最后通过相关性分析进行筛选，得到最终的既具有经济意义又富含预测能力的因子。

针对问题二，通过使用沪深 300 和中证 500 的月度收益率代表大小盘股指，利用已有研究得到对于大小盘风格轮动较为重要的解释变量（主要包括宏观经济数据与大小盘收益率之差相关的指标），进而利用这些解释变量进行内生化的风格轮动建模。通过 CatBoost、LightGBM、XGBoost、随机森林等机器学习模型，利用规定时间段内的因子数据，建立模型对月度超额收益率进行滚动训练预测，并将机器学习模型回测结果与等权重线性的多因子模型回测结果进行对比。

针对问题三，通过仓位控制控制最大回撤。由于要求最大回撤<10%前提下进行收益最大化的建模，因此需要对整体仓位进行限制。通过控制最大回撤小于 10%确定总资产中分配于股票市场的资金，并将总资产中其他资金投资于国债等无风

险标的。

综上所述，本文量化投资方案的亮点在于：一是创新了风格轮动解决思路，使用内生化的风格轮动策略量化宏观经济变化对于大小盘风格的影响，同等风险下提高了选股的收益；二是使用了较为新颖的机器学习算法以及集成学习思路，模型具有较高的样本外预测准确率；三是将选股模型与资产组合理论相结合，获得更稳健的超额收益。

关键词：多因子选股；风格轮动；集成学习；CatBoost；LightGBM

Abstract

China's financial market is becoming more and more mature. In the era of big data and information technology, the proportion of quantified products in the domestic market is gradually rising. The way of quantified investment can obtain excess returns. The concept of quantified investment has also been recognized by more investors, so it is used more and more. Quantitative stock selection is one of the main contents of quantitative investment, and multi-factor model is one of the most widely used quantitative stock selection models. After years of research and development at home and abroad, it has become a mature and systematic theory in quantitative investment. In recent years, it has played a good role in China's capital market. Style rotation is also an important part of quantitative stock selection. Style phenomena include growth, value style rotation and market, small-disk style rotation. This paper aims at modeling and analysis of small-disk style rotation. At present, there are some shortcomings in quantitative stock selection in China, such as single strategy, performance differentiation and so on. Therefore, it is particularly important to explore and excavate new ways of quantitative stock selection and promote the development of quantitative investment.

According to previous studies, the phenomenon of large and small market rotation is obvious in China, and the return on style is quite different. If we can choose the right stock pool at the right time, we can get a higher return on investment. Based on this, this paper improves the traditional method of quantitative stock selection, and resolves the problems involved in the "Teddy Cup" competition. The ideas of this paper are as follows:

In view of the first problem, first by sharpe ratio combined with manual screening filter factor, calculation of single factor to choose a strategy corresponding to the sharpe ratio, screen out the sharpe ratio greater than 1 factor, and according to the theory of economic significance and financial manual screening, the sharpe ratio less than 1 but economic significance is relatively important variable is stored in the candidate factor; Adaptive Lasso was used for factor screening; Finally, the factors with both economic

significance and prediction ability are obtained through correlation analysis.

To solve the second problem, by using the monthly return of Shanghai-Shenzhen 300 and China Securities 500 to represent the large and small stock indices, and using the existing research, we get the more important explanatory variables (mainly the indicators related to the difference between macroeconomic data and small stock returns), and then use these explanatory variables to build an endogenous style rotation model. Through the machine learning models such as CatBoost, LightGBM, XGBoost and Random Forest, the rolling training prediction analysis of the model is established by using the factor data in a specified period of time, and the results of the machine learning model and the multi-factor model with equal weight linearity are compared.

Aiming at the third problem, the maximum withdrawal is controlled by position control. Because of the requirement of maximum withdrawal less than 10% for profit maximization, it is necessary to restrict the overall position. By controlling the maximum withdrawal less than 10%, the funds allocated to the stock market in the total assets are determined, and other funds in the total assets are invested in risk-free targets such as Treasury bonds.

In summary, the highlights of this paper's quantitative investment plan are as follows: firstly, it innovates the idea of style rotation, uses the endogenous style rotation strategy to quantify the impact of macroeconomic changes on the style of large and small plates, and improves the returns of stock selection under the same risk; secondly, it uses novel machine learning algorithms and integrated learning ideas, and the model has a high out-of-sample prediction accuracy. The third is to combine stock selection model with portfolio theory to obtain more robust excess returns.

Keywords: Multi-factor Stock Selection; Style Rotation; Integrated Learning; CatBoost; LightGBM

目 录

第1章　绪论

1.1 研究背景及意义

一、多因子选股模型的发展

自 19 世纪奥地利学派发动“边际革命”以来，经济学从外貌上越来越像自然科学。作为一门精确的语言，数学提高了经济学的思想交流效率，加速了经济学的发展和传承。以多因子模型为例，1993 年 Eugene F.Fama 和 Kenneth R. French 发表了三因子模型，开创了多因子模型的先河，他们认为股票市值、账面市值比和市场风险这三个因子可以显著地解释股票价格的变动。Fama 的学生 Clifford S.Asness 随后发现了动量因子，经过细致的测试发现可以利用多因子模型获利，他在芝加哥大学取得金融学博士后立刻奔赴高盛建立了模型，在世界各国开展投资，获得了巨大成功，Asness 离开高盛后创立了对冲基金 AQR，目前管理着约 600 亿美元资产。随后 BGI 公司的 Richard C. Grinold 和 BARRA 公司的 Barr Rosenberg 进一步发展了多因子模型。

每个人脑海中都有一个多因子模型，包括散户和基本面投资经理。信奉价值投资的基本面基金经理会选择估值低、基本面较好的股票，也许还会考虑过去一段时间的涨跌幅，这就涉及了 3 个因子，然后选择上述三个方面表现都不错的股票买入。量化模型大致分为两派：P quant 和 Q quant。前者通常是股票量化投资，使用统计工具，后者通常是利率/汇率衍生品量化投资，使用随机数学作为工具。Q quant 最擅长预测利率/汇率的未来走势，而这类方法恰好可以预测因子的未来效果，筛选出未来有效因子之后，投资收益和信息比都显著提高，从而静态多因子变成了动态多因子。2010 年以来，A 股大量市场参与者已经按照多因子模型进行投资，它们的交易行为改变了市场，使得同质化的模型不再那么有效。

二、集成学习的发展

集成学习是典型的实践驱动的研究方向，它一开始先在实践中证明有效，而后才有学者从理论上进行各种分析，这是非常不同于大名鼎鼎的 SVM。SVM 是先有理论，然后基于理论指导实现了算法。这是机器学习研究中少有的理论指导的

创新案列。首先回顾下集成学习中最主流的随机森林的发展历程。1995 年，AT&T bell 实验室的香港女学者 Ho Tin Kam 最早提出了 RF，那个时候还不叫 Random Forests，而叫 RDF（Random Decision Forest），她主要是采用 Random Subspace 的思想使用 DT（Decision Tree）来构建 Forest。随后的几年里，又有一批人相继提出了大大小小的一些类似或改进的工作。到了 2001 年，统计学家 Breiman 已开始在机器学习界站稳脚跟。他在 RDF 基础上又引入了 Bagging 技术，并提出了沿用至今的 Random Forests。2005 至 2015 这十年里，集成学习方面的论文陆续有发出。大多工作都是围绕一个特定的算法做分析，始终没有一个大一统的理论站稳脚跟。回顾集成学习理论的发展历程，从 bias-variance 分解角度分析集成学习方法，人们意识到：Bagging 主要减小了 variance，而 Boosting 主要减小了 bias，而这种差异直接推动结合 Bagging 和 Boosting 的 MultiBoosting 的诞生。值得一提的是，我国学者在集成学习领域并不落后，以南大周志华教授为代表的学者的一系列工作走在了世界前列，如选择集成技术、集成聚类技术、半监督集成技术等等。周志华老师还最早将 Ensemble Learning 翻译为“集成学习”，是国内这一领域的先行者。

三、研究意义

从量化投资的角度来看，本文的研究意义分为以下三个方面：

第一，从对历史的认识来看，基本面分析看似全面却并不见得准确。人的大脑已开发的功能有限，难以正确处理纷繁复杂的海量信息。某些信息被主观放大，另外一些信息则会被忽略，这很容易导致人们的认知出现偏差甚至是错误，这将对未来的投资产生误导。而计算机对于输入的全部信息都会平等地加以考察。对每个因素所发挥的历史作用都能进行精确的测量，也就是说，它在有限的信息范围内能做到准确全面的处理。当然，准确全面的程度有赖于使用计算机的人的能力， 但从方法论的角度来说，它无疑是最精确的。

第二，从投资决策方面来说，基本面派难以做到足够的客观，主观感性的影响无处不在。即使经历相似的投资者在面对同样的信息时也会得出不同的判断，同一个人在不同环境中也可能作出完全迥异的操作，显然人为主观因素(包括喜好、心情、性格等)都产生了非常重要的影响。当然这并不是要否定主观感性，

而是想说明人为主观很可能会使得投资者放弃理性的思考，扭曲对客观事实的理解。而冷冰冰的计算机程序足以克服人性的弱点，它能够非常忠实地执行模型开发者所完成的理性的研究成果，而不受其他因素的干扰。同样的信息输入，它得出的结论是唯一的、明确的，并且足够客观、足够理性。

第三，量化投资可以大大减轻人脑的负荷，帮助人们进行更高效的投资。计算机程序可以同时处理大量的信息。例如数量选股模型可以在输入千万个数据后快速批量地输出股票组合，而人脑如果要选出同样的组合恐怕需要好几个月的辛勤劳作，却并不见得能取得更好的成绩。另外计算机还能不知疲倦地工作，这会显著提高投资者把握机会的几率。

因此，开展量化方面的投资和研究是非常有必要的，它将对传统投资起到非常好的补充和提升作用。我们不可因为长期资本管理公司的破产就产生恐惧心理，而致因噎废食。量化模型是很优秀的投资工具，结果好坏的关键在于开发者和使用者如何运用，而不应归咎于量化手段本身。

与此同时，在进行量化投资选择的时候是会有自己的投资偏好的。有些投资者比较偏好成长股，有时候又会比较偏好价值股。可能在某些阶段比较偏好小盘股，在另外的时期又比较偏好大盘股。市场风格就是由于投资者的这种不同的交易行为产生的。所以如果能够在投资中紧跟市场风格变化行动，就会比一直持有的效果要好的多。因此当多因子选股结合风格轮动策略将会使得收益更高。

综上所述，风格轮动策略和多因子模型在市场上都已经得到实证检验和应用，在本文以多因子模型为主，风格轮动策略为辅并采用集成学习算法 CatBoost 等三者联合使用，一是比较符合国内金融市场的主要特征，二是充分发挥各自的优势进行互补，三是符合量化投资的时代潮流。

1.2 研究内容和技术路线

1.2.1 研究内容

本文的研究内容主要有以下两个方面：

一、内生化大小盘风格轮动

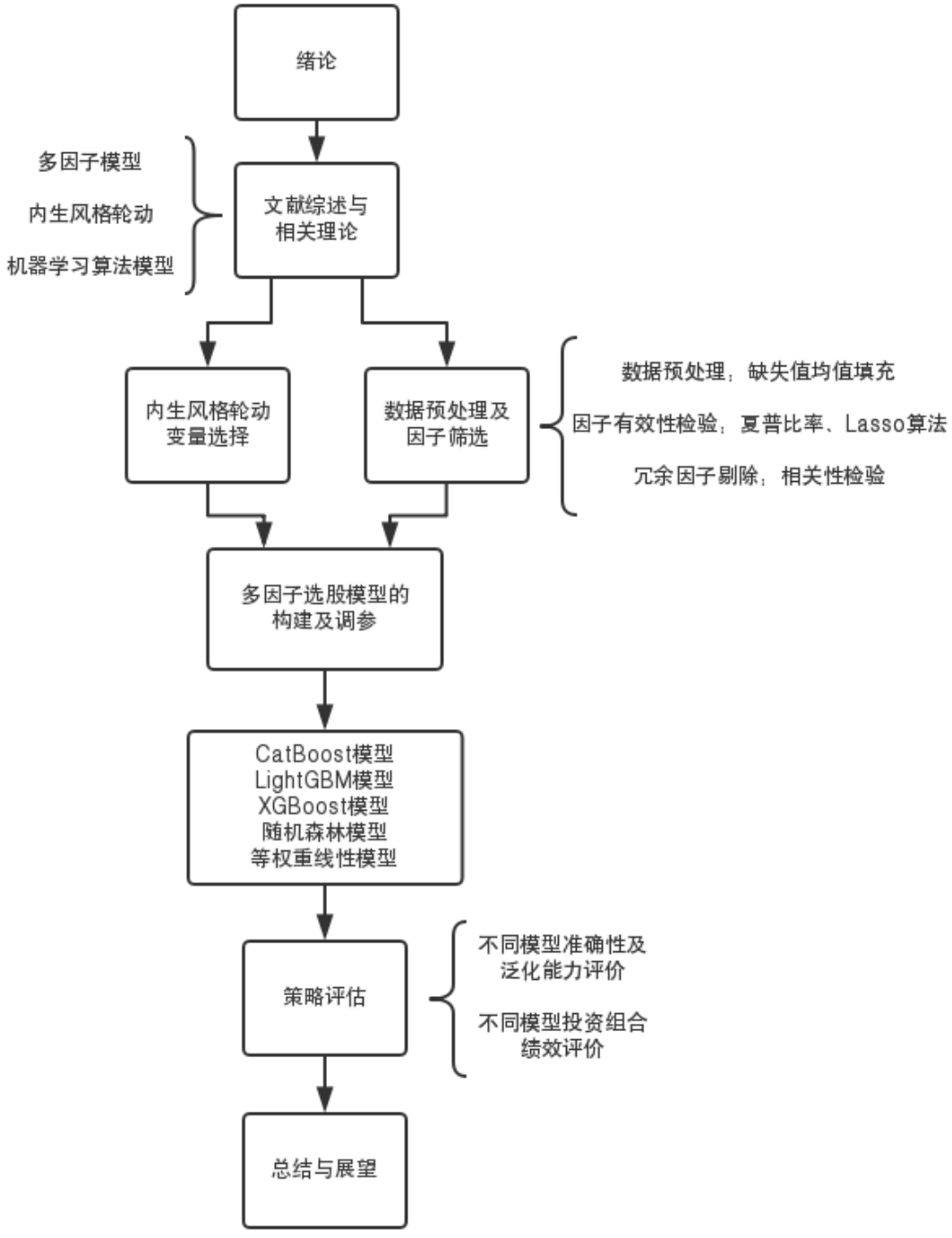
通过使用沪深 300 和中证 500 的月度收益率代表大小盘股指，利用已有研究得到对于大小盘风格轮动较为重要的解释变量（主要包括宏观经济数据与大小盘收益率之差相关的指标），进而利用这些解释变量进行内生化的风格轮动建模。大小盘收益的高低存在着一个轮动现象，大小盘强势轮流交替，有时大盘股收益明显要高于小盘股，有时小盘股收益明显要高于大盘股，有时大小盘交替趋势明显，风格转换迅速，呈震荡趋势，大小盘转换频繁。为了更好的看出转换趋势，得出投资策略，本策略通过宏观经济指标建立内生的风格轮动模型进行选股。

