学校代码：10730　　密级：公开
分类号：O213

兰州大学

硕 士 学 位 论 文

（ 专 业 学 位 ）

论文题目（中文）　机器学习在实证资产定价中的应用
——以中国股票市场为例

论文题目（英文）　**Application of Machine Learning in Empirical Asset Pricing:Case Study of China Stock Market**

作 者 姓 名　焦清泽

类 型 领 域　专业学位类型 • 应用统计

研 究 方 向　金融大数据

教 育 类 型　学历教育

指 导 教 师　严定琪　副教授

合 作 导 师

论文工作时段　2021 年 7 月　至　2023 年 3 月

论文答辩日期　2023 年 5 月

校址: 甘肃省兰州市城关区天水南路 222 号

机器学习在实证资产定价中的应用——以中国股票市场为例

中文摘要

资产定价的基本目标为衡量风险溢价, 将机器学习模型与实证资产定价相结合, 极大的的改善了传统实证资产定价模型对于风险溢价衡量的准确性. 本文以国内 A 股市场为例, 通过将影响股票收益率的特征引入机器学习算法, 在对比了 R^2_{OOS} 之后, 证实了神经网络模型、树形模型以及降维线性模型在各个股票集合中预测收益率的能力均强于传统线性模型, 同时证实了在国内股票市场中, 国有企业、小盘股以及散户投资者是国内 A 股市场出现异象的主要原因所在. 这些异象导致了在 A 股中, 中小板市场的股票反而更具可预测性, 创业板市场的股票可预测性较弱. 我们运用机器学习模型对影响股票收益率的特征因素进行进一步挖掘, 筛选出在不同股票集合中对股票收益率贡献较大的个股因子与宏观经济因子, 证实了流动性因子与动量因子是影响股票收益率的关键因子, 并且发现神经网络模型对个股因子识别的准确度较高. 证实了宏观经济因子对股票的影响与股票集合的划分无关, 在以所有股票样本为集合的数据集中, 通货膨胀因子、股权扩张因子以及信用利差因子是对股票收益率贡献较大的宏观经济因子.

关键词：实证资产定价；股票；机器学习；风险溢价

APPLICATION OF MACHINE LEARNING IN EMPIRICALASSET PRICING:CASE STUDY OF CHINA STOCK MARKET

Abstract

The basic goal of asset pricing is to measure the risk premium. Combining machine learning models with empirical asset pricing has greatly improved the accuracy of traditional empirical asset pricing models for risk premium measurement. In this dissertation, taking the domestic A-share market as an example, by introducing the characteristics that affect the stock return rate into the machine learning algorithm, after comparing R^2_{OOS}, it is confirmed that the neural network model, tree model, and reduced dimension linear model are better than traditional linear models in predicting returns in various stock sets. At the same time, it is confirmed that Chinese state-owned enterprises, Small-cap stocks and retail investors in the domestic market are the reasons for the anomalies in the A-share market. These anomalies lead to stronger predictability of stocks in the small and medium enterprise board market, and weaker predictability of stocks in the ChiNext market. We use the machine learning model to mine the characteristic factors that affect the stock return rate, and screen out the individual stock factors and macroeconomic factors that contribute greatly to the stock return rate in different stock collections, and confirm that the liquidity factor and the momentum factor are the key factors that affect the stock return rate, and found that the neural network model has a high accuracy in identifying individual stock factors. It has been confirmed that the impact of macroeconomic factors on stocks is independent of the division of the stock set. In the dataset with all stock samples as the set, inflation factors, equity expansion factors and credit spread factors are the macroeconomic factors that contribute significantly to stock returns.

Keywords: Empirical Asset Pricing; Stock; Machine Learning; Risk Premiums

目 录

第一章　绪论

1.1　研究综述

1.1.1　研究背景

股票作为金融市场最具代表性的金融产品，具备金融产品的三个最基本的特征，即为收益性、流动性以及风险性. 随着股票市场的不断发展与完善，稳定的市场环境、健全的投资机制以及完善的股份制公司发展模式吸引着越来越多的投资者参与股票投资. 金融产品收益性的特征，满足了投资者资本逐利的心态，对企业的价值投资、短线投机以及风险对冲分别对应着市场中投资者、投机者和套利者三种投资者类型；流动性的特征赋予了股票的价值，价值的多元化特征使得金融产品在市场中得以流通；风险性的特征，则让投资者在逐利的过程中，始终保持理性的判断，投资风险的规避及其对风险的控制则是使得资本市场投资变得精彩纷呈的原因. 上述三类基本特征，使得金融产品的投资者必须在投资过程中权衡收益与风险之间的关系.

在股票市场中，传统的投资方法运用基本面分析与技术分析来判断单只股票在市场中的表现. 在基本面分析中，宏观经济分析与产业周期分析是市场整体情况的判断工具；公司财务与价值的分析，则是着力于公司的经营管理情况，可以见微知著的审视公司的发展情况以便判断公司股票是否具有投资价值. 技术分析则专注于股票的市场行为，通过对股票价格数据及其指标的观察来判断股票价格的未来变化趋势. 可以看到，传统的投资方法以其可操作性强与操作难度较小而被投资者广泛使用，但是传统投资方法的局限性体现在以下几点. 其一是传统投资方法很大程度依赖于理论的假设，实际的市场并非能够很好的满足假设条件；其二是证券市场价格的变动表现为非常复杂的形式，凭借单一的基本面分析技巧或是技术指标，很难做出最优的投资决策；其三是要凭借着传统投资方法，选出心仪的股票进行投资，需要较多的精力，所以在面对市场中众多的优质股票而言，其投资效率不高. 近年来，随着计算机性能的提高与机器学习技术的发展，研究者在面临上述传统投资方法的缺陷时，考虑如何在保证较高收益的前提下，尽可能的降低投资风险，同时可以提高选股与决策的效率，进而运用计算机进行资产投资并辅助决策成为了主流投资方式，催生出了量化投资. 而研究者所提出的新的高效投资方式，其实质是通过资本资产定价理论的启发，将实证资产定价模型运用于股票价格的预测和收益率的预测中来. 而股票收益率的预测在现代金融学理论中，更多的是将其定义为股票风险溢价的衡量.

实证资产定价起源于上世纪七十年代, 在经历了不断的发展、验证、完善、创新之后, 线性多因子模型的研究价值越来越受到人们的重视, 但是随着市场的不断发展与研究者提出越来越多影响股票收益率的因素, 传统的线性多因子模型被用于股票收益率估计时有以下几个问题. 一是线性模型的局限性使得模型难以拟合出因子之间的非线性因素, 并且传统的组合排序与回归方法无法处理高维因子, 使得模型的拟合效果不尽人意; 二是经过实证验证的影响股票收益率的因子预测稳健型不高, 在稳定预测一段时间之后预测效果就会急剧下降; 三是因子的数量不断增多, 通过不断改变其线性组合而从众多因子中选取出预测效果较好的因子及其组合并非易事.

市场的繁荣体现在产品的丰富程度上, 在实证资产定价领域, 股票因子的不断发掘与创新造就了股票“因子市场”的繁荣. 经过实证以及金融理论验证的股票因子对于我们构造因子模型, 进行实证资产定价有着一定的帮助, 但是绝大多的因子并非在实际应用中有着良好的表现, 并且大部分因子难以持续有效. Cochrane 在 2011 年美国金融协会演讲中提出了“因子动物园”概念[1], 要求研究者能够识别出提供美国股票平均回报独立信息的公司特征, 这些公司特征也被称为是横截面收益预测因子. 从“因子动物园”中发掘出我们想要的有效因子, 成为了现代实证资产定价领域的研究方向. 此外, 行为金融学作为现代金融理论研究的前沿, 不断有学者将涉及行为金融学的因子引入实证资产定价模型, 使得基于现代行为金融理论的因子模型成为了改进上述传统实证资产定价模型的一个思路. 另外一个克服上述传统线性模型缺陷的思路就是引入机器学习模型挖掘有效因子. 因此近年来, 对于实证资产定价的前沿研究, 大都集中于行为金融理论下的因子研究与机器学习框架下的因子研究. 本文的研究集中于后者, 将机器学习模型引入实证资产定价, 测算股票的风险溢价.

1.1.2　研究意义

本文将机器学习与实证资产定价相结合, 证实了机器学习模型相较于传统的线性定价模型, 在具有较多潜在影响股票收益率的变量和回归函数范式时, 机器学习模型更能应对自如. 机器学习模型的这种适应性体现在, 我们无需指定要引入模型的股票因子, 只需将需要考察的因子集中的因子统统输入模型, 便可以有效衡量股票的风险溢价, 并且可以确定含有信息量最大的预测变量, 这有助于我们对实证资产定价的经济机制进行更有效的研究. 实证资产定价的许多特质, 都使得机器学习方法在实证资产定价中能够发挥其优势, 本文中引入的实证研究正揭示了这种优势所在.

其次, 为了探究上述优势, 我们将国内 A 股市场股票作为实证资产引入我们

的研究. 我们与 Gu 等人[2] 的研究作对比时发现国内 A 股市场相较于美股市场存在" 异象". 国内股票市场无论是市值规模还是融资规模, 均处于世界第二位, 特别是近两年以来在疫情以及西方贸易保护主义的双重影响下, 国内股票市场颇受国际市场青睐, 但是国内 A 股市场始终区别于国外市场的一点是, 国内 A 股市场中有大量规模相对较大的国有企业. 我们在本文的研究中, 证实了国有企业对于股票市场的潜在影响大于其他非国有企业的影响, 这种潜在影响力体现在股票未来收益率的可预测性之中. 除了国有企业之外, A 股市场中规模较大的散户投资者也对 A 股市场出现异象做出了贡献. 据统计, 截止 2022 年底, 全国在 A 股开设账户的投资者达到 2.11 亿户, 其中个人投资者数量占比为 99.76%, 而机构投资者占比仅为 0.2% 左右, 大量的散户投资者在 A 股市场中参与投资, 由散户投资者所造成的集体投资行为, 加剧了 A 股市场的流动性. 因此我们也进一步证实了, 个人投资者的集体投资行为也是国内股票市场存在异象的原因之一.

我们将机器学习模型运用于实证资产定价问题, 不仅有效衡量股票的风险溢价, 还能挑出对股票收益率影响显著的因子, 此外, 针对不同的股票类型集合, 我们还给出了该类型股票中包含信息量最大的几类因子, 以便投资者在进行投资时按照所要投资股票的类型选择相应的因子, 这不仅丰富了投资者的决策方式，而且有助于我们了解股票的市场行为. 参与资本市场投资的投资者中, 本身具有专业投资经验与专业投资知识的投资者占比较小, 调查显示具有丰富投资经验的投资者只占总体投资者的 47.64%, 再加上运用机器学习模型进行股价预测时要涉及多方面的基础知识, 所以建立一个针对不同股票类型选出有效因子的“黑箱”模型, 可以简化使用门槛使得机器学习在实证金融领域得到广泛应用. 其中最为简单可行办法的就是, 通过将多个影响股票价格的个股因子与宏观经济因子引入机器学习模型之中, 通过模型来鉴别出对未来股票价格影响最大的因子, 以便投资者从令人眼花缭乱的因子库中, 选择出最具解释力的因子, 作为投资前的参考以辅助投资者决策.