二、多因子选股模型的构建

首先基于十二大类指标，选择较为普遍的指标作为模型的候选因子，通过计算单因子选股策略对应的夏普比率，筛选出使得夏普比率大于 1 的因子；接下来根据经济意义和金融理论手动筛选，将夏普比率小于 1 但是经济意义较为重要的变量也保存在候选因子中；再进行因子的有效性检验，通过 Adaptive Lasso 进行因子筛选，即衡量因子和因子回报是否显著进行有效因子的选取；最后通过相关性分析进行进行冗余因子的剔除，得到最终的既具有经济意义、富含预测能力又不具有较强共线性的因子。

选取 2016 年 1 月至 2018 年 8 月间沪深 300 指数和中证 500 指数总计的 800 只成分股月度数据作为样本数据，通过对数据进行预处理及因子筛选后，使用 LightGBM 算法框架进行参数调优和样本训练得到最优模型来对股票分类，达到获取优质股票的目的，组成股票投资组合并进行评价，得出可行性的投资策略。

1.2.2 技术路线


绪论
多因子模型
内生风格轮动
机器学习算法模型
文献综述与
相关理论
内生风格轮动
变量选择
数据预处理及
因子筛选
数据预处理，缺失值均值填充
因子有效性检验，夏普比率、Lasso算法
冗余因子剔除，相关性检验
多因子选股模型的
构建及调参
CatBoost模型
LightGBM模型
XGBoost模型
随机森林模型
等权重线性模型
策略评估
不同模型准确性及
泛化能力评价
不同模型投资组合
绩效评价
总结与展望

1.3 本文的主要贡献

一、从金融背景的应用角度来看，国内的多因子选股投资策略往往是以单一的多因子模型为主，既缺少量化投资的多元性又没有考虑到整体市场的即时性，未充分考虑风格轮动的重要性。考虑到以上问题，本文在多因子选股的基础上结合了大小盘风格轮动策略，并以多因子模型为主，大小盘风格轮动策略为辅，共同选出优质的股票组合。

二、从研究方法的创新角度来看，首先，本文在模型算法的使用上有所创新，使用的是较为新颖的 CatBoost、LightGBM 算法框架作为多因子选股模型的算法，并与集成学习算法随机森林以及 XGBoost 进行建模效果比对；其次，本文没有采用传统的多因子选股方法，即一般使用打分法或者线性回归，一方面是因为该方法很难挖掘出其中复杂的非线性关系，另一方面由于选取的数据样本数据量较大，很难实现打分法或者线性回归，因此本文采用机器学习算法构建多因子选股模型来对股票进行分类，进而在分类的基础上实现了多因子的量化选股目标，解决股票投资中选择优质股票的问题。

第 2 章 文献综述和相关理论

2.1 文献综述

本文将从量化投资、多因子选股、大小盘风格轮动三个方面进行研究，其中量化投资最重要的策略就是量化选股，而多因子模型和风格轮动模型皆属于量化选股投资策略的范畴。

2.1.1 量化投资

在国内相关文献方面，2012 年电子工业出版社出版了丁鹏编著的《量化投资:策略和技术》[1]，这是国内第一本有关量化投资策略的著作，介绍了量化投资的各个方面的内容，包括策略和理论两部分。策略部分包括:量化选股、量化择时、股指期货套利、商品期货套利、 统计套利、期权套利、算法交易和资产配置等;理论部分包括:人工智能、数据挖掘、小波分析、支持向量机、分形理论、随机过程及 IT 技术等。同年，郭剑光翻译了国外量化投资的书籍《打开量化投资的黑箱》[2]，全书介绍了许多量化投资交易策略,包括阿尔法模型、风险模型、交易成本模型。在国外相关文献方面，量化投资最早起源于 20 世纪中期，马科维茨（1952）发表《投资组合选择理论》[3]，建立了均值-方差模型，在维持期望收益尽可能大的情况下，努力减少交易中的风险。这一理论的建立成为后来诸多量化模型的逻辑出发点。在此基础上，Sharpe(1964)，Lintner(1965)建立了资本资产定价模型[4-5]，将目标资产进行了风险资产和无风险资产的划分，这些也成为之后进行证券风险研究时的基本模型和研究出发点。Mossin(1966)以 CAPM 为基础，提出了超额收益的概念，即经风险收益调整后的资产总收益与考虑承受市场风险的风险贴水之后的期望收益率的差[6]，这就是所谓的 Alpha 收益。

2.1.2 多因子选股

国内外学者对于多因子选股模型的研究大多数还停留在传统方法的基础上。

在国外相关文献方面,量化投资在欧美等成熟的金融市场已经形成了比较完善的理论体系同时在量化选股的理论和实证方面也有深入的研究。Fama 和 French(1993)通过对美国股市的研究，提出三因素模型[7]，把股票回报的影响因素归纳为三个，分别是市场资产组合、市值因子、账面市值比因子，有效地解释了上市公司股票收益波动情况。Asness(1997) 研究上市公司的基本面数据对股票收益的影响，研究结果表明，上市公司股价波动于其近期的基本面数据的变化有关[8]。Mohanram(2004)在多因子量化选股模型的基础上，从盈利能力、增长的稳定性和财务的稳健性三个方面选取了 9 个指标对 PB 排名前 1/5 的股票进行打分选股，构造的投资组合具有良好的市场表现[9]。Richard Tortoriello（2009）就美国标普 500 的因子选股模型研究整理成书籍《Quantitative Strategies for Achieving Alpha》[10]，这本书分类列举出可用于量化选股的各因子的具体指标，而且对各因子如何作用于股票市场进行了理论解释。在此基础上筛选出美国股市的有效因子，并对单因子、多因子及多因子选股模型进行归纳，形成了基于不同行业有效因子及因子权重的列表。

国内量化投资起步晚，对多因子选股模型研究也是近些年比较多，而且研究主要集中在券商和基金公司等机构投资者。学术界对多因子量化选股的研究不多，研究方法近似。潘凡(2011)使用多因子模型进行选股，对众多基本面和技术面因子进行筛选后最后得出 9 个有效因子，并且投资组合在样本期表现较好，说明多因子选股模型是有效的[11]。刘毅(2012)运用打分法的选股方式,从 25 个可能影响股票收益率的因素中做了因子实证检验，最终选择了 8 个较好的因子对 A 股市场的股票进行了实证研究，构建了价值、成长、价值成长和价值成长质量四个策略模型并进行评估，最终得出结论是价值成长质量多因子模型表现最好[12]。孙守坤（2013）将多因子模型与行业轮动结合在一起，希望能够找到最优的投资组合策略。在他认为我国股市的行业轮动比较明显，在多因子模型的基础上结合行业轮动能增加投资组合获得超额收益的可能性。他通过对多因子模型进行研究，将股票池中的股票进行分排序并选取排名前 30 名的股票作为投资规模构建投资组合可以取得超额收益的结果[13]。王昭栋(2014)利用当时在中国证券市场使用的主流多因子模型即打分法和回归法来对两个模型的评估和比较[14]。

对于机器学习方法在量化选股中的应用，张晨宇(2017) 使用量化的方法分

析上市公司的部分财务指标和交易数据并采用数据挖掘技术中的一些聚集函数来对股票数据进行分析和预测,研究结果表明经过平台系统选出的股票收益率的增长大于同期沪深 300 指数[15]。李文星,李俊琪（2018）利用半监督 K-means 算法应用于多因子选股模型中。研究表明,相比传统的聚类模型,改进的模型具有较强的泛化能力,模型在处理样本线性不可分等问题上具有明显的优势并且可以选出较优的股票组合[16]。陈满祥等人（2018）从股价变动和财务因子之间的关系利用 logistic 回归模型分别预测个股在近期的财务报表公布后上涨概率,并对其进行从高到低的排序,最终获得上涨概率最高的一只股票作为投资标的[17]。

2.1.3 大小盘风格轮动

我国学者们专门针对大小盘风格轮动的研究相对较少,而仅把大小盘风格轮动作为风格投资的一个研究方面。因而在本文,第一，因为在我国关于大小盘轮动的针对性研究多见于券商的研究报告中,将关注券商研究人员对大小盘轮动的研究情况；第二，将关注我国学者对大小盘风格轮动策略进行量化选股的研究情况。

第一，国内关于大小盘风格轮动的研究主要集中在券商的研究报告当中，其中相对比较有代表性的研究报告包括：曹源（2010）通过选取货币、宏观经济和市场情绪等因子建立多元回归模型，对 A 股的大小盘轮动规律进行研究，发现在 2007 年的牛市中，大小盘轮动策略能够很好地抓住大盘股的投资机会，而在 2010 年时大小盘轮动策略又可以捕捉小盘股的轮动特点，从而抓住投资机会[18]。国金证券的杨勇（2015）从价格和交易量动量的角度构建大小盘轮动模型进行预测，通过计量分析，在历史数据集上计算得出模型最优参数为 20 个交易日。尽管模型的结果是统计意义上的最优，并且模型信号具有一定滞后性，还是可以结合模型信号，对大小盘风格做出较为准确的推断[19]。

第二,一些金融学者和机构研究者开始利用大小盘风格轮动策略进行量化选股，以希望取到更好的选股效果进而增加获得超额收益的可能性，其中最有代表性是曹力（2010）研究了大小盘轮动的情况，采用对数差的形式定义小盘股对大盘股的相对优势指标,使用类似布林线的上下轨和双阈值来判断大小盘转换的将

征和趋势，并且结合中国股市的总情况，分析不同时间段投资大盘股还是小盘股更加合适[20]。

本文将建立内生风格轮动量化选股模型中选择解释变量部分建立于前人的研究基础之上，即利用前人研究进行变量选择，进而利用最先的算法进行模型融合与变量内生化。

2.2 多因子模型概述

2.2.1 多因子模型介绍

在量化交易中，多因子策略是一种常被提及且应用广泛的选股策略。我们会经常使用某种指标或者多种指标来对股票池进行筛选，这些用于选股的指标一般被称为因子。顾名思义，多因子模型是指使用多个因子，综合考量各因素而建立的选股模型，其假设股票收益率能被一组共同因子和个股特异因素所解释。多因子模型的优点在于，它能通过有限共同因子来有效地筛选数量庞大的个股，在大幅度降低问题难度的同时，也通过合理预测做出了判断。

各种多因子模型核心的区别第一是在因子的选取上，第二是在如何用多因子综合得到一个最终的判断。一般而言，多因子选股模型有两种判断方法，一是打分法，二是回归法。打分法就是根据各个因子的大小对股票进行打分，然后按照一定的权重加权得到一个总分，根据总分再对股票进行筛选。回归法就是用过去的股票的收益率对多因子进行回归，得到一个回归方程，然后再把最新的因子值代入回归方程得到一个对未来股票收益的预判，然后再以此为依据进行选股。

2.2.2 多因子选股步骤

多因子选股模型的建立过程主要分为候选因子的选取、选股因子有效性的检验、有效但冗余因子的剔除、综合评分模型的建立和模型的评价及持续改进等五个步骤。

第一，候选因子的选取，候选因子的选择主要依赖于经济逻辑和市场经验，但选择更多和更有效的因子无疑是增强模型信息捕获能力，提高收益的关键因素之一。本文首先使用夏普比率进行筛选，筛选出单因子策略中使夏普比率大于 1

的因子，并在剩余的因子中选出经济意义较强的因子，进入下一步的筛选环节。

第二，选股因子有效性的检验，一般检验方法主要采用排序的方法检验候选因子的选股有效性。但本文使用的是非线性的机器学习方法进行多因子选股模型的建立，故采用下文详细介绍的 Lasso 法进行筛选。

第三，有效但冗余因子的剔除，不同的选股因子可能由于内在的驱动因素大致相同等原因，所选出的组合在个股构成和收益等方面具有较高的一致性，因此其中的一些因子需要作为冗余因子剔除，而只保留同类因子中收益最好，区分度最高的一个因子。本文中使用相关性分析进行冗余因子删除。

第四，综合评分模型的建立和选股，综合评分模型选取去除冗余后的有效因子，根据选股评分方法，本文利用机器学习的方法，根据需要选择排名靠前的股票。

第五，模型的评价及持续改进，一方面，由于量化选股的方法是建立在市场无效或弱有效的前提之下，随着使用多因子选股模型的投资者数量的不断增加，有的因子会逐渐失效，而另一些新的因素可能被验证有效而加入到模型当中；另一方面，一些因子可能在过去的市场环境下比较有效，而随着市场风格的改变，这些因子可能短期内失效，而另外一些以前无效的因子会在当前市场环境下表现较好。

另外，计算综合评分的过程中，各因子得分的权重设计、交易成本考虑和风险控制等都存在进一步改进的空间。因此在综合评分选股模型的使用过程中会对选用的因子、模型本身做持续的再评价和不断的改进以适应变化的市场环境。

2.3 内生风格轮动策略理论概述

为了做到较为准确的得到大小盘轮动，要清晰的看出大小盘运动的趋势，而要观察大小盘的运动趋势，就需要先找出一个能反映大小盘优势比较趋势的指标，也就是相对优势指标，进而利用相对优势指标的滞后值，对未来股票的月度超额收益率进行建模和预测。在捕捉往复轮动趋势的研究中，针对大小盘的预测问题，多数基金公司采取的相对优势指标的具体计算方法如下公式：

$$r_t = R_{t,1} - R_{t,2} \qquad (2\text{-}1)$$

$$R_{t,j} = \sum R_{d,j} \quad (2\text{-}2)$$

其中R_t^1是小盘股在 t 月份的收益率，R_t^2是大盘股 t 月份的收益率,具体计算方式如公式 2-2 所示，$R_{d,j}$为每天的收益率，r_t为小盘月度收益率减去大盘月度收益率的差值。从整体来看，小盘相对优势用同一时间指标做差来表示。

表 1 解释变量及其意义

解释变量	意义
L1_mean_pe	市场整体 PE 的一阶滞后
L3_indus_tbhb	工业增加值同比环比的三阶滞后
L2_CPI_tbhb	CPI 同比环比的二阶滞后
L_M2tbhb	M2 同比环比的一阶滞后
L_confidence	市场信心指数的一阶滞后
L_house	国房景气指数（综合指数）的一阶滞后
L_rate	短期贷款利率的一阶滞后
L_government	财政预算支出：当月同比的一阶滞后
L_exchange_rate_usdcny	平均汇率：美元兑人民币（月）的一阶滞后
last_3_diff_rev	前三个月累计收益率之差
L2_L_advantage	小盘与大盘月度收益率差值的二阶滞后
L1_L_advantage	小盘与大盘月度收益率差值的一阶滞后

结合前人的研究，筛选出重要的宏观经济变量如表 1 所示，将配合风格轮动中相对优势指标的滞后值对未来一个月的每一只股票月度收益率进行预测。整体而言，市场的整体估值越高越有利于小盘板块。我们认为短期市场环境会影响到投资者对大小盘风格的偏好，当市场整体估值加大，表明投资者情绪比较乐观，此时投资者会对风险偏高的小盘股更加偏好。

上升的通货膨胀带来货币紧缩，因此大盘股表现相对好些。对比 CPI 与大小盘相对强弱发现，CPI 越大越有利于大盘板块。PPI 以及 CPI 增加时通胀压力增大，央行采取紧缩的货币政策可能性加大，由于大盘股相对而言财务更稳健，冲击相对较小。此外，由于大小盘的轮动与经济周期存在一定的联系，而通货膨胀在一定程度上能够反映经济周期的更替。

工业增加值同比环比(滞后 3 期)增长越高，大盘股的走势越好。这个关系是符合经济学直观的，因为小盘股更容易受到市场情绪，资金面等因素的影响，这些噪声降低了实体经济与小盘股之间的相关系数。大盘股则相对不容易受到市场情绪的影响，而更多地受实体经济的影响，二者的相关系数也比较高。