1.2　文献综述

1.2.1　国外研究现状

实证资产定价作为金融经济学领域内最重要的理论之一, 解决了如何量化资产未来的价值或者价格. 马科维茨资产投资组合理论奠定了现代金融学的基石, 自其提出以来, 金融学界对未来资产价值以及价格的研究方兴未艾. 从 CAPM 模型、套利定价理论到如今广为应用的 q-factor 模型, 其为传统的实证资产定价理论. 机器学习和深度学习作为新兴的解决各种与数据有关问题的工具, 在经济与

金融中的应用随着近年来计算机算力的不断提高而如雨后春笋般涌现, 从量化投资到资产定价, 都不乏机器学习和深度学习的研究文献, 以此极大地推进了经济金融领域的发展.

而实证资产定价理论中最典型而且研究最多的问题则是衡量资产的风险溢价, 也就是说资产定价的基本目标就是了解风险溢价行为. 但是问题的难点恰恰就在与风险溢价的测算. 股票作为一种具有代表性的实证资产, 所表现出的投资便利性与高流动性吸引着众多的投资者参与其中, 所以预测股票未来的收益率, 进而测算股票的风险溢价成为了近年来研究的主流方向. 在 2002 年, Fama 和 French[3] 认为股票风险溢价难题的根源在于测算平均股票收益率与无风险收益率之间的差额. 所以准确的预测股票未来收益率是衡量股票价值的重要一环, 而衡量股票价值就意味着要在众多影响股票的因素中去粗取精的发现其关键的影响因素. Basal 和 Yaron[4] 在 CCAPM 模型的基础上, 研究了美国股市的长期收益率的可预测性, 并且证实了股票风险溢价的可观测性. Fama 和 MacBeth[5] 在 1973 年所提出的 Fama-MacBeth 两步截面回归法的目的在于检验 CAPM 模型, 在金融领域中也常被用于考察多因子模型中的实证资产的因子暴露与截面收益率的关系. 而后 Fama 和 French 在其 Three-Factor Model(1993)[6] 的基础上, 构造并检验了 Five-Factor Model(2015)[7] 在实证资定价中的表现, 尽管证明了在实证资产中 Five-Factor Model 表现更好, 但是 Five-Factor Model 也难以捕捉小型股票的收益率. Giglio 和 Xiu[8] 指出, 即使因子并非全部都被指定或者可观测, 但可以用 PCA 将此类隐性因子的因子风险溢价和风险敞口估计出来, 文献中将其称之为 Three-Pass Estimator 方法, 用此方法对风险溢价进行有效的测量, 并实证验证了其有效性.

除此之外, 有大量的研究文献都是基于衡量股票的风险溢价而描述股票收益率预测模型的优劣, 进而探究股票的实际价值与价格的关系, 从而让投资者可以在投资行为中获得收益, 甚至于获得超额收益. 但是, 机器学习与深度学习的应用给研究者提供了全新的思路与方法. 与此同时 Welch 和 Goyal[9] 指出线性的预测模型并不能对股票风险溢价进行很好的预测, 文献中使用 OLS 模型对 30 年以来的数据进行了重构检验, 检验结果表明绝大多数线性模型即使在样本内 (IS) 数据中的表现也不是很好, 同样样本外 (OOS) 数据的检验结果更加不尽人意. 除了线性因子模型在测定风险溢价时的有效性遭到质疑以外, Bai 和 Serena[10] 指出因子模型理论和经验有效性的核心是得到正确的因子数量和因子类别. 尽管 Green 等人[1] 认为有些因子在面对不同的情况时, 有着较高的实证价值并且表现出较好的预测稳健性, 但此文献中指出自 2003 年开始, 这些因子组合的收益可预测性急剧下降. 因此, 随着近年来因子数量的不断增长, 从众多因子中选择出对实证资产价

格影响显著且稳健的因子并非易事.

而机器学习算法的优点与特性正好有助于解决上述的问题. Gu 等人[2] 认为, 与传统的实证资产定价理论相比, 机器学习在处理股票数据时, 提供了更为广泛的潜在预测变量列表和更为丰富的规范函数形式, 所以这种灵活的方法也很好的推动了风险溢价研究的前沿. Gu 等人[2] 在文献中指出机器学习算法的目的是解决统计预测中的高维问题以及与之伴随的选择防止预测模型过拟合的正则化方法. 以此为出发点, 近年来研究者提出的基于机器学习算法的实证资产定价模型大概可以归纳为以下两个方向.

其一, 即为在传统的线性模型的基础上, 加之以机器学习算法的防止过拟合现象的正则化条件, 构造带有惩罚的广义线性模型, 最为常见的就是带有 Lasso 和 Huber 惩罚项的线性模型. Rapach 等人[11] 在文献中应用 Lasso 预测了全球股市收益. Fan 等人[12] 提出的带有 Diverging Parameter 的 Huber 惩罚项, 为高维数据和厚尾分布数据提供了使用 Huber 损失函数的理论依据. Chen 等人[13] 运用偏最小二乘回归 (PLS) 与主成分分析 (PCA) 提取出了刻画投资者情绪的有效指标, 认为投资者情绪指标的预测能力要优于常见的收益率预测因子的预测能力, 因此表明了投资者情绪对于估计市场风险溢价的重要性. Rapach 等人[14] 提出将单一线性模型的预测结果组合起来, 即取平均得出最终预测的结果, 此类模型称之为混合线性模型, 模型对样本外 (OOS) 数据进行重构检测发现其对股票风险溢价的预测效果要高于所有的单一预测模型. Leung 和 Tam[15] 提出了用弹性网 (Elastic Net) 模型来估计统计套利风险溢价, 并考察统计套利风险与统计套利风险溢价的关系.

其二, 就是以机器学习和强化学习算法为主的非线性预测模型, 非线性预测模型又主要以树类模型和神经网络模型为代表. Ozbayoglu 等人[16] 对近年来机器学习和深度学习算法在金融领域中的应用做了全面的调查与概括, 这种显著优于经典模型的非线性模型在金融业中得到了广泛的应用, 其中涉及了算法交易、风险评估、投资组合管理以及资本资产定价等领域. Khandani 等人[17] 提出了一种基于 CART 的树型预测模型, 来对消费者信用违约和拖欠行为进行预测. Moritz 和 Zimmermann.[18] 提出了一种基于树的条件投资组合排序, 此种模型可以处理高维变量以及变量之间潜在的非线性以及相互作用. Hutchinson 等人[19] 运用 RBF 神经网络模型对 S&P 500 指数期权进行预测, 并证实预测效果优于 Black-Scholes 定价公式. 而 Yao 等人[20] 用 BP 神经网络模型对日经指数期权进行了预测, 实证发现也有较好的预测性能. Heaton 等人[21] 运用几种常见的深度学习模型, 比如自编码器模型和 LTSM 模型, 来用于投资组合选择和进行风险管理. Gu 等人[22] 将机器学习中的自编码器模型与运用于资产定价之中, 旨在捕捉影响实证资产价格变

化的的潜在因子. Chen 等人[23] 通过将三种简单的神经网络模型结合并且在模型引入无套利条件, 构建了一种新的估计个股收益率的资产定价模型, 在实证检验中证实该模型具有较好的表现.

1.2.2　国内研究现状

我们将目光转向国内市场, 国内金融市场作为新兴市场随着近几年的迅猛发展, 市场体量已经相当可观, 金融制度也日趋完善. 基于对于国内金融市场的有效性研究, 学者们借鉴国外市场的实证资产研究机制, 将实证资产定价理论运用于中国国内市场并根据国内市场的特殊情况对资产定价理论作出了符合实际情况的改进. 1998 年, 在 Fama 和 French[24] 以及 Rouwenhorst[25] 的研究文献中指出, 在美国市场上基于横截面数据的预测模型可以推广到其他国家或者地区的市场之中, 以此作为推广资本资产定价理论的依据来研究中国国内市场. Wang 和 Xu[26] 运用 Fama-French 三因素模型检验了影响国内 A 股市场收益率的有效因子为规模因子, 也证实了实证资产定价模型在国内市场的有效性, 尽管效果有所减弱. Drew 等人[27] 指出用传统的资产定价模型很难描述中国 A 股市场, 并且证实了 A 股市场具有明显的季节性因素, 这是传统定价模型中的市场因子所无法解释的. Wang 和 Chin[28] 考察了中国 A 股市场的历史成交量和收益率的交互作用在股票收益率预测中的表现, 发现了此种交互作用在中国股市预测中的表现与参与对照的美股市场结果不相一致, 作者认为造成此种差异的原因在于中国股市拥有其特有的特征. Chen 等人[29] 在文献中用传统的实证资产定价模型来研究中国股市的可预测性, 选取 18 个已被实证有效的预测因子分别预测中国股市和美股的收益率并进行横向比较, 发现中国股市的可预测性较低, 作者通过实证检验将其解释为中国股票价格的可解释性相对较小, 这也与中国股市的异质性有关. Cakici 等人[30] 对中国股票收益率可预测性进行了综合分析, 运用分析收益可预测性的投资组合方法与基于横截面数据的回归方法, 得出了对中国股市收益率预测有显著影响的因子相较于其他参与横向对比的国外市场有所差别.

尽管在中国股票市场中, 对市场有效性的验证以及国内外市场同质性的研究证明了可以将经典的实证资产定价理论运用于对中国市场的研究, 但是上述所列举的文献大多表明中国股票市场存在着异质性, 简单的对经典实证资产定价理论的复刻并不能较好的预测国内股市的收益率, 进而不能对实证资产的风险溢价进行有效且合理的估计, 于是学者证实了一些符合中国市场的特有因子并考虑构建符合中国市场特点的实证资产定价模型. Pan 等人[31] 提出中国股市因一系列特质因素的影响表现出典型的投机性特征, 作者在文献中检验了中国股市中投机交易对股票收益率的影响, 开发了一个与交易量相关的变量, 即异常周转率 (atr), 以

此作为中国股市特有的因子. Liu 等人[32] 基于 Fama-French 三因子模型, 构建了包含中国市场特有因子的三因子模型 (CH-3), 与经典的三因子模型进行了横向对比, 证实了该因子模型在中国市场有较好的表现. Liu 等人[33] 同样基于中国股市, 构建了适合中国市场的四因子模型, 其构建策略是在 Liu 等人的三因子模型的基础上, 添加了趋势因子, 并且该模型的有效性在实证中得到了验证.

国内市场中所表现出的特性对传统的实证资产定价理论构成了挑战, 需要更加符合中国市场的资产定价模型或因子来进行建模, 以便更加合理的解释国内市场的风险溢价现象以及准确的衡量其风险溢价. Leippold 等人[34] 在文献中归纳总结出中国股市所具有的三类相较于国外市场的异质性特征, 分别为: i 散户主导; ii 政府影响; iii 卖空限制. 也正是此三类特征造成了国内市场有别于国外市场的异质性. 面对以上三类特征, 研究者们提出的基于经典实证资产定价理论的线性模型在实证检验中发挥了非常有限的效力, 同国外市场的实证资产定价研究一样, 近几年也开始在国内市场用机器学习与强化学习等非线性模型来进行实证资产的定价研究, 以期更好的对中国市场的实证资产进行合理的定价与预测. Cao 等人[35] 首次将神经网络模型用于对中国股票市场的预测, 并通过与传统的线性模型进行比较发现, 神经网络模型具有较好的预测性能. Wang 等人[36] 构建了将时间序列模型和神经网络模型混合的预测模型, 来对未来股票的价格做出预测, 并以深圳综合指数和道琼斯工业指数为例分别做了实证检验, 其预测表现较好. Chen 等人[37] 运用 LSTM 神经网络模型对中国股市进行收益率预测研究, 并且证实该模型相较于其他模型其预测准确率大幅提高. Zhang 等人[38] 提出了一种进行股票收益率预测因子选择的新算法, 因果特征选择 (CFS) 算法, 并与经典的机器学习的特征选择算法做了对比, 发现在股票收益率预测因子的选择上, 有较好的表现. Yuan 等人[39] 在文献中运用了不同的机器学习与强化学习算法对影响股票收益率的因子进行筛选, 并对国内 A 股市场进行股票收益率的预测, 发现当随机森林算法应用于因子选择和收益率预测时有较好的表现.