考虑到数据可得性问题，本文对数据进行了相应的滞后处理，从而保证在进行预测的时候，模型需要的数据均为可以获得的数据。

2.4 算法介绍

2.4.1 Adaptive-Lasso 算法

Lasso 是近年来被广泛应用于参数估计和变量选择的方法之一，并且在确定的条件下，使用 Lasso 方法进行变量选择已经被证明是一致性的。Lasso 是由 Tibshirani 提出的将参数估计与变量选择同时进行的一种正则化方法。Lasso 参数估计被定义如下：

$$\hat{\beta}(\text{lasso}) = \underset{\beta}{\text{argmin}}^{2} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (2\text{-}3)$$

其中，λ为非正数正则参数，$\lambda \sum_{j=1}^{p} |\beta_j|$称为惩罚项。

Lasso 方法虽然可以解决最小二乘法和逐步回归局部最优估计的不足，但是其自身需要满足一定的苛刻条件。Hui ZOU 提出改进 lasso 方法-Adaptive-Lasso，该方法通过为不同系数赋予不同权重，定义如下：

$$\hat{\beta}^{*(n)} = \underset{\beta}{\text{argmin}} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_{\text{n}} \sum_{j=1}^{p} \hat{\omega} |\beta_j| \qquad (2\text{-}4)$$

其中，权重 $\hat{\omega}_{\text{j}} = \frac{1}{\left|\hat{\beta}_{\text{j}}\right|^{\gamma}} (\gamma > 0)$，j=1，2，.....，p，$\hat{\beta}_{\text{j}}$为由普通最小二乘法得出的系数。

2.4.2 决策树算法模型

决策树算法是一种常见的机器学习方法，通过对训练数据进行测算，实现对新数据进行分类和预测的算法，它是后面集成学习算法的基础，因此也需要进行建模。简单来讲决策树算法就是通过对已有明确结果的样本数据进行分析，寻找数据中的特征或规则，并以此为依据对新数据的结果进行预测。决策树算法的基本分类如表 2 所示：

表 2　决策树算法分类表

决策树算法	算法描述
ID3 算法	核心是在决策树的各级节点上，使用信息增益方法作为属性的选择标准，来帮助确定生成每个节点时所应采用的合适属性
C4.5 算法	C4.5 决策树生成算法相对于 ID3 算法的重要性改进是使用信息增益率来选择节点属性。而 C4.5 算法可以克服 ID3 算法存在的不足：ID3 算法只适用于离散的描述属性，而 C4.5 算法既能够处理离散的描述属性，也可以处理连续的描述属性
CART 算法	CART 决策树是一种十分有效的非参数分类和回归方法，通过构建树、修建树、评估树来构建一个二叉树。当终节点是连续变量时，该树为回归树；当终节点是分类变量时，该树为分类树

决策树算法涉及到的重要概念：

信息熵是度量样本集合纯度最常用的指标，假定当前样本集合 D 中第 k 类样本所占的比例为p_k（k=1,2，…，$|\gamma|$），则 D 的信息熵定义为：

$$Ent = -\sum_{k=1}^{|\gamma|} p_k \, log_2 \, p_k \tag{2-5}$$

Ent（D）的值越小，则 D 的纯度越高。

考虑到不同的分支结点所包含的样本数不同，我们给分支结点赋予权重是的样本数越多的分支结点的影响越大，因此将信息增益表示为：

$$Gain(D，a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2-6}$$

信息增益率使用分裂信息值将信息增益规范化。信息增益率定义为下式

$$Gain_{ratio(D,\ a)} = \frac{Gain(D,\ a)}{IV(a)} \tag{2-7}$$

基尼指数用来算则划分属性，公式如下

$$Gini(D) = \sum_{k=1}^{|\gamma|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\gamma|} p_k^2 \tag{2-8}$$

2.4.3 XGBoost 算法框架模型

XGBoost 是一种梯度提升算法、残差决策树，其基本思想为：一棵树一棵树地逐渐往模型里加入，每加一棵 CRAT 决策树时，要使得整体的效果（目标函数有所下降）有所提升。使用多棵决策树（多个单一的弱分类器）构成组合分类器，每构造新的一颗决策树时，根据已经建立的决策树残差，使用二阶泰勒展开式来优化损失函数，得到新决策树的权值。它的训练速度快，因为自动地运用 CPU 的多线程进行并行计算，同时在算法精度上也进行了精度的提高。

这里的决策树是分类与回归树 CRAT（classification and regression tree,CART）。CART 决策树是二叉树，内部结点特征的取值为“是”和“否”，左分支是取值为“是”的分支，右分支是取值为“否”的分支。CART 会把输入根据属性分配到各个叶子节点上，而每个叶子节点上面会对应一个分数值（权值）。

XGBoost 是在 GBDT 等提升算法基础上进行优化的算法。优化目标函数需要尽量使得预测值接近真实值，同时保证模型的泛化能力。可以通过最小化损失函数并加上控制模型复杂度的惩罚项也称为正则化项，实现优化目标函数以达到误差和复杂度综合最优。其原理如下：

目标函数如下第一个公式，由误差函数 L(x)和复杂度函数 Ω(x)组成。损失函数 Ω(x)为以下第二个公式所示。带入目标函数可得公式第三个公式

$$Obj(x) = L(x) + \Omega(x) \tag{2-9}$$

$$Obj(x) = \sum_i l(y_i, \hat{y}_i) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \left\| W_j \right\|^2 \tag{2-10}$$

其中 l 是可导且凸的损失函数，用来衡量 y ^与 y 的相近程度。通过每一步

增加一个基分类器 f_m，贪婪地去优化目标函数，使得每次增加都使得损失变小。然后让后一次迭代的基分类器去学习前一次遗留下来的误差。这样可以得到用于评价当前分类器 f_m 性能的评价函数,如下式。

$$Obj_m = \sum_{i=1}^{n} l[y_i, \hat{y}^{m-1} + f_m(x_i)] + \Omega(f_m) \tag{2-11}$$

这个算法又可以成为前向分步优化。为了更好更快的优化此函数，可以在 f_m=0 附近二阶泰勒展开，泰勒展开的形式为公式。

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \tag{2-12}$$

进一步定义 g_i， h_i 表达式，如公式所示。

$$g_i = \partial_{\hat{y}^{(m-1)}} l(y_i, \hat{y}^{(m-1)}), h_i = \partial^2{}_{\hat{y}^{(m-1)}} l(y_i, \hat{y}^{(m-1)}) \tag{2-13}$$

最后可得到目标函数,在剔除常数项后可以得到最终的表达式,如公式所示。

$$Obj_m = \sum_{i=1}^{n} [g_i + f_m(x_i) + \frac{1}{2} h_i f_m{}^2(x_i)] + \Omega(f_m) \tag{2-14}$$

XGBoost 主要的优点有一方面 XGBoost 对代价函数进行了二阶泰勒展开，得到一阶和二阶导数并在代价函数中加入了正则项，用于控制模型的复杂度。从权衡方差偏差来看，它降低了模型的方差，使学习出来的模型更加简单，防止过拟合；另一方面，XGBoost 借鉴了随机森林的做法，支持列抽样，不仅防止过拟合，还能减少计算。主要缺点有在每一次迭代的时候，都需要遍历整个训练数据多次。如果把整个训练数据装进内存则会限制训练数据的大小；如果不装进内存，反复地读写训练数据又会消耗非常大的时间。首先，空间消耗大。这样的算法需要保存数据的特征值，还保存了特征排序的结果，这里需要消耗训练数据两倍的内存；其次，时间上也有较大的开销，在遍历每一个分割点的时候，都需要进行分裂增益的计算，消耗的代价大；最后，对缓存优化不友好。在预排序后，特征对梯度的访问是一种随机访问，并且不同的特征访问的顺序不一样，无法对缓存进行优化。

2.4.4 随机森林算法模型

随机森林算法原理的具体过程如下表所示：

表 3　随机森林算法流程示意表

算法：随机森林
（1）从原始训练集中使用 Bootstraping 方法随机有放回采样选出 m 个样本，共进行 n_tree 次采样，生成 n_tree 个训练集；
（2）对于 n_tree 个训练集，分别训练 n_tree 个决策树模型；
（3）对于单个决策树模型，假设训练样本特征的个数为 n，那么每次分裂时根据信息增益/信息增益比/基尼指数选择最好的特征进行分裂；
（4）每棵树都一直这样分裂下去，直到该节点的所有训练样例都属于同一类。在决策树的分裂过程中不需要剪枝；
（5）将生成的多棵决策树组成随机森林。对于分类问题，按多棵树分类器投票决定最终分类结果；对于回归问题，由多棵树预测值的均值决定最终预测结果。

首先进行数据预处理及因子筛选，确定最终的有效因子作为模型的预测变量；其次，构建的响应变量，其中股票表现优异设为 1， 表现一般设为 0；接下来，将数据集分为训练集和测试集；然后，建立模型进行训练；最后，根据测试集的数据预测结果预测股票的分类，并判断模型的准确度，该过程中进行参数调优。

第一，随机森林适合大样本、非线性的金融数据。在本文沪深 300 指数 300 只成分股（或中证 500 指数 500 只成分股）的样本数据，数据样本量较大，同时由于股价受多种因素影响，表现出“没有规律”的随机游动特性，很难挖掘出其中复杂的非线性关系，而随机森林算法与本文所研究的数据情况一致。第二，在对股票进行分类时，因子指标的数据维度较多，而随机森林有着良好的处理高维数据的能力，而且该算法的泛化能力好，容忍不平衡性和特征遗失。

优点：（1）实现比较简单，训练速度快，泛化能力强，可以并行实现，因为训练时树与树之间是相互独立的；（2）它能够处理很高维度（特征很多）的数据，并且不用做特征选择，因为特征子集是随机选取的；（3）在训练完后，它能够给出哪些特征比较重要；（4）对于不平衡的数据集来说，它可以平衡误差；（5）如果有很大一部分的特征遗失，仍可以维持准确度。

缺点：（1）随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟；（2）相比于单一的决策树，它的随机性让我们难以对模型进行解释。

2.4.5 LightGBM 算法框架模型

1、GOSS 算法

GOSS（基于梯度的单边采样）方法的主要思想就是，梯度大的样本点在信息增益的计算上扮演着主要的作用，也就是说这些梯度大的样本点会贡献更多的信息增益，因此为了保持信息增益评估的精度，对样本进行采样的时候保留这些梯度大的样本点，而对于梯度小的样本点按比例进行随机采样即可。

在 AdaBoost 算法中，每次迭代时更加注重上一次错分的样本点，即将上一次错分的样本点权重增大，而在 GBDT 中并没有本地的权重来实现这样的过程，所以在 AdaBoost 中提出的采样模型不能应用在 GBDT 中。但是，每个样本的梯度对采样提供了非常有用的信息。也就是说，如果一个样本点的梯度小，那么该样本点的训练误差就小并且已经经过了很好的训练。一个直接的办法就是直接抛弃梯度小的样本点，但是这样做的话会改变数据的分布和损失学习的模型精度。GOSS 的提出就是为了避免这两个问题的发生。下面就是 GOSS 算法流程示意表：

表 4　GOSS 算法流程示意表

算法：GOSS（基于梯度的单边采样）
输入：训练数据，迭代步数 d，大梯度数据的采样率 a，小梯度数据的采样率 b， 损失函数和弱学习器的类型（一般为决策树）；
输出：训练好的强学习器；
（1）根据样本点的梯度的绝对值对它们进行降序排序；
（2）对排序后的结果选取前 a*100%的样本生成一个大梯度样本点的子集；
（3）对剩下的样本集合（1-a）*100%的样本，随机的选取 b*（1-a）*100%个样本点，生成一个小梯度样本点的集合；
（4）将大梯度样本和采样的小梯度样本合并；
（5）将小梯度样本乘上一个权重系数；
（6）使用上述的采样的样本，学习一个新的弱学习器；

（7）不断地重复（1）~（6）步骤直到达到规定的迭代次数或者收敛为止。

通过上面的算法可以在不改变数据分布的前提下不损失学习器精度的同时大大的减少模型学习的速率。从上面的描述可知，当 a=0 时，GOSS 算法退化为随机采样算法；当 a=1 时，GOSS 算法变为采取整个样本的算法。在许多情况下，GOSS 算法训练出的模型精确度要高于随机采样算法。另一方面，采样也将会增加弱学习器的多样性，从而潜在的提升了训练出的模型泛化能力。

2、EFB 算法

Lightgbm 实现中不仅进行了数据样本采样，也进行了特征抽样，使得模型的训练速度进一步的提高。但是该特征抽样又与一般的特征抽样有所不同，是将互斥特征绑定在一起从而减少特征维度。主要思想就是，通常在实际应用中高维度的数据往往都是稀疏数据，可以设计一种几乎无损的方法来减少有效特征的数量。尤其，在稀疏特征空间中许多特征都是互斥的可以安全的将互斥特征绑定在一起形成一个特征，从而减少特征维度。Lightgbm 作者将互斥特征绑定在一起使用的是基于直方图（histograms）算法。

直方图算法的基本思想是先把连续的特征值离散化成 k 个整数，同时构造一个宽度为 k 的直方图。在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。由于基于直方图的算法存储的是离散的“箱子”而不是连续的特征值，可以通过让互斥特征驻留在不同的“箱子”中来构造特征捆绑。

直方图算法并不是完美的。由于特征被离散化后，找到的并不是很精确的分割点，所以会对结果产生影响。但在不同的数据集上的结果表明，离散化的分割点对最终的精度影响并不是很大，甚至有时候会更好一点。原因是决策树本来就是弱模型，分割点是不是精确并不是太重要；差一点的切分点也有正则化的效果，可以有效地防止过拟合；即使单棵树的训练误差比精确分割的算法稍大，但在 Gradient Boosting 的框架下没有太大的影响。

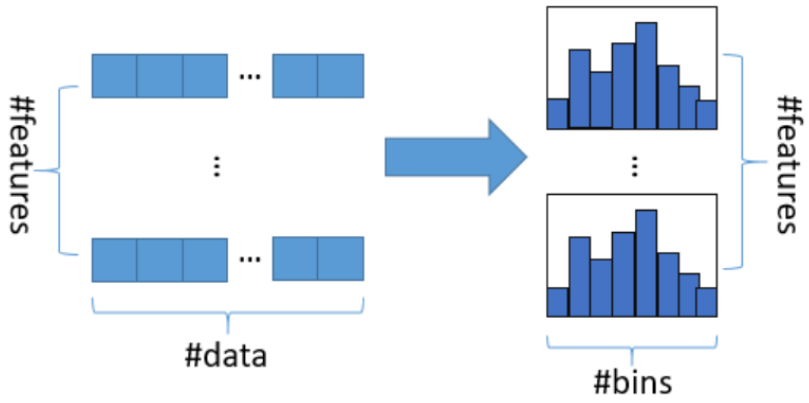


图 2　直方图算法示意图

3、leaf-wise 的决策树生长策略

LightGBM 抛弃了大多数 GBDT 工具使用的按层生长，即采用 level-wise 决策树生长策略，而使用了带有深度限制的按叶子生长 leaf-wise 决策树生长策略。level-wise 过一次数据可以同时分裂同一层的叶子，容易进行多线程优化，不容易过拟合。但实际上 level-wise 是一种低效的算法，因为它不加区分的对待同一层的叶子，带来了很多没必要的开销。因为实际上很多叶子的分裂增益较低，没必要进行搜索和分裂。leaf-wise 则是一种更为高效的策略，每次从当前所有叶子中，找到分裂增益最大的一个叶子，然后分裂，如此循环。因此和 level-wise 相比，在分裂次数相同的情况下，leaf-wise 可以降低更多的误差，得到更好的精度。leaf-wise 的缺点是可能会长出比较深的决策树，产生过拟合。因此 LightGBM 在 leaf-wise 之上增加了一个最大深度的限制，在保证高效率的同时防止过拟合。如下是 level-wise 和 leaf-wise 决策树生长策略示意图：