随着近年来国内金融市场体量不断发展壮大, 学者对中国国内市场的实证资产定价研究也逐渐增多, 目光越来越多的集中于国内股票市场, 尤其是 A 股的风险溢价测算上. 但由于中国股市有着其独特的异质性特征, 进而也催生出众多分析论证中国股市异像进而发掘出中国股市特有因子的文献, 也不乏构建符合中国股市特有的异质性预测模型的文献. 将上述两种方向相结合, 发掘与构建对股票以及投资组合收益率预测有显著效果的因子与模型, 也是近年来研究的新方向与突破点. 以 Leippold 等人[34] 的研究为例, 将国外市场对实证资产定价的研究新理论, 即以机器学习和深度学习等非线性模型为主导的方法, 应用于对中国市场的实证资产定价的研究中, 这也成为近年来国内实证资产定价研究的重要方向.

1.3 研究内容

1.3.1 研究目标

本文以国内 A 股市场为研究对象, 探究机器学习模型在实证资产定价中所发挥的作用. 而资产定价的基本目标则是了解风险溢价行为, 但是风险溢价的衡量比较困难, 所谓风险溢价, 实际上就是已经确定的收益与所承担风险收益之间的差额. 已经确定的收益即为无风险收益率, 在本文中我们用十年期国债收益率来代替, 而风险收益则是股票价格的变化所引起的收益率的增减. 对股票而言, 由于股票的风险溢价即为股票收益率与无风险利率的差额, 无风险利率一般为常数, 所以研究股票风险溢价时, 收益率与无风险利率的差额可以仍然以股票收益率代替, 衡量股票风险溢价即为预测股票在未来一段时间的价格. 从衡量股票的风险溢价为出发点, 本文的研究目标大致为以下几个方面.

首先, 风险溢价测量结果的准确性以机器学习模型的预测效果来评判, 既然是机器学习在实证资产定价中的应用, 就要探究哪种机器学习模型在用于资产定价时有着更好的效果. 机器学习模型既然有其线性模型无法比拟的优点, 所以文中也以较大的篇幅介绍了机器学习模型较线性模型的优势所在.

其次, 本文以国内 A 股市场为研究对象, 目的是为了探究 A 股市场相较于美股市场的差异所在. 在有效市场假说下, 国内股票市场虽是规模与市值位居世界第二, 仅次于美股市场, 但是研究者发现其仍未达到半强式有效市场所定义的范畴. 在 A 股市场与美股市场对比中所表现出的种种“异象”, 促使我们考量 A 股市场出现“异象”的原因.

最后, 机器学习模型对股票未来一段时间的收益率如果得到了完美的观察, 我们需要解释参与回归的股票因子中, 哪些因子所包含的信息量较大, 即为找出对股票收益率影响最大的因子, 以便对这些因子进一步分析和验证其有效性. 此外, 不同的股票类型集合之中, 其对股票收益率影响最大的因子有所不同, 而在对应集合中找这些因子, 也是本文的研究目标之一.

1.3.2 研究思路

根据本文的研究目标, 我们选取了截止 2022 年 4 月 A 股市场的所有股票作为研究对象, 并搜集了股票的收益率数据及其 100 个个股因子数据, 为了在模型拟合时更加接近股票的市场化行为, 我们还搜集了可能影响股票价格的 13 个宏观经济因子以及所有要考察股票的行业分类数据, 分别对上述股票的收益率数据与因子数据做同质化处理之后作为回归模型的因变量与自变量. 数据整理妥当后, 我们选取用于做模型拟合及预测的机器学习算法, 我们选取了应用广泛的几类机器学习模型, 并且为每一类机器学习模型选定两种具有代表性的算法. 简单线性

模型作为基准模型, 将传统的线性资产定价模型与机器学习模型在股票未来收益率预测中的差异性做了对比, 我们选择了普通最小二乘回归 (OLS) 模型与三因素的 OLS(OLS-3) 模型；带惩罚的线性模型中我们选取了最具代表性的 LASSO 与弹性网 (Enet) 模型, 引入正则化项来控制模型训练中的过拟合现象；降维的线性模型中我们选择了主成分回归 (PCR) 与偏最小二乘回归 (PLS), 以作为消除变量间相关性的手段；集成学习中我们选择了随机森林 (RF) 与梯度提升回归树 (GBRT) 模型, 此两种机器学习算法可以很好的处理非线性数据集；最后我们用神经网络模型, 进一步证实多层神经网络模型在实证资产定价领域内应用的成效. 有了数据集与选定机器学习模型之后, 我们逐步开展研究.

首先, 将处理好的股票数据集带入模型, 测算股票未来一段时间的收益率, 然后借助样本外预测评价指标 R^2_{OOS}, 横向的对比机各个器学习模型在实证资产定价领域的表现, 纵向的对比同一机器学习模型在不同类型的股票集合中的预测差异. 我们将股票按照其市值、前十大股东股权占比、股权性质以及板块市场依次划分成不相交的集合类, 探究不同集合间的整体差异以及机器学习模型在各个集合上的差异.

其次, 不同机器学习模型在不同类型的股票集合中有较大的预测差异, 对不同的股票集合, 我们探讨了哪些股票因子对股票收益率的可预测性所做出的贡献最大. 具体的方法是, 将参与模型拟合的股票个股因子逐一设置为零, 通过 R^2_{OOS} 减小的幅度来定量的考量个股因子对于模型拟合的重要程度.

最后, 将参与回归的宏观经济因子也做与上述中的个股因子相同的处理, 定量的考察宏观经济因子对股票收益率的影响, 确定出对股票收益率综合影响最大的宏观经济因子.