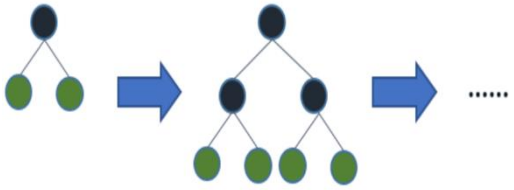


图 3　level-wise 决策树生长策略

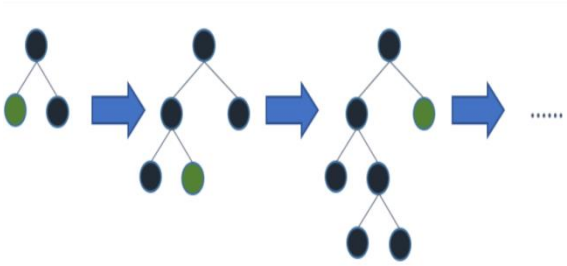


图 4　leaf-wise 决策树生长策略

同以上的两个集成算法的模型一样，首先进行数据预处理及因子筛选，确定最终的有效因子作为模型的预测变量；其次，构建的响应变量，其中股票表现优异设为 1， 表现一般设为 0；接下来，将数据集分为训练集和测试集；然后，建立模型进行训练；最后，根据测试集的数据预测结果预测股票的分类，并判断模型的准确度，该过程中进行参数调优。

LightGBM 算法框架的主要优势是：首先，LightGBM 引入的 GOSS 算法不仅在许多情况下训练出的模型精确度要高于随机采样算法。而且，GOSS 算法采样也将会增加弱学习器的多样性，从而潜在的提升了训练出的模型泛化能力。其次，EFB 算法中的直方图方法中离散化的分割点即并不是很精确的分割点也有正则化的效果，可以有效地防止过拟合；最后，和 level-wise 相比，在分裂次数相同的情况下，leaf-wise 可以降低更多的误差，得到更好的精度，速度较快。结合以上优势，首先，LightGBM 适合大样本、非线性或者线性的金融数据。在本文证 500 指数 500 只成分股的样本数据，数据样本量较大，同时由于股价受多种因素影响，表现出“没有规律”的随机游动特性，而 LightGBM 算法框架与本文所研究的数据情况一致。第二，在对股票进行分类时，因子指标的数据维度较多，而 LightGBM 也有着良好的处理高维数据的能力，而且该算法的泛化能力好，不仅能降低过拟合，还能减少计算，支持并行，速度较快。

2.4.6 CatBoost 算法模型

CatBoost 可赋予分类变量指标，进而通过独热最大量得到独热编码形式的结果（独热最大量：在所有特征上，对小于等于某个给定参数值的不同的数使用独热编码）。

特征组合：

在数据集中，组合的数量随类别特征个数成指数型增长，在算法中不太可能考虑所有。在当前树考虑新的拆分时，Catboost 以贪婪的方式考虑组合。

1、 第一次分裂不考虑任何组合在树上；

2、 对于下一次分类，在有所有类别特征的数据集的当前树，Catboost 包含了所有的组合和分类特征。组合值即被转换为数字；

3、 Catboost 还可以生成数值和类别特征的组合:在树中选择的所有分裂视为具有两个值的类别，并在组合中也类似使用。

这种做法可以降低类别特征中低频次特征带来的噪声。在回归问题中，计算优先级一般通过对标签值取平均得到。针对二分类问题，优先级一般基于正类样本的先验概率得到。针对类别型特征，也可以将两个类别型特征组合来，在 CatBoost 中，第一次分割时不考虑类别型特征的组合，下面的分割中将所有类别型特征之间的组合考虑进来，组合后的特征就会变成数值型。

CatBoost 也会把分割得到的两组值作为类别型特征参与后面的组合。在 GBDT 中，构建下一棵树包含两步，即选择树的结构以及树结构固定之后设置叶子节点的值。在 CatBoost 中，对于每个样本 Sample，都单独构建一个利用该样本之前的样本点的梯度估计得到的模型 Model，针对这些模型 Model，估计该样本 Sample 的梯度，然后利用新模型 Model_new 重新对样本 Sample 打分。

CatBoost 模型依赖于样本的排序，利用多种样本排序可以训练得到多种模型，这样可以减少过拟合现象。CatBoost 具有两大优势，其一，它在训练过程中处理类别型特征，而不是在特征预处理阶段处理类别型特征；其二，选择树结构时，计算叶子节点的算法可以避免过拟合。另外，在 CatBoost 中，将浮点型特征，统计值以及 one-hot 编码的特征二值化，所有这些二值化特征放入向量中，然后在打分过程中利用二值化的特征计算模型的输出。

其算法建立的步骤为下面几步：

(1) 数据样本加载。

(2) CatBoost 模型的创建。使用 Python 提供的 train_test_split 库将数据集划分为训练集与验证集进行训练，设置 CatBoost 的基本参数，树的深度，学习率，迭代次数，L2 正则化系数，其他参数均为默认值。

(3) CatBoost 模型训练。使用 Python 提供的 CatBoostClassifier 模块，

用训练样本数据来训练数据进行分类。

在数据集中，组合的数量随类别特征个数成指数型增长，在算法中不太可能考虑所有。在当前树考虑新的拆分时，Catboost 以贪婪的方式考虑组合。

1、 第一次分裂不考虑任何组合在树上；

2、 对于下一次分类，在有所有类别特征的数据集的当前树，Catboost 包含了所有的组合和分类特征。组合值即被转换为数字；

3、 Catboost 还可以生成数值和类别特征的组合:在树中选择的所有分裂视为具有两个值的类别，并在组合中也类似使用。

这种做法可以降低类别特征中低频次特征带来的噪声。在回归问题中，计算优先级一般通过对标签值取平均得到。针对二分类问题，优先级一般基于正类样本的先验概率得到。针对类别型特征，也可以将两个类别型特征组合来，在 CatBoost 中，第一次分割时不考虑类别型特征的组合，下面的分割中将所有类别型特征之间的组合考虑进来，组合后的特征就会变成数值型。

CatBoost 的建树过程：

初步计算 splits

对每个数值特征二值化，选择可能的分桶方式，结果用于选择树结构。binarization method (feature_border_type) 和 number of buckets (border_count) 参数设置在初始参数中。

将 categorical 特征转化为 numerical 特征 (也可以使用 one-hot 编码) 有以下三步：

(1)随机重排输入对象；

(2)将 target 从浮点数转换为整数 (Regression/ Classification/ Multiclassification 不同)；

(3)将 categorical 特征转化为 numerical 特征 (与初始参数有关)：

〖ctr〗_i= (countInClass+prior)/(totalCount+1)

Borders:

i∈[0,k]

计算第 i 个 Bucket 的 ctr ()，countInClass 是 target 超过当前对象 i 的次数(按照随即重排后的顺序)，totalCount 是到当前对象为止的匹配次数，

prior 是初始参数(常数)。

〖ctr〗_i= (countInClass+prior)/(totalCount+1)

Buckets:

计算第 i 个 Bucker 的 ctr(),countInClass 是 target 和当前对象相等的次数，totalCount 和 prior 同上。

〖ctr〗_i= (countInClass+prior)/(totalCount+1)

BinarizedTargetMeanValue:

countInClass 是当前 cat 的 target 总数和最大 target 总数的比，totalCount 和 prior 同上。

〖ctr〗_i= (countInClass+prior)/(totalCount+1)

Counter:

curCount 是训练集中具有当前特征的对象总数，maxCount 是训练集中最大对象总数，prior 同上。对于测试集，curCount 取决于选择的计算方法：PrefixTest (具有当前特征的训练集中的对象数和到目前为止的测试集中的对象数)，FullTest (具有当前特征的训练集和测试集中的对象数)，SkipTest (具有当前特征的训练集中的对象数)。maxCount 取决于选择的计算方法：PrefixTest (训练集和到目前为止测试集中的对象)，FullTest (训练集和测试集)，SkipTest (训练集)。prior 同上。

选择树结构

贪婪算法,找出所有可能分割方式,计算每种方式的惩罚函数,选择最小的，将结果分配个叶节点。后续叶节点重复此过程。在构建新树前进行随机重排，按梯度下降方向构建新树。

第 3 章 数据预处理及因子筛选

3.1 数据获取

本文通过 Autotrader 的 Python 接口获取 2016 年 1 月至 2018 年 9 月间的沪深 300 与中证 500 成分股的日数据，每月的因子数据为当月每天因子数据的简单平均。沪深 300 以及中证 500 成分股为 2019 年 4 月 8 日时获取到的成分股，每个月每只股票的的收益率为每日收益率的累加，虽然这样的累加方式会丢失开盘跳空对收益的影响，但是该方法得到的月度收益率与真实的月度收益率之间差别可以忽略。

为了保证数据的可得性，本文使用本月的因子数据对下一个月的月度收益率进行建模，并通过十二个月份预测一个月份的方式进行滚动预测。具体的预测时间为 2016 年 1 月至 2018 年 8 月（由于比赛方提供的数据最后一个月份为 2018 年 9 月，无法获取 2018 年 10 月的数据）。

在机器学习领域里有一句名言：数据和特征决定了机器学习的上限，而模型和算法的应用只是让我们逼近这个上限。这个说法形象且深刻的提出前期的数据预处理和特征分析的重要性。在第三章后面的小节里将着重介绍本文的数据预处理过程和因子特征分析过程即候选因子的选取、因子有效性检验及选取和冗余因子的剔除。

3.2 候选因子的选取

股票的超额收益是被各个因子影响着，因此在多因子选股模型的构建中关于候选因子的选取就显得极为重要。候选因子越多，对股票的收益分析就会越全面，构建的模型效果也就会更好，因此候选因子的选取必须考虑足够的维度，同时也要具备合理的经济解释意义。

候选因子应该具备以下三个特点：可得性、普遍性以及区分性。具备可得性的数据更符合量化投资的数据化特性，只有普遍性才能使得所研究的所有股票都符合，具备较强区分性的因子能减少冗余因子的存在进而可以更好地将优质股票

筛选出来。

基于之前的学者对多因子量化选股模型的研究并考虑到数据的可得性、普遍性和区分性，本文从单因子选股对应夏普比率配合手动筛选、Adaptive Lasso 方法、相关性分析和预测重要性四个方面选取候选因子。具体实现方法为：

3.2.1 夏普比率结合手动筛选

表 5 单因子选股回测夏普比率表

因子名称	夏普比率	因子名称	夏普比率
ROE	1.58	BasicEPS	1.13
ARTDays	1.51	EPS	1.13
RevenueTTM	1.45	DilutedEPS	1.13
TRevenueTTM	1.45	SalesExpenseTTM	1.12
ARTRate	1.44	TaxRatio	1.12
ROIC	1.37	NetAssetPS	1.12
CostTTM	1.29	OperatingRevenueGrowRate5Y	1.11
RetainedEarningsPS	1.27	UndividedProfitPS	1.10
TCostTTM	1.24	OperCashInToCurrentLiability	1.09
CTOP	1.24	OperateNetIncome	1.08
OperatingCycle	1.21	AssetImpairLossTTM	1.07
NIAPCut	1.20	NegMktValue	1.07
NetProfitAPTTM	1.18	LFLO	1.07
OperatingProfitPS	1.18	OperateProfitTTM	1.06
ROAEBIT	1.17	TProfitTTM	1.06
DividendPS	1.17	NetProfitTTM	1.06
NIAP	1.15	EBITPS	1.06
RetainedEarnings	1.13	AccountsPayablesTRate	1.04
TEAP	1.13	SurplusReserveFundPS	1.04
GrossProfitTTM	1.13		

首先使用单因子选股对应夏普比率以及配合手动筛选进行因子筛选。通过建

立单因子选股模型，将单因子正序以及倒序排列，从上证 50 中筛选出前 10 只股票进行月初买入月末卖出的交易回测，回测时间段为 2016 年 1 月至 2018 年 9 月。通过筛选出夏普比率大于 1 的因子，因子名称以及对应夏普比率如表 5 所示；然后通过手动筛选，从剩下的因子中累计筛选出具有显著经济意义的因子，从而第一步得到共计 119 个因子。

手动从夏普比率小于 1 的因子中筛选因子的意义在于：首先能够避免过拟合的问题，因为因子本期的有效并不能保证未来的持续有效，尤其是当一个因子被越来越多的交易者使用的时候；其次是对于模型的经济意义解释更方便，因为手动筛选出的因子在经济意义上更具有重要性，使用频率更高，计算方法更成熟。

3.2.2 使用 Adaptive Lasso 方法进行因子筛选

为了进行有效的特征选择，筛选出最具有统计意义的一组特征变量，本文采用 Adaptive-Lasso 变量选择模型进行特征变量的有效性筛选。Adaptive-Lasso 变量选择模型，既能够剔除存在共线性关系的变量，同时也体现出 Adaptive-Lasso 对多特征变量进行变量选择的优势。根据本文各个特征变量的因子系数值，将特征系数为 0 的因子在后续的建模过程中进行删除。通过 LASSO 方法从夏普比例以及手动筛选的因子中累计筛选出 30 个因子，如表 6 所示：

表 6　因子 Adaptive LASSO 系数表

因子名称	LASSO 系数	因子名称	LASSO 系数
AD	9.07E-07	GrossProfitTTM	-1.02E-14
TangibleAToInteBearDebt	2.28E-07	NegMktValue	-2.16E-14
RSTR21	1.84E-07	TEAP	-3.01E-14
cm_ARC	1.11E-07	NetTangibleAssets	-7.96E-14
InterestCover	1.77E-09	OperateProfitTTM	-9.52E-14
NIAPCut	1.21E-12	TCostTTM	-2.18E-13
TProfitTTM	1.19E-12	OperateNetIncome	-3.91E-13
SalesExpenseTTM	4.46E-13	NIAP	-7.94E-13
NetProfitAPTTM	2.24E-13	NetProfitTTM	-1.69E-12
RevenueTTM	1.92E-13	VSTD10	-9.82E-11

RetainedEarnings	1.63E-13	PEG5Y	-3.53E-09
AssetImpairLossTTM	5.46E-14	ARTRate	-4.73E-09
MONEYFLOW20	3.86E-14	ForwardPE	-2.81E-08
CostTTM	3.03E-14	PEG3Y	-3.07E-08
TRevenueTTM	-3.48E-15	PE	-4.18E-08

3.2.3 使用相关性分析筛选

虽然利用 Adaptive-Lasso 筛选出了有效因子，但是不同的选股因子可能由于内在的驱动因素大致相同等原因，所选出的组合在个股构成和收益等方面具有较高的一致性，因此其中的一些因子需要作为冗余因子剔除，只保留同类因子中收益最好、区分度最高的一个因子。冗余因子的剔除过程如下：计算上一步筛选后的有效因子间的相关系数矩阵，一般来说，相关系数取绝对值后，0-0.09 视为没有相关性，0.1-0.3 为弱相关，0.3-0.7 为中等相关，0.7-1.0 为强相关。在本文中相关性阈值取 0.7，相关性大于 0.7 的两个有效因子视为高度相关将进行冗余因子的剔除过程，即如果两个变量的相关性大于 0.7，这时删除因子 Lasso 系数较小的那个变量，保留较大的变量，具体分析时还应该全面考虑各个因子的经济含义。

使用相关性分析计算前两步筛选出的因子，去掉相关性比较强的因子，从而保证在模型预测能力几乎不受影响的前提下，有效的起到降维的作用，提高运算效率和实际应用价值。

变量的相关系数热力图如图 6 所示，由于因子总数较多，因此没有将具体的相关系数显示在图上。通过设置临界值 0.7，筛选出相关性较弱的因子作为建模的因子。

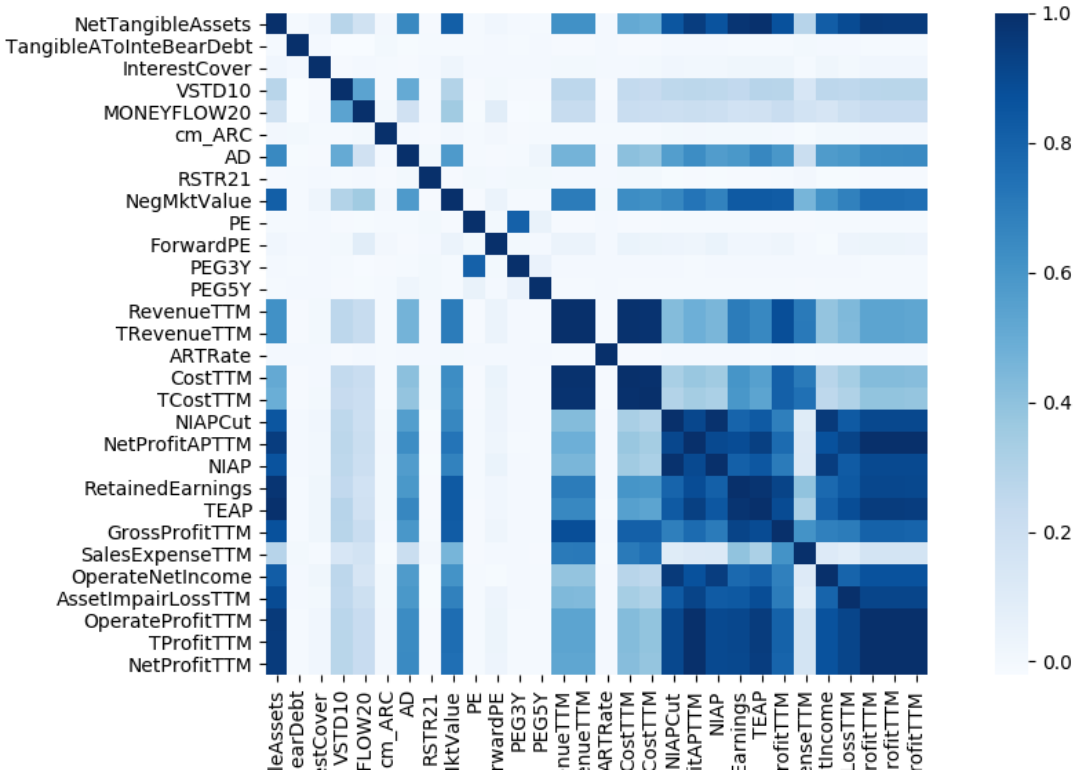


图 5　变量的相关系数热力图

通过相关性筛选，得到最终的 13 个相关性小于 0.7 而且因子与收益之间有较为显著关系的最终因子，结果如图 7 所示：

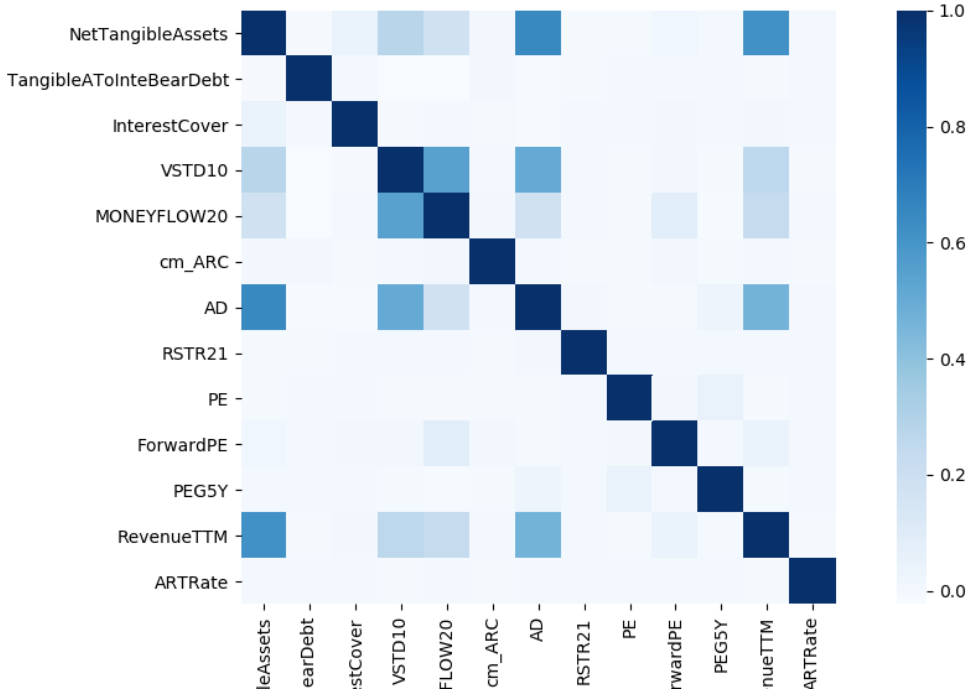


图 6　最终因子的相关系数热力图

通过图 7 可以看出，任意两个因子之间的相关系数均小于 0.7，并且绝大多数因子之间的相关系数小于 0.1，从而认为因子之间的相关性比较弱，又由于这些因子筛选自对于模型较为重要的因子范围，因此有理由认为经过相关系数筛选出的因子为既具有解释能力又不具有较强线性相关的有效因子。从而认为因子筛选起到了在保证预测能力变化不大的前提下有效降维的目的。

3.3 数据预处理

在进行多因子量化选股模型的构建时，大量的原始数据存在着一部分不完整即缺失、不一致等现象，如果没有进行合理的数据清洗步骤会严重地影响到后续模型的构建结果，导致建模结果出现严重偏差，所以建模前的数据清洗过程就显得尤为重要。当数据清洗结束后再进行一系列的数据集成、变换、规约等处理，这个完整的过程就叫做数据预处理。本文进行数据预处理是要提高量化选股训练和测试数据的质量，为了让数据更好地适应后续集成学习算法模型，数据的预处理过程如下：

3.3.1 缺失值处理

缺失值的处理方式对于后续建立模型具有重要的影响。由于本文使用的数据来源与 Autotrader 的日度数据合成而来，从而在缺失值比例上相对较小，通过做缺失值的比例图可以更加客观的分析。通过图 8 可以看出，仅有两个因子的缺失比例超过了 0.2%，从而有理由认为缺失值所占的比例不足以对结果产生显著的影响。为了尽可能控制缺失值对模型的影响，通过均值填充的方式处理缺失值，将会是比较合理的选择。

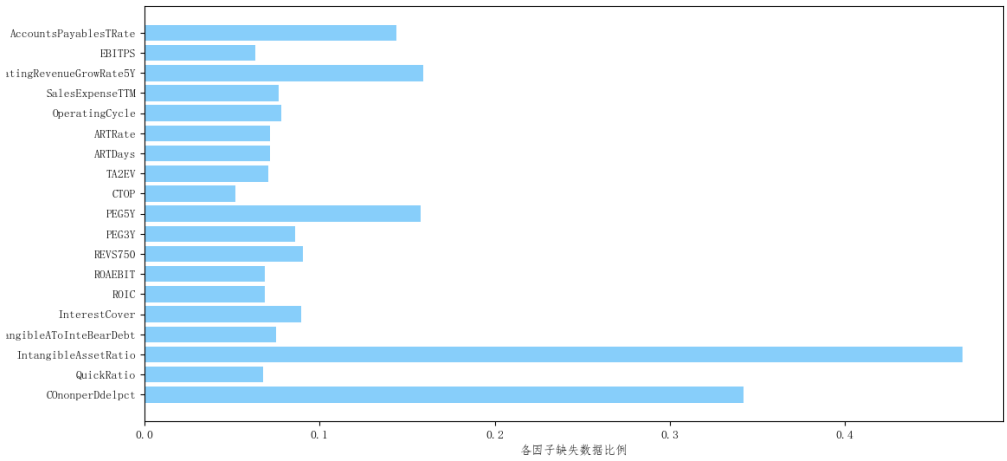


图 7　缺失值比例图

3.3.2 定义预测因变量

股票的超额收益率是指，在一定时期内，股票的收益率超越市场收益率的部分。为了尽可能接近于业界对于超额收益的界定，建模时使用最广泛的沪深 300 指数的收益率作为市场收益率，从而月度超额收益率的计算方式为每只股票的月度收益率减去该月对应的沪深 300 的月度收益率。

多因子选股最终的目的是取得可观的、超越基准指数收益的优质股票，本文将每只股票的月度超额收益率作为衡量模型的 Y 值。通过机器学习模型，对下一期的超额收益率进行预测，预测后排序，然后选择预测的超额收益率最高的一定数量的股票作为交易标的。

第 4 章 基于风格轮动的多因子选股策略研究

4.1 大小盘轮动分析及预测

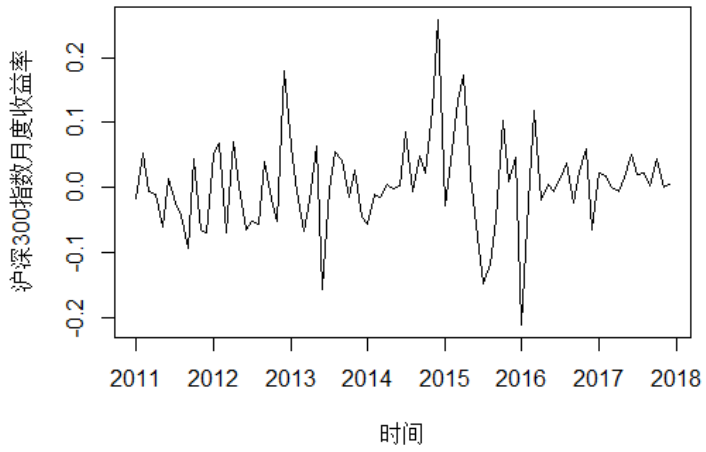


图 8　沪深 300 指数月度收益率时间序列图

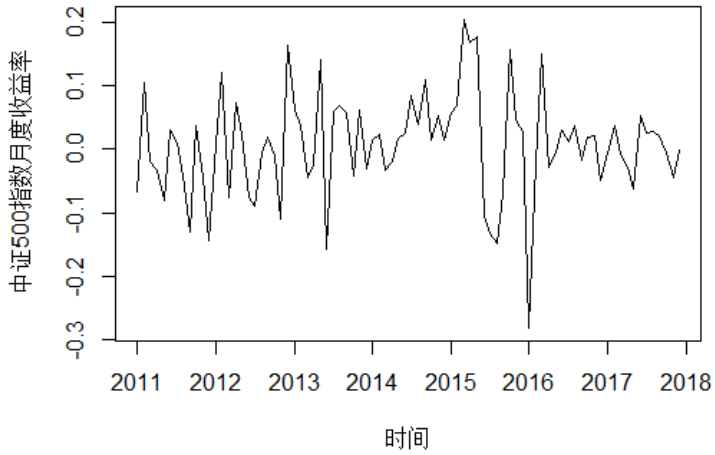


图 9　中证 500 指数月度收益率时间序列图

从图 7 和图 8 沪深 300 指数和中证 500 指数月度收益率时间序列图可以看出，两个股指的月度收益率的波动均非常剧烈，没有表现出明显的上升趋势或者下跌趋势，因此仅仅从两个股指的月度收益率时间序列图无法判断并预测出在下一个时间阶段的收益趋势走向，因此在下文中将引入小盘相对优势指标进行分析。

为了做到较为准确的得到大小盘轮动，要清晰的看出大小盘运动的趋势，而要观察大小盘的运动趋势，就需要先找出一个能反映大小盘优势比较趋势的指标，也就是相对优势指标。从而，相对优势指标的滞后值将会是对于选股模型重要而

且有效的解释变量。

从整体来看，小盘相对优势用同一时间点指标做差来表示。结合前人的研究，筛选出重要的宏观经济变量如表 1 所示，对未来一个月的小盘优势进行预测。宏观经济变量由于各种原因会有少量的缺失值，因此本文使用插值法进行缺失值填充，也即使用该变量上一个非缺失值与下一个非缺失值的平均值进行填充。另外，对于宏观经济变量的选择中，本文主要使用宏观经济指数的同比环比进行数据处理，从而消除季节性影响。

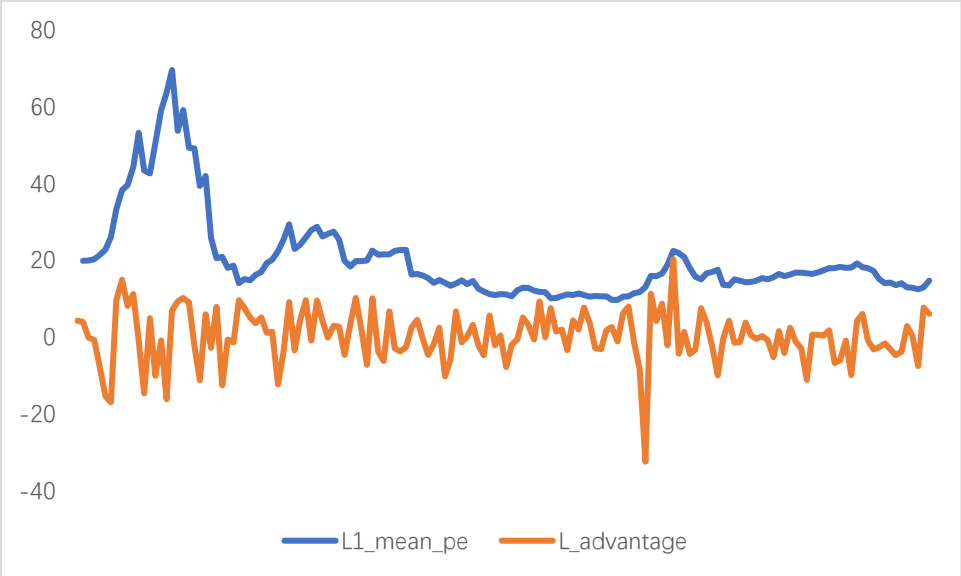


图 10　市场 PE 与小盘优势

市场 PE 的一阶滞后与小盘高于大盘收益率的差额的折线图如图 11 所示，通过图 11 可以看出，市场 PE 与小盘优势之间具有正相关的关系，可以理解为，在估值较高的市场中，股市参与者对市场较为热情，投资倾向于激进，持股周期更短，所以小盘股更可能获取更高的收益。