1.4　章节安排

本文参考 Leippold 等人[34] 的研究结果, 通过预测股票未来的收益率代替风险溢价的测量, 在证实了机器学习模型在实证资产定价领域体现出其灵活性优势的同时也具有经济意义, 也进一步证实了国内 A 股市场中存在的异象. 文章的章节安排如下.

第一章绪论部分通过对研究背景和研究意义的阐述, 指出了文章所研究课题的现实意义及其经济意义, 并且通过文献综述部分，阐明了金融经济学界近年来应用机器学习模型在实证资产定价领域中的应用及推广.

第二章研究方法部分, 我们主要介绍了本文中用于估计股票风险溢价的机器学习算法, 以及总体回归拟合模型的设定. 通过将套利定价理论模型做了适当的变形, 从而将资产的超额收益描述为一个加性预测误差模型, 并且适用于文中所

提到的所有机器学习算法的总体模型.

第三章研究设计部分主要介绍了本文中作为回归因变量的股票因子. 与此同时, 我们在参考文献所做研究的基础上, 扩充了样本数据, 更新了个股因子数据以及宏观经济因子数据. 数据包含了 1997 年 1 月至 2022 年 4 月间的 4932 只股票的基本信息以及月度收益率, 其跨度达 304 个月; 个股因子数据扩充至 100 个, 宏观经济因子数量也扩充至 13 个. 使得机器学习模型更有机会挖掘出对股票价格具有潜在影响力的因子.

第四章实证分析部分, 我们将股票数据及其因子数据带入模型, 同时将股票数据集合按照不同性质划分, 然后将个股因子通过逐一归零的方法, 选出不同性质集合中对各自集合影响力较大的因子. 不仅对不同股权性质的股票做了划分, 同样地, 我们也对不同板块市场进行划分, 通过对各个不同的板块市场分别运用 13 类机器学习模型进行拟合, 揭示不同板块市场之间的可预测性差异, 并通过逐一归零法选出各个板块市场之中对股票价格影响最大、包含信息量最多的个股因子. 在本章的研究中, 我们还分析了宏观因子在不同的股票集合中的表现, 但是发现在不同股票集合中, 宏观经济因子的差异并不显著, 所以我们认为, 宏观经济因子只在市场指数层面产生影响. 因此, 同样地我们通过归零法选择出了对于整体股票价格影响最大的几种宏观经济因子.

第二章　研究方法

在上一章中, 我们讨论了实证资产定价理论发展的大致历程, 从传统的简单线性模型逐步扩展至复杂的非线性模型, 且随着研究方法的不断创新与应用领域的拓宽, 实证资产定价理论早已突破了资本资产定价理论在其模型设定与研究方法上的约束, 从原来以传统经济金融学角度的研究方法为主, 转而向以数据为驱动的统计学角度发展. 在此前的章节中我们也提到, 为了准确的了解股票资产的风险溢价行为, 研究者们提出其有效的办法是尽可能准确的预测股票未来预期收益率. 而以股票为代表的诸多实证资产的预期收益由许多因素决定, 单单就反应个股情况的股票因子而言, 其种类繁多令人眼花缭乱. 所以, 研究资产定价中高维以及非线性的特性才能推动风险溢价研究的前沿, 这就意味着我们采用机器学习与深度学习方法进行因子选择与模型拟合符合实际情况.

本章中, 我们简单概述本文中所用到的机器学习算法与总体模型. 本章第一部分的每个小节从统计模型设定、最优化策略选择、最优化计算算法三个方面来介绍机器学习算法. 首先, 模型设定描述了此机器学习算法的一般函数形式, 这个函数形式反映了模型用于描述风险溢价的方法；其次, 最优化策略选择描述了此模型用于估计模型参数的目标函数；最后, 最优化方法论述了用于求解目标函数的最佳迭代算法. 由于各种机器学习算法的适用情况与应用环境的不同, 文中并未应用全部的传统机器学习算法, 只是针对我们所研究的问题, 选择几个具有代表性且在此类问题中表现较好的机器学习算法. 本章第二部分介绍了适用于文中所提到的所有机器学习算法的总体模型, 将资产的超额收益描述为一个加性预测误差模型, 同时以多元线性回归 (OLS) 模型为例, 简单证明了我们所选取的总体模型与传统多因子模型之间的关系.

2.1　机器学习预测算法

2.1.1　简单线性模型

从最简单的线性模型入手, 将其作为我们与其他模型比对的基准. Fama 和 French[7] 等人的工作是从金融经济学领域出发找出恰当的因子, 构建预测股票预期收益率的线性模型. 而我们在本文中所采用的线性模型, 包含了所有的预测因子, 将此模型记为 OLS 模型. 将找出的所有因子视作因子池, 这一假定等同于因子动物园 (Zoo of factor)[1] 的说法. 可以想象, 将所有的因子视作线性回归的自变量带入模型后, 由于因子间的相关性、模型自身的线性性设定以及自变量的高维特

征等影响, 模型对股票收益率的预测效果难免差强人意, 但是作为与其余机器学习模型比较的基准, 它可以强调非线性模型在股票收益率预测中良好的表现. 我们假定原始的预测因子与股票的未来预期超额收益率之间满足如下线性关系:

$$f(\boldsymbol{z}_{i,t};\theta)=\boldsymbol{z}_{i,t}^{'}\theta, \tag{2-1}$$

其中 $f(\boldsymbol{z}_{i,t};\theta)$ 表示股票 i 的预期超额收益率的期望, 函数 $f(\cdot)$ 描述了股票的 $p\times 1$ 维预测因子向量 $\boldsymbol{z}_{i,t}^{'}$ 与预测因子的系数矩阵 θ 之间的关系. 实际上, 在线性模型中我们表示股票的预期超额收益率会以如下的式子为标准:

$$r_{i,t+1}=\boldsymbol{z}_{i,t}^{'}\theta+\varepsilon_{t+1,i}. \tag{2-2}$$

上式中, $r_{i,t+1}$ 为股票 i 在未来 $t+1$ 期的预期超额收益率, $i=1,\ldots,N$ 且 $t=1,\ldots,T$, $\varepsilon_{t+1,i}$ 表示股票 i 在 $t+1$ 期的预测误差. 式 (2-1) 仅仅是考察股票未来预期超额收益率期望的一个式子, 它是式 (2-2) 的一个简化.