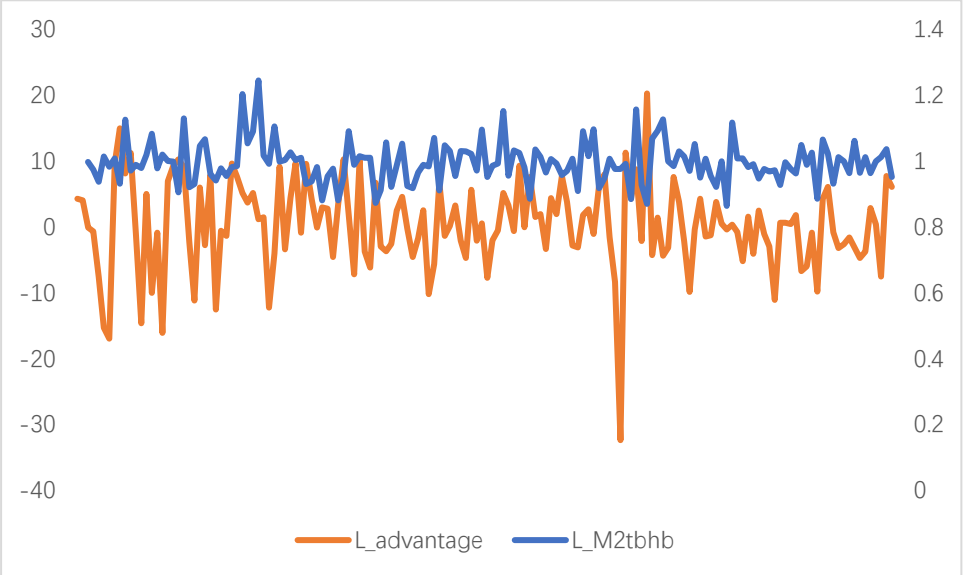


图 11　M2 同比环比与小盘优势

M2 的一阶滞后与小盘高于大盘收益率的差额的折线图如图 12 所示，通过图 12 可以看出，M2 与小盘优势之间具有正相关的关系，可以理解为，在货币发行量较大的市场环境中，投资于无风险产品将获取更低的收益，并且市场的资金总量也增加了，从而导致可能出现小盘股涨幅更大的投机性现象。

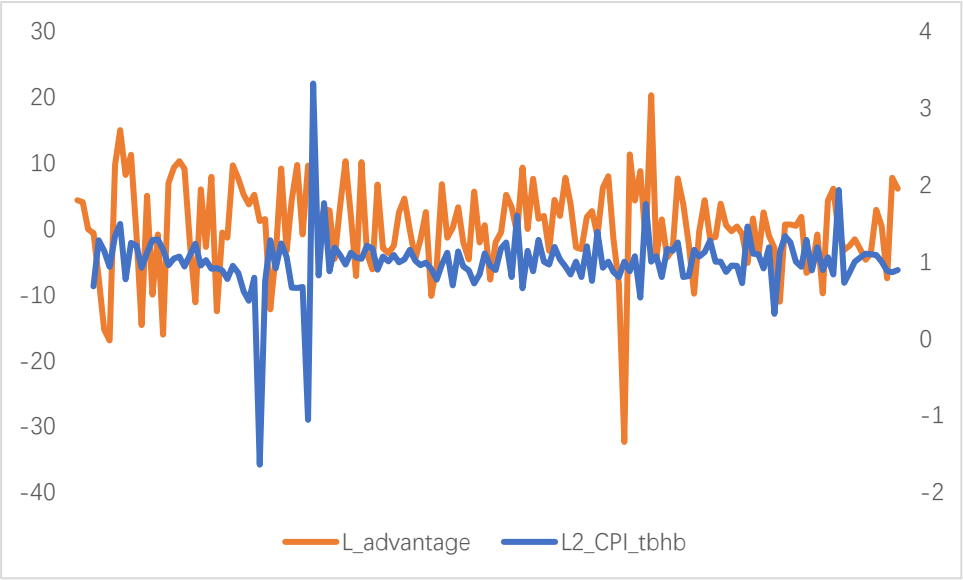


图 12　CPI 二阶滞后与小盘优势

CPI 的二阶滞后与小盘高于大盘收益率的差额的折线图如图 12 所示，通过图 12 可以看出，CPI 的二阶滞后与小盘优势之间具有正相关的关系，可以理解为，在通货膨胀较大的市场环境中，投资于无风险产品将获取更低的实际收益，从而

导致可能出现小盘股涨幅更大。

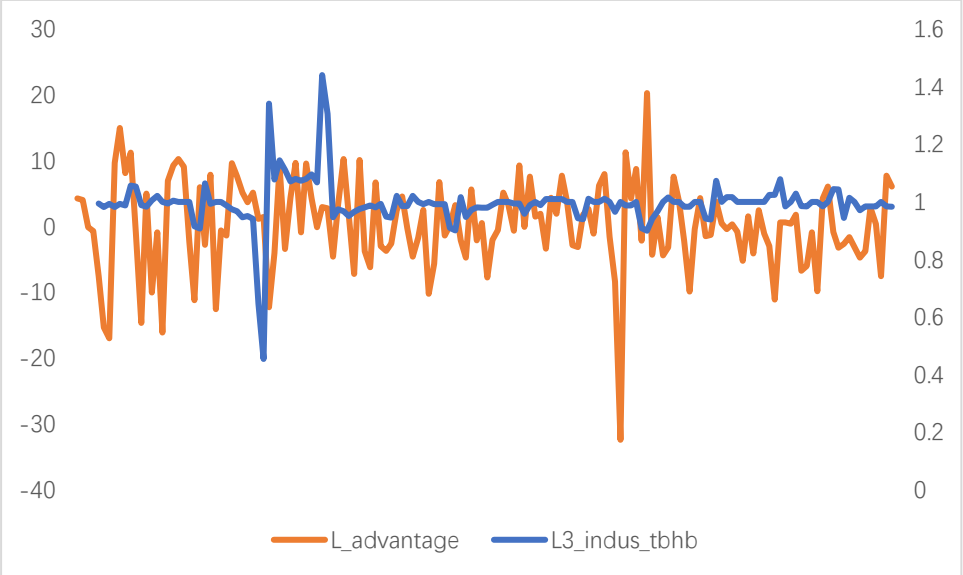


图 13　工业增加值同比环比 3 阶滞后与小盘优势

工业增加值同比环比 3 阶滞后与小盘高于大盘收益率的差额的折线图如图 12 所示，通过图 12 可以看出，工业增加值同比环比 3 阶滞后与小盘优势之间具有负相关的关系，可以理解为，在经济稳健增长的市场环境中，大盘股对应的股票更具有长期持有的价值，从而导致可能出现大盘股涨幅更大。

通过利用表 1 中宏观经济变量以及相对优势变量的滞后值相关变量，将风格轮动因素内生于模型，对下一期的股票超额收益进行建模预测。通过建立模型进行滚动预测，具体使用 12 个月的数据作为模型的训练集，对第 13 个月份进行滚动预测。

4.2 多因子选股模型的构建

在这一节中将选择沪深 300 和中证 500 所有股票在 2016 年 1 月到 2018 年 9 月的样本数据作为总数据，通过 12 个月的数据作为训练集，接下来的第 13 个月份作为测试集，进行滚动预测。本节将分为两个部分，第一部分是 CatBoost、LightGBM、随机森林和 XGBoost 四个模型的建模效果对比，第二部分是根据第一部分确定的最优模型进行多因子选股模型的构建。

4.2.1 CatBoost、LightGBM、随机森林、和 XGBoost 建模效果对比

要进行 CatBoost、LightGBM、随机森林和 XGBoost 这四个模型的建模效果对比，并与普通的等权重多因子模型进行了比较。首先要了解机器学习领域最常用的建模效果指标即回归准确性度量指标：RMSE。

一、RMSE（Root Mean Squard Error）均方根误差

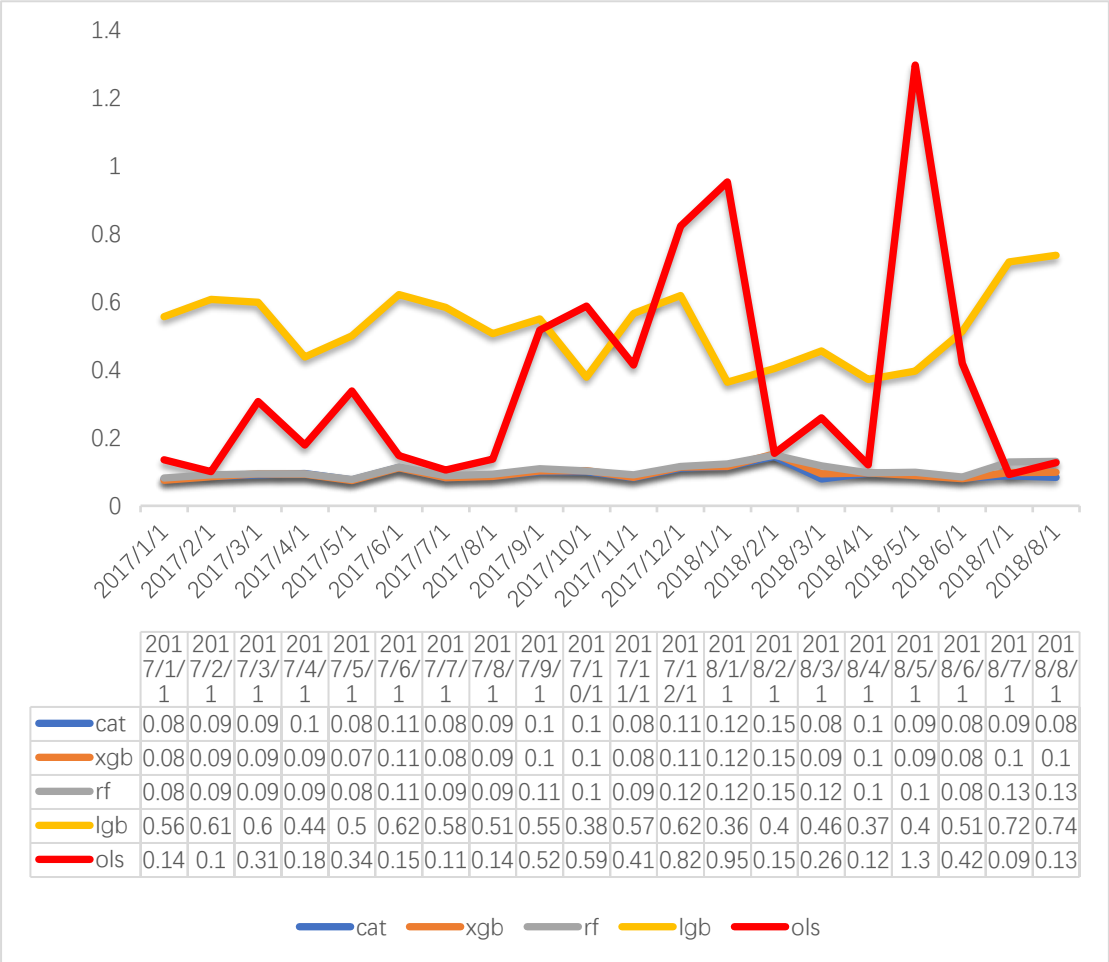
本文所做的是利用模型来进行回归拟合，进行预测的实证分析研究，所以用到的衡量指标也相对直观。回归模型中最常用的评价模型是 RMSE（平方根误差），其定义如下：

$$RMSE=\sqrt{\frac{\sum_{i=0}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{4-1}$$

其中，y_i 表示的是第 i 个样本的真实值，$\hat{y}_i$ 表示的是第个样本的预测值，n 表示的是样本的个数。RMSE 使用的是平均误差，它对于平均值来说它对异常值比较敏感，由于本文数据中几乎不包含异常值，因此 RMSE 对模型的评价在本案例中会较为准确。

二、模型稳定性比较

在模型比较中仅仅只关心单个模型的 RMSE 是远远不够的，还需要长期关注各个模型的稳定性差别，这样才能说明某个算法模型较其他模型的优异性。四个模型的 RMSE 值曲线如下图，其中 X 轴代表时间，Y 轴代表各个算法模型的 RMSE 值。



	2017/1/1	2017/2/1	2017/3/1	2017/4/1	2017/5/1	2017/6/1	2017/7/1	2017/8/1	2017/9/1	2017/10/1	2017/11/1	2017/12/1	2018/1/1	2018/2/1	2018/3/1	2018/4/1	2018/5/1	2018/6/1	2018/7/1	2018/8/1
cat	0.08	0.09	0.09	0.1	0.08	0.11	0.08	0.09	0.1	0.1	0.08	0.11	0.12	0.15	0.08	0.1	0.09	0.08	0.09	0.08
xgb	0.08	0.09	0.09	0.09	0.07	0.11	0.08	0.09	0.1	0.1	0.08	0.11	0.12	0.15	0.09	0.1	0.09	0.08	0.1	0.1
rf	0.08	0.09	0.09	0.09	0.08	0.11	0.09	0.09	0.11	0.1	0.09	0.12	0.12	0.15	0.12	0.1	0.1	0.08	0.13	0.13
lgb	0.56	0.61	0.6	0.44	0.5	0.62	0.58	0.51	0.55	0.38	0.57	0.62	0.36	0.4	0.46	0.37	0.4	0.51	0.72	0.74
ols	0.14	0.1	0.31	0.18	0.34	0.15	0.11	0.14	0.52	0.59	0.41	0.82	0.95	0.15	0.26	0.12	1.3	0.42	0.09	0.13

图 14　模型的稳定性比较图

通过模型的均方根误差（RMSE，图 15）可以发现，Catboost、XGBoost、随机森林模型明显优于 OLS 模型，从而可以认为机器学习模型提升了因子对于预测的表现。另外机器学习模型在模型的稳定性方面相较具有很强的优势，从而可以推断机器学习模型将会在选股方面相对于等权重线性模型有巨大的优势。

4.2.2 基于机器学习模型的多因子选股模型的构建

一、模型参数调优

机器学习模型参数众多，为了提高模型的精度，同时提升模型的泛化能力，即在模型过拟合和欠拟合中寻求最优，调参过程不可缺少。

对于基于决策树的模型，调参的方法一般都需要如下步骤：首先选择较高的学习率，大概 0.1 附近，加快收敛的速度；其次，对决策树基本参数调参；最后，

降低学习率，提高准确率。

网格搜索亦被称为穷举法，其主要思路为将需要进行调整的参数在一定范围内划分为网线，一个交点对应一对参数组合，其中有可行点、不可行点，且在可行点之中又存在优劣之分。网格搜索对所有的参数组合进行搜索、训练，并依照模型的评价标准给出最优目标函数所在的交点，该点所对应的各参数的值即为最优参数值。此法可以同时搜索需要的多个参数值，以获取评价函数达到最优的参数组合。

结合模型待调参数来看，该算法框架中所涉及参数大多存在一定的取值范围，多在 0~1 之间，且从实际意义来看，参数的选取实质上是离散而非连续值，故而采用网格搜索法参数寻优较为合适。

本文需要调整的参数设定为八个，模型优化评价指标设为均方根误差。

通过各参数之间的组合，能够对模型的精度提高和过拟合进行有效的控制，调参方案如下：

第一步：先固定每个参数的初始值，再调整最优化。其中设置的初始值见各算法框架参数结果详情表。

第二步：对 max_depth 和 num_leaves 进行参数调优。调参方法如下：首先以 max_depth 的初始值 6 为基准，前后步长设置为 1，取出两个参数值 5 和 7，即该参数的可选参数值设为 5，6，7 三个选项。接下来，因为 LightGBM 使用的是 leaf-wise 算法，因此在调节树的复杂程度时，使用的是 num_leave 而不是 max_depth。大致换算关系：num_leaves = 2^(max_depth)，但是它的值设置应该小于 2^(max_depth)，否则可能会导致过拟合，分别对应设置三个参数 30，50，80。最后，将 max_depth 和 num_leave 的参数值进行组合，利用网格搜索得到最优解，当 max_depth 和 num_leave 分别取 7 和 80 时为最优解，最优解分数达到-1.860,转化为均方误差为 1.3638。

第三步:对 min_data_in_leaf 和 min_sum_hessian_in_leaf 进行参数调优。这一步中，首先用 max_depth 和 num_leaves 的最优参数替换初始参数，其他参数依旧为初始参数，接下来的步骤与第二步一致，min_data_in_leaf 以 20 为初始值，一般的调参幅度为 2，min_sum_hessian_in_leaf 以 0.001 为初始值，其余值设置为 0.002，0.003；最后得出结果当 min_data_in_leaf 和

min_sum_hessian_in_leaf 依旧为 20 和 0.001 时候达到最优解，即这里调参之后，优化分数不变。

第四步：对 feature_fraction 和 bagging_fraction 进行参数调优。方法同上，保持前两步的已调优参数不变，剩余参数依旧保持初始值，一般对 feature_fraction 和 bagging_fraction 这两个参数的前后调节幅度为 0.1，由于这两个参数初始值均为 0.8，故选择备选参数均为：0.7，0.8，0.9，当 feature_fraction 和 bagging_fraction 为 0.7 和 0.7 时，最优解分数达到-1.8541,转化为均方误差为 1.3618，均方误差有所下降。

第五步： 对 learning_rate 和 num_boost_round 进行参数调优。使用较高的学习速率是因为可以让收敛更快，现在降低学习率。由于学习速率的常见值范围在 0.01 到 0.1 之间，本文选择其他备选参数为 0.01，同时设定当连续 500 次迭代不能得到最优的结果时停止迭代，并以最优的结果所在轮次作为本次训练的最优迭代次数即最优次数为 266 次。并且均方误差又有所下降为 1.3527。

在进行以上五步的调参过程后，LightGBM 算法框架模型的参数调整汇总结果详情如下表所示：

表 7 LightGBM 算法框架调参结果详情表

参数名称	参数简称	参数含义	初始值	调参结果
学习率	learning_rate	为了加快收敛的速度，选择较高的学习率，大概 0.1 附近	0.1	0.01
迭代次数	num_boost_round	boosting 的迭代次数	500	266
最大深度	max_depth	树的最大深度。这个值是用来避免过拟合的。max_depth 越大，模型会学到更具体更局部的样本,但深度越大可能过拟合。	6	7
叶子数目	num_leaves	由于 LightGBM 使用的是 leaf-wise 策略生长算法，调节树的复杂程度	50	80

叶子节点中最小的数据量	min_data_in_leaf	调大可以防止过拟合，值取决于训练数据的样本个数和叶子数目。设置较大可避免生成一个过深的树，但有可能导致欠拟合。	20	20
最小叶子节点中的样本权重和	min_sum_hessian_in_leaf	某观测叶子节点中所有样本权重之和的最小值，用于防止过拟合问题：较大的值能防止过拟合，过大的值会导致欠拟合问题	0.001	0.001
样本采用比例	bagging_fraction	这个参数控制每棵树随机采样的比例。减小这个参数的值，算法就会更加保守，避免过拟合。但是，如果这个值设置的过小，它可能会导致欠拟合。	0.8	0.7
特征采用比例	feature_fraction	每棵树随机选取特征的比例，进行特征子抽样，防止过拟合，提高训练速度	0.8	0.7

二、多因子选股模型的构建

本文通过对机器学习模型确定最优参数之后，进行滚动训练和预测，最终得到基于机器学习算法框架的多因子选股模型，本文以表现最优的 CatBoost 模型作为基本的结果解释模型。

1、各个特征对模型预测结果的贡献度

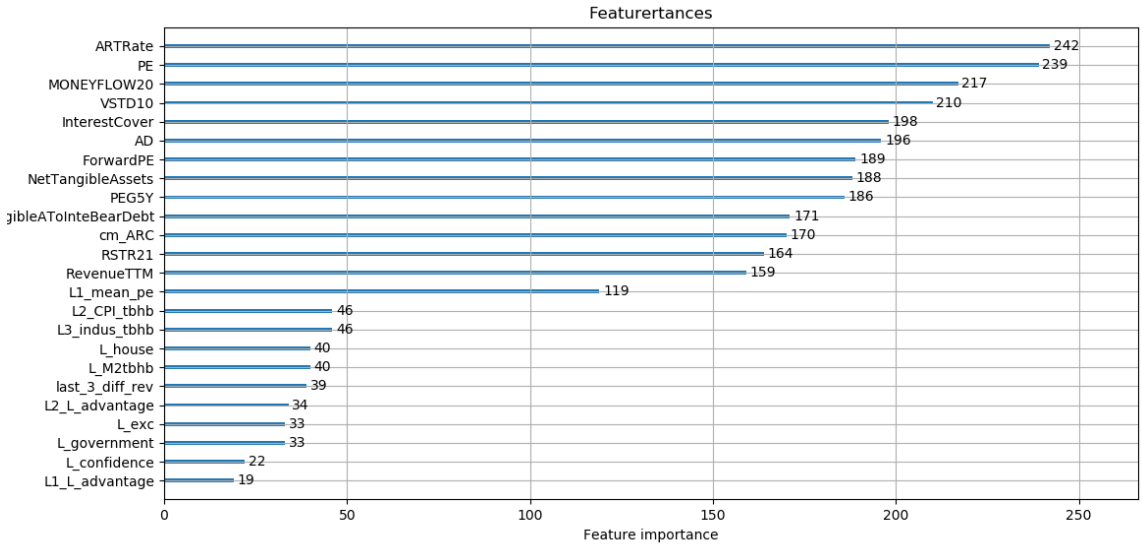


图 15　特征重要性输出图

各个特征对模型预测结果的贡献度从高到低的排序可以看出，排名前十三的特征变量的贡献度相对较高，对建模的影响较大。 ARTRate、PE、MONEYFLOW20、VSTD10 、 InterestCover 、 AD 、 ForwardPE 、 NetTangibleAssets 、 PEG5Y 、 TangibleAToInteBearDebt、cm_ARC、RSTR21、RevenueTTM 分别代表的是：应收账款周转率（质量类）、市盈率（价值类）、20 日资金流量（情绪类）、10 日成交量标准差（情绪类）、利息保障倍数（质量类）、累积/派发线（动量类）、动态 PE（价值类）、有形净资产/有形资产净值（基础科目衍生类）、市盈率/归属于母公司所有者净利润 5 年复合增长率（价值类）、有形净值/带息负债（质量类）、筹码分布的均值因子（情绪类）、21 天相对强势（动量类）、营业收入（基础科目衍生类）。下面将从这些特征的经济含义和意义来一一分析这十个重要的特征变量是如何影响本文的量化选股的。

贡献率排名第一的是应收账款周转率，属于质量类指标，对于预测下一月的超额收益率而言，质量类指标发挥了重要的作用，不仅贡献率最高的因子属于质量类，而且质量类因子在经过筛选后保留的也相对比较多。另外，价值类因子对于预测下一月的超额收益率也发挥了重要作用，通过因子筛选共筛选出 3 个价值类的因子放入模型，数量相对较多，而且比较符合经济学意义。

对于股票超额收益率更高的公司而言，长期的公司背景应更加符合市场未来的需求，短期应具备价格快速波动的基础条件。对于一个长期的发展而言，质量类的指标反映了公司的价值的大小，这决定了投资者对于未来的信心；对于短期

快速波动而言，可以反映短期快速波动潜力的指标是价值类、情绪类、动量类指标，短期内能够快速波动的股票往往是相对低估值、市场情绪较为高涨、具有了开始波动的初始动能的股票。

另外不可忽视的是宏观经济指标，也就是内生化的风格轮动变量，对于预测股票未来超额收益率的重要作用。宏观经济变量由于相对于公司的指标变化较慢，因此单纯从预测重要性的角度看将不如指标更有重要性。反映在重要性方面，可以看到虽然宏观经济变量重要性在数值上不如指标，但是又没有数量级上的很大差距。但是由于宏观经济对于整个市场的潜移默化的影响，市场环境会逐步发生变化，为了适应市场的变化，考虑到宏观经济变量的模型会在预测未来超额收益率方面具有更好的样本外表现。由于宏观经济数据具有季节性的特点，本文使用的宏观经济数据为环比同比数据为主，另外考虑到环比同比不一定能够准确解释季度差异，本文通过建立 12 个月份预测未来 1 个月份的超额收益率解决宏观经济数据具有季度差异的问题。通过模拟回测，验证了我们在考虑风格轮动后会得到更高收益的结论。

4.3 投资组合的构建与绩效评价

4.3.1 投资组合的构建

本文基于训练样本数据集训练机器学习模型，建立动态的量化选股算法模型，之后将测试数据集的特征带入训练好的模型，进而预测该股股票的超额收益率，超额收益率预测值越大，那么代表该股票成为强势股的可能性也就越大，所以本文股票组合的选取主要是根据预测股票超额收益率大小。筛选股票的另一个重要的因素就是股票组合中的股票数目。在国外一般基金的选股个数控制在 60 只上下，在本文为了更好的分散风险，也选用 60 只股票作为最终的股票组合个数，同时采用的投资方式为等权重投资。

本文选股的方式为先利用机器学习算法框架模型对回测样本进行超额收益率的预测，再将所有中证 500 和沪深 300 成分股的预测超额收益率进行降序排序，并按照等权重的投资方式选取前 60 只股票，作为选出的最终股票组合。

4.3.2 投资组合的绩效评价理论

当确立投资组合后，可以通过对比收益率以及风险率来对选股组合在回测时间段的表现作出判断，进行评价。重点就是要平衡风险和收益的关系，在收益率不变的情况下，要追求风险最小；在已知固定风险的情形下，要努力做到收益最大化。

对于一个股票组合的评价来说主要分为两个部分，分别为：收益率指标，包括总收益率、年化收益率以及年化超额收益率和风险度量指标，包括夏普比率和信息比率。

一、收益率指标

（1）总收益率

总收益率是指在一定的投资时期范围内，期末所获得的收益总和与期初投资净值的比率。总收益率的公式如下：

$$\mathrm{R}=\frac{N_2-N_1}{N_1}\times 100\% \qquad (4\text{-}2)$$

其中：N_1代表期初投资净资本，N_2代表期末投资净资本

（2）年化收益率

总收益从整体的角度出发，在相对较长的投资时间中，总收益往往会在数值上随着投资时间的增加而增加，并没有总指导的意义。但将其标准化，即年化收益率，虽然是一种理论收益率，不是真实的收益率，但是一个很好的可以横向比较的收益指标。

由于本文的数据均为月度数据，即年投资时期为 12 个月度，每个月度对应的收益率为R_i，i=1,…,12

年化收益率的公式如下：

$$R_{year}=\frac{\sum_{i=1}^{12}R_i}{n}\times 12 \qquad (4\text{-}3)$$

（3）年化超额收益率

年化超额收益率综合了上述两个收益率的优点，既体现了年度性，又剔除了市场的影响，即剔除考察期中证 500 指数的年化收益率。在本文中年化超额收益率的计算公式如下：

$$R_{over}=R_{year}-R_{300} \qquad (4\text{-}4)$$

R_{year}代表投资组合的年化收益率，R_{300}代表沪深 300 指数的年化收益率，两个指标的时期一致。

二、风险度量指标

（1）夏普比率

夏普比率比较的是投资组合的收益率与市场无风险收益率。他们之间的差值为正值时表明投资组合的收益率超过了市场无风险收益率。说明这种投资策略要好于银行存款，值得投资，因此夏普比率越高，说明单位风险内收益越高，该投资组合更优质。在本文中，夏普比率指标以国债收益率为基准。

夏普比率公式如下：

$$S_p = \frac{\overline{r_p} - r_f}{\sigma_p} \tag{4-5}$$

其中，$\overline{r_p}$代表已做平均化处理的收益率(总收益率)，r_f代表市场无风险利率，σ_p代表各收益率的标准差。

（2）信息比率

信息比率用来衡量超额风险带来的超额收益，比率越高说明超额收益越高。

信息比率计算公式：

$$\mathrm{IR} = \frac{R - R_{sh}}{\sigma} \tag{4-6}$$

其中：$R - R_{sh}$代表超额收益率，R 代表资产组合收益率，R_{sh}代表基准收益率，在本文为沪深 300 指数收益率，σ 代表主动风险，一般用超额收益率的标准差即可衡量主动风险。

4.3.3 不同模型的绩效评价对比

当确立投资组合后，可以通过对比收益率以及风险率来对选股组合在回测期间的表现作出判断，进行评价。重点就是要平衡风险和收益的关系，在收益率不变的情况下，要追求风险最小；在已知固定风险的情形下，要努力做到收益最大化。

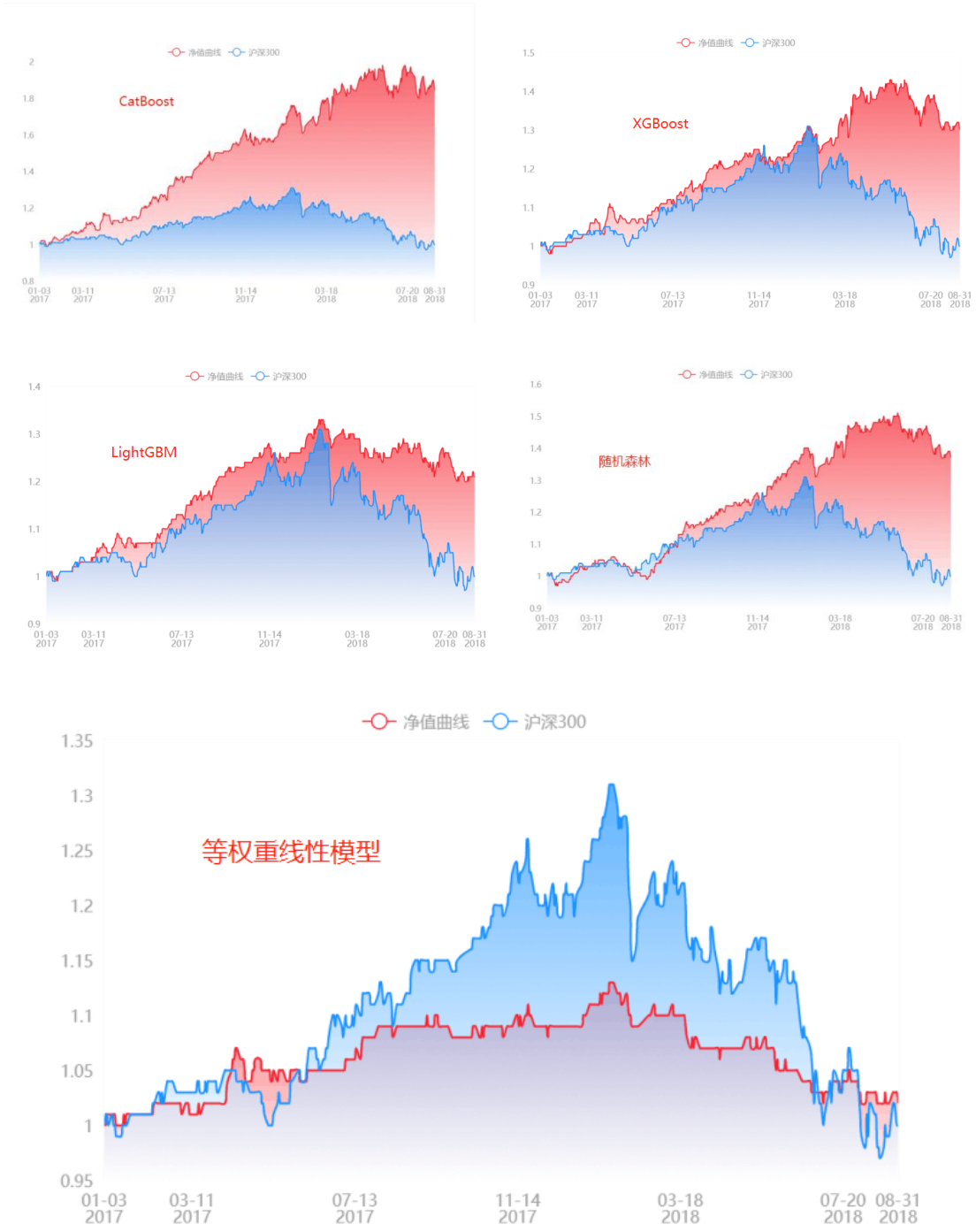


图 16　回测净值曲线对比

图 16 为不同模型回测对应的净值曲线，通过对比可以看出，CatBoost 模型对应的净值曲线相对其他模型更加平稳，并且获取到了更高的收益。为了进行客观的评判，接下来将采用上一节介绍的评判标准对不同模型进行客观的对比。