在金融数据中, 尤其是股票收益率数据的分布, 常常会呈现出尖峰厚尾的特性, 出现此种特性的原因在于金融数据频繁出现的异常值, 异常值对模型的拟合也带来了较大的影响, 最小二乘估计中的 l_2 损失函数使得用于估计模型参数的目标函数容易收到异常值的影响, 所以在本文中, 我们采用 Huber 损失函数[40] 来代替 l_2 损失, 即有:

$$L_H(\theta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}H\left(r_{i,t+1}-f(\boldsymbol{z}_{i,t};\theta);M\right), \tag{2-3}$$

其中,

$$H(x;M)=\begin{cases}x^2, & \text{if } |x|\le M,\\ 2M|x|-M^2, & \text{if } |x|>M.\end{cases} \tag{2-4}$$

可以看出, Huber 损失函数是由较小误差时的平方损失与较大误差时的绝对值损失混合构成, 实际是 l_2 损失的改进, 其中 M 作为判断是否出现异常值点的阈值, 是一个可以调整的参数, 当 $|x|\le M$ 时, Huber 损失即为 l_2 损失. 在选择目标函数式 (2-3) 的最优化方法时, 我们考虑到 Huber 损失函数在任意点是可导的, 所以既可以用求偏导的方法来找最优化参数, 也可以用梯度下降法实现相同的目的. 此外, Huber 损失函数不仅适用于线性回归模型, 也适用于其他类型的机器学习模型.

本文中除了将所有的预测因子作为自变量引入线性模型之外, 还另外考察了只含有规模、账面市值比和动量这三个因子作为自变量的线性模型, 在其后的论述中我们将其记为 OLS-3 模型. 选择这三个因子作为自变量来建模是因为, 这三个因子是传统的多因子模型中应用最为广泛的因子, 分别对应其传统多因子模型中的规模、价值和动量[41].

2.1.2 带惩罚的线性模型

当高维协变量参与到普通线性回归模型之中时, 由于变量间线性相关所引起的多重共线性问题和无正则化项产生的过拟合现象, 使得回归的结果通常变的不可靠. 所以, 我们考虑了加入正则化项的线性模型, 应用较为广泛的正则化线性模型有岭回归、LASSO、弹性网等. 带有惩罚的线性模型的总体模型设定与 OLS 模型一样, 区别只在于将 OLS 模型中用于估计模型参数的目标函数加入了正则化项, 其用于估计参数的目标函数的统一形式可以表示为:

$$L(\theta;\cdot) = L(\theta) + \phi(\theta;\cdot), \tag{2-5}$$

其中, $L(\theta)$ 表示 OLS 模型中的损失函数, 带惩罚的线性模型之间以正则化项 $\phi(\theta;\cdot)$ 取不同的函数形式作为区分. 与此同时, 带惩罚的线性模型中的 l_2 损失项 $L(\theta)$ 也可以用 Huber 损失 $L_H(\theta)$ 代替, 以避免异常值对模型拟合的干扰.

首先考虑 LASSO 模型, 研究者基于时间序列数据和独立同分布数据已经对其统计特性做了大量的研究. LASSO 自 1996 年首次被 Tibshirani[42] 提出以来, 在高维数据的处理中表现出良好的性能, 它可以将引入模型的不相关的协变量系数收缩为零, 从而达到变量选择的目的. LASSO 回归中的正则化项为 l_1 惩罚, 其损失函数形式如下:

$$L_H^{\text{LASSO}}(\theta) = L_H(\theta) + \lambda \sum_{j=1}^{P} \left|\theta_j\right|, \tag{2-6}$$

其中, $L_H(\theta)$ 为式 (2-3) 所示的 Huber 损失函数, λ 为控制正则化项大小的超参数. 式 (2-6) 为 LASSO 回归中用于估计参数的目标函数, 而在超参数被指定的情形下, 由于 LASSO 回归中的目标函数带有绝对值, 故惩罚项不是连续可导, 所以我们使用坐标轴下降法或者最小角回归法来寻找目标函数的最优化参数.

另一个在本文中要考虑的带惩罚的线性模型为弹性网 (Elastic Net)[43], 弹性网模型是将岭回归与 LASSO 的正则化项结合起来的正则化线性模型, 相比较单一的岭回归或是 LASSO 模型, 弹性网模型不仅可以起到变量选择的作用, 而且相较于 l_1 惩罚的不可导性, 其有着计算上的优势. 弹性网模型的损失函数如下所示:

$$L_H^{\text{Enet}}(\theta) = L_H(\theta) + \lambda \sum_{j=1}^{P} \left(\rho\theta_j^2 + (1-\rho)\left|\theta_j\right|\right). \tag{2-7}$$

上式中, λ 和 ρ 均为超参数, 其中 ρ 的取值范围为 $[0,1]$, 决定了 l_1 惩罚与 l_2 惩罚在弹性网模型的损失函数中的权重. 可以看出当 $\rho = 0$ 时, 式 (2-7) 则等价于 LASSO 回归的损失函数; 当 $\rho = 1$ 时, 式 (2-7) 即退化为岭回归的损失函数. 与 LASSO 回归类似, 弹性网回归在求解目标函数时可用坐标轴下降法来迭代求解最优化参数.