（1） 年化超额收益率

年化超额收益率在业界由于其代表性而被广泛采用，因此本文将对年化超额收益率进行模型间对比分析。年化超额收益率的对比柱状图如图 17 所示。

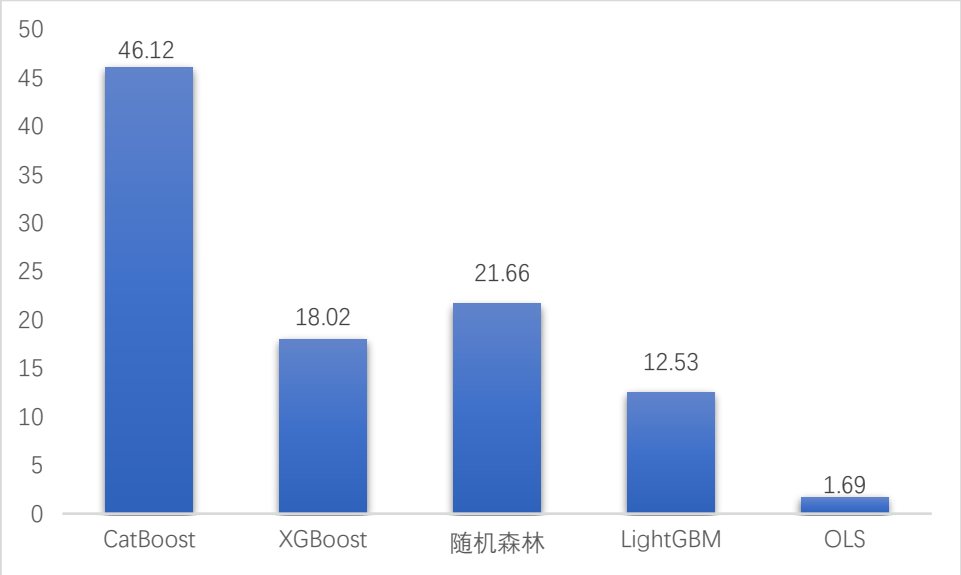


图 17　年化超额收益率对比

通过年化超额收益率的柱状图可以发现，各个利用本文筛选出的因子进行选股的模型均可以取得大于 0 的超额收益率，机器学习模型的超额收益率显著高于等权重线性模型的超额收益率，CatBoost 机器学习模型的效果又显著优于其他模型。

由于市场行情会受到很多事件的冲击，市场的代表性收益率将能够较为客观的反映其他事件对于整个股票市场的冲击，本文使用的超额收益率进行建模可以排除整体性事件的影响，并且超额收益率也是一个较好的反映基金的资金管理能力的指标。

对于实际的应用场景而言，仅考虑年化超额收益率是远远不够的，在一定风险下获取更高收益的基金才能够获得投资者的青睐，为了满足实际需求，本文也将充分考虑风险与收益的权衡。

（2） 风险收益权衡下的综合评判

风险收益权衡下的评判标准主要包括：夏普比率、信息比率、阿尔法、贝塔。综合各个指标的模型间对比如图 18 所示，可以看出，CatBoost 模型在综合各个指标方面均处于最优的状态。从而可以认为，CatBoost 模型在取得最大超额收益的同时兼顾了风险。

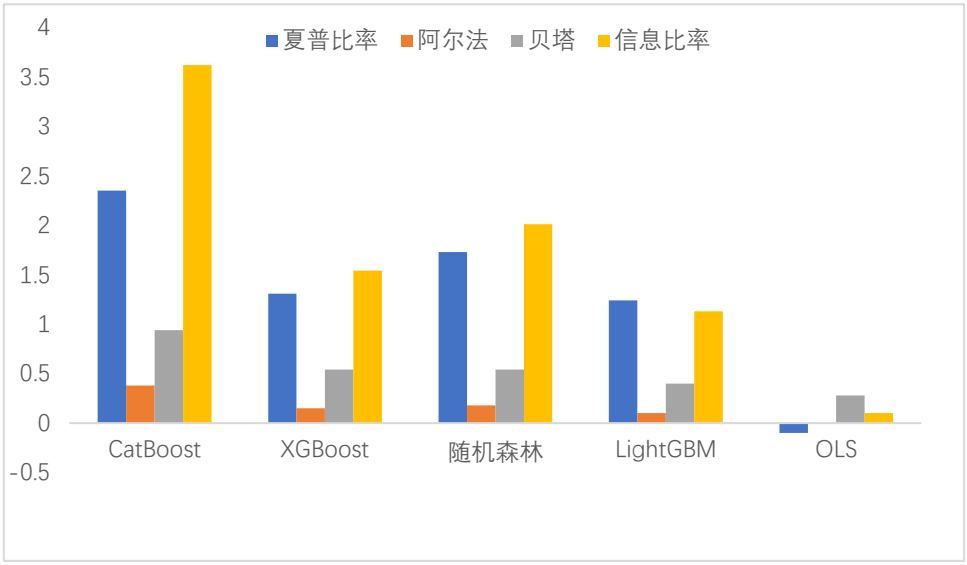


图 18　风险权衡下的模型对比

第 5 章 总结与后续研究建议

5.1 全文总结

本文通过结合内生化大小盘风格轮动策略和机器学习算法进行多因子量化选股，主要目的旨在选出优质的股票组合，最终得到以下结论：

一、通过选取沪深 300、中证 500 的月度收益率数据，进行月度收益的时间序列直观描述分析，但是收益率的波动剧烈且运行趋势不明显，进而构建小盘相对优势指标并使用小盘优势指标的滞后值和宏观经济数据滞后值作为内生化的风格轮动解释变量，放入机器学习模型预测个股未来一个月的超额收益率。通过建模时对变量重要性计算，可知宏观经济变量对于模型更好的预测能力具有重要的作用。

二、在进行 CatBoost、LightGBM、随机森林和 XGBoost 四个机器学习模型的建模效果对比以及与等权重线性模型进行对比时可以发现，由预测能力指标 RMSE 来看，CatBoost、XGBoost、随机森林更胜一筹，LightGBM 次之，等权重线性模型预测能力较差且不够稳定，从而初步断定机器学习模型在选股方面会更有优势。

三、对机器学习多因子选股模型的准确性及泛化能力评价，第一，准确性评价，当本文的多因子选股模型的参数调至最优的时候，RMSE 绝对数值很低，说明本文选用的模型准确率是极佳的；第二，泛化能力评价，经过不同测试集的 RMSE 对比可知，RMSE 波动性较小，始终处于较低水平，说明该模型具备较强的泛化能力。

四、在进行投资组合的绩效评价时得出以下结论：CatBoost 模型的年化收益率远超过基准指标沪深 300 指数的收益率，说明本文采用的基于机器学习算法尤其是 CatBoost 的多因子选股模型并结合大小盘风格轮动策略的投资方案的收益是十分可观的。与此同时，在风险权衡评价指标如夏普比率等指标的对比下，CatBoost 依然具有绝对的优势。

综上所述，本文所提出的基于内生风格轮动和机器学习算法的多因子选股投资策略可以作为投资者的一种投资策略。

5.2 后续研究建议

本文的多因子选股投资策略研究虽然在所选取的样本上进行预测得到了较好的结果，但是仍然存在研究的不足之处，具体如下：

一、本文是以沪深 300、中证 500 分别代表大盘股指和小盘股指，然而在股票市场中代表大小盘的股指有很多，只选用了一组股指代表大小盘显得单一，可以尝试引入另一组股指进行对比分析研究。

二、在投资组合的构建中，本文为了分散投资风险，选择股票的时候只考虑等权重模型，可以尝试引入变权重模型，进行对比，从而量化选出最优的投资组合模型。

三、对于机器学习算法的应用上，可以尝试编制或使用现阶段通过 AT 不能获取的新的因子，进行建模训练，进一步提升建模的效果。

参考文献

[1] 丁鹏.量化投资[M].电子工业出版社.2012.

[2] Rishi K.Narang.郭剑光译.打开量化投资的黑箱[M]北京：机械工业出版社，2012.2,15-17.

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

[4] William F.Sharpe.Capital Asset Prices:A Theory of Market Equilibrium under Conditions of Risk [J].Journal of Finance 1964.19:425-442.

[5] Lintner J. The valuation of risky assets. The selection of risky investment in stock portfolios capital budgets. Reviews of Economics. Statistics. 1965.47:13-47.

[6] Mossin J.Equilibrium in a capital asset market. Econometrica.1966.

[7] Fama,E.F. and French, K.P.,1992,The cross-section of expected stock returns[J], Journal ofFinance,(47),427-465.

[8] Asness C S.The interaction of value and momentum strategies[J].Financial Analysts Journal(March/April,1997,(2):29--39).

[9] Mohanram P S.Separating Winners from Losers among Low Book-to-Market Stocks using Financial Statement Analysis [J]. Social Science Electronic Publishing, 2004,10(2-3):133-170.

[10] Tortoriello R. Quantitative strategies for achieving alpha [M]. McGraw Hill, 2009.

[11] 潘凡.基于有效多因子的多因子选股模型[R]. 安信证券研究报告，2011,1.

[12] 刘毅.因子选股模型在中国市场的实证研究[D]. 复旦大学，2012.

[13] 孙守坤. 基于沪深 300 的量化选股模型实证分析[D].复旦大学,2013.

[14] 多因子选股模型在中国股票市场的实证分析[J]. 王昭栋.山东大学.2014(02).

[15] 张晨宇.基于数据挖掘技术的 A 股市场选股系统设计与实现[J].信息技术与信息化,2017(05):23-27.

[16] 李文星,李俊琪.基于多因子选股的半监督核聚类算法改进研究[J].统计与信息论坛,2018,33(03):30-36.

[17] 陈满祥,吴冕,吴昊,李雯.基于财务驱动因子的 logistic 预测选股模型[J].经贸实践,2018(12):161+163.

[18] 曹源. A 股风格轮动规律探寻[R].国都证券，2010.

[19] 杨勇.基于价格和成交量的大小盘风格轮动研究[R].上海:国金证券, 2015.

[20] 曹力.大小盘轮动策略研究[R].华泰联合证券，2010(2).

[21] Ntoutsi I，Kalousis A，Theodoridis Y.A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees.SDM，2008: 810—821

[22] Perner P. How to compare and interpret two learnt Decision Trees from the same Domain? 2013 27th International Conference on． AdvancedInformation Networking and Applications Workshops(WAINA).New York: IEEE，2013: 318—322

[23] Perner P.How to interpret decision trees? Industrial Conference on Data Mining. Berlin: Springer，2011: 40—55

[24] 王日升,谢红薇,安建成.基于分类精度和相关性的随机森林算法改进[J].科学技术与工程,2017,17(20):67-72.

[25] Rokach L. Decision forest: twenty years of research. Information Fusion，2016; 27: 111—125

[26] 魏静靓.风格投资——大小盘风格轮动策略实证分析[J].中国证券期货,2012(04):12-13.

[27] 谢晓闻,方意,梁璐璐.人民币汇率对大小盘股指影响的异质性及成因研究[J].中南财经政法大学学报,2013(5).

[28] 王敬,刘阳.证券投资基金投资风格:保持还是改变？[J].金融研究,2007(8):120-130.

[29] 尹向飞,张曦.关于中国股市风格的实证研究[J].统计与决策, 2009(13):89-91.

[30] 黄付生,吴启权.基于经济周期的大小盘股票风格研究[J].北京理工大学学报, 2010,12(5):23-26.

[31] 肖峻.关于中国股市中风格投资与风格动量的研究[J]. 经济科学, 2006(6):49-57.

[32] 李颖.陈方正.风格投资理论研究[J].经济社会体制比较, 2005(5).

[33] 周志华.机器学习:= Machine learning[M].清华大学出版社.2016.

[34] 加雷斯·詹姆斯,丹妮拉·威滕,特雷弗·哈斯帖.统计学习导论——基于 R 应用[M].机械工业出版社.2017,12.

[35] 李航.统计学习方法[M].清华大学出版社.2012.

[36] 曹正凤，纪宏，谢邦昌.使用随机森林算法实现优质股票的选择[J]. 首都经济贸易大学学报，2014（2）：21-27.

[37] Ming-Chieh Wang, Ming Fang and Jin-Kui Ye, Financial Integration of Large- and Small-Cap Stocks in Emerging Markets, 2013, (49):17-31.

[38] Switzer, The Behavior of Small Cap vs. Large Cap Stocks in Recessions and Recoveries: Empirical Evidence For the United States and Canada [J]. North American Journal of Economics and Finance,2010, (21):332-346.

[39] Chen, H, De Bundt, M, Style Momentum within the S&P500 index [J]. Journal of Empirical Finance,2004, (11): 483-507.

[40] 徐永春. 基于ＳＶＭ技术的套期保值模型的实证分析⑴. 统计与决策313(16).

[41] 杨喻钦.基于Alpha策略的量化投资研究[J] .中国市场. 2015(25): 83-84.

[42]王淑燕，曹正凤，陈铭芷.随机森林在量化选股中的应用研究[J].运筹与管理.2016(03):163-168.

[43] 李姝锦，胡晓旭，王聪.浅析基于大数据的多因子量化选股策略[J]. 经济研究导刊.2016(17): 106.

[44] 何亚莉.论量化投资对中国资本市场的影响[J].现代商贸工业. 2016(19): 120-121.

[45] 董素娟.国内量化产品分类及现状[J].新经济. 2016(06): 43-44.

[46] 陈健，宋文达.量化投资的特点、策略和发展研究[J].时代金融. 2016(29): 245-247.

[47] 凌士勤, 石川. 基于多因子选股的 Alpha 策略设计[J]. 商情, 2016（29）.

[48] 王辰.投资基金风格轮动策略分析[J]. 现代经济信息, 2014(19):321-321.

[49] 范青亮,王婷.企业并购、人力资本与风险溢价—以沪深300指数成分股公司为例[J].中国经济问题,2016(2):82-98.

[50] 赵庄,侯晓丽,郑丰.基于财务指标的选股系统设计与实现[J].农业网络信息,2012(10):57-58.

[51] Ando T, Bai J. Panel Data Models with Grouped Factor Structure Under Unknown Group Membership[J]. Journal of Applied Econometrics,2016,31(1).

[52] 徐慧丽.基于随机森林的多阶段集成学习方法[J].高师理科学刊,2018(2).

[53] 曹正凤.随机森林算法优化研究[D].首都经济贸易大学,2014.

[54] 张潇,韦增欣.随机森林在股票趋势预测中的应用[J].中国管理信息化,2018(3):120