2.1.3 降维线性模型

带惩罚的线性模型一般采用系数收缩或是变量选择的方法处理高维协变量问题, 模型中最终出现的仍是被选择出的原始变量. 而降维的方法则是提取原始变量的信息生成新的回归变量, 再将生成的回归变量带入线性模型参与模型的拟合. 这种通过原始变量的线性组合而生成新的回归变量的做法, 可以有效的减少噪音以更好的分离出原始变量中的有用信息, 而且可以有效的消除原始变量中的相关性. 主成分回归 (PCR) 和偏最小二乘回归 (PLS) 作为两种具有代表性的降维线性回归方法, 自然也被应用于本文中来处理高维股票因子变量.

在陈述降维线性模型之前, 我们首先将表示股票未来预期超额收益的函数式 (2-2) 扩展至向量的形式, 即有:

$$R = Z\boldsymbol{\theta} + E, \tag{2-8}$$

其中, R 表示股票预期超额收益的 $NT \times 1$ 维向量, Z 为股票的预测因子向量 $\boldsymbol{z}_{i,t}^{'}$ 按照时间序列堆叠而成的 $NT \times p$ 维矩阵, E 表示残差序列, 为 $NT \times 1$ 维. 接着, 我们引入降维线性模型的模型设定, PCR 与 PLS 都是经典的降维线性模型, 两种方法在其基本模型的设定上差别不大. 假定我们所要考察的用于预测股票超额收益率的因子数目为 p, 降维之后模型的自变量的维数变为 $M(M < P)$, 所以降维线性模型的模型设定为:

$$R = (Z\Omega_M)\boldsymbol{\theta}_M + \tilde{E}, \tag{2-9}$$

上式中, $\boldsymbol{\theta}_M$ 是 $M \times 1$ 维向量, 表示降维线性模型中自变量的系数, 并且设 $\Omega_M = (\omega_1, \omega_2, \ldots, \omega_M)_{p\times M}$ 表示对原始自变量即股票的预测因子进行线性变换的系数矩阵, 线性变换后的股票因子由于进行了线性组合从而得到新的回归变量 $Z\Omega_M$, 进而也实现了降维的目的.

在 PCR 模型中, 运用主成分分析来生成原始变量的不同权重的线性组合, 即主成分, 我们选择前 $M(M \ll p)$ 个主成分作为新的回归变量, 以计算第 j 个线性组合为例, 简要说明主成分的求解方法:

$$\begin{aligned} \omega_j = \quad & \arg\max_{\omega} \operatorname{Var}(Z\omega) \\ \text{s.t.} \quad & \omega'\omega = 1 \\ & \operatorname{Cov}(Z\omega, Z\omega_l) = 0, l = 1, 2, \ldots, j-1. \end{aligned} \tag{2-10}$$

可以看出, 求解第 j 个主成分实际上是带约束的最优化问题. PCR 模型面临的问题在于, 求解原始自变量的主成分时没有将因变量考虑进去, 所以主成分的求解完全脱离了因变量的约束. 而 PLS 模型相较于 PCR 模型, 其原理是考虑自变量与因变量相关性情形下的最小二乘估计算法. 而两种模型的不同之处则在于迭

代求解原始变量线性组合的方法, PLS 的主要目的在于求解与预测目标具有最大关联的 Z 的 M 个线性组合, 尽管如此, PLS 求解线性组合的方法仍然可以写成类似于式 (2-10) 的形式:

$$\begin{aligned} \omega_j = \quad & \arg\max_{\omega} \mathrm{Cov}^2(R, Z\omega) \\ \text{s.t.} \quad & \omega'\omega = 1 \\ & \mathrm{Cov}\,(Z\omega, Z\omega_l) = 0, l = 1, 2, \ldots, j-1. \end{aligned} \tag{2-11}$$

上式即为 PLS 模型中求第 j 个 PLS 分量的方法, 将求出的分量与原始变量组合得到新的回归变量, 我们尽可能的选取合适的 PLS 分量数目, 即为确定 M 的值, 以保证在消除原始自变量之间多重共线性的同时能够充分的保留原始自变量与相应变量的信息.

在得到了降维线性模型 PCR 与 PLS 各自的降维方法以及模型设定之后, 在实际的应用中当 M 确定时, 对式 (2-9) 采取简单的最小二乘回归即可进行模型参数的估计以及用模型进行预测. 因此, 降维线性模型的最优化策略选择和最优化计算算法仍然遵从简单线性回归模型的设定.

2.1.4　集成学习模型

对于众多的机器学习模型, 评判其模型好坏的方法即为考察其模型的拟合能力及泛化能力, 就单一的树形模型来讲, 如 CART 决策树模型, 模型自身的拟合能力和泛化能力不可兼得, 如果想要在某方面得到提升, 就必须以牺牲另一方面作为代价, 而集成学习模型则克服了这一弊端. 所谓集成, 顾名思义就是将若干个单一的弱学习器进行组合, 构成一个集成的强学习器, 不仅能够处理分类问题, 还可以处理回归问题. 因此, 集成学习秉承着“众人拾柴火焰高”的原则, 在机器学习领域有着较为突出的地位.

在集成学习中, 我们将构成总体模型的单个模型称为“个体学习器”, 按照个体学习器生成的方法以及各个个体学习器之间关联度的大小, 集成学习模型可以分为序列化集成学习模型和并行化集成学习模型. 并行化方法的代表模型如 Bagging、随机森林 (RF) 和极端随机树模型 (ET) 等, 而序列化方法的代表模型如 AdaBoost、提升树 (Boosting Tree) 以及梯度提升树 (GBDT) 等. 我们设定个体学习器皆为同质化的决策树模型, 首先来看随机森林 (RF) 模型.

随机森林 (RF) 之所以可以归类为并行化的集成学习模型, 是因为生成的森林中各个树之间的关联程度较弱, 大致可以将它们看成是相互独立的. 回归问题中随机森林的模型设定为:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b). \tag{2-12}$$

上式中 $\hat{f}_{rf}^{B}(x)$ 为随机森林模型的预测结果, B 为森林中树的个数, $T(x;\Theta_b)$ 代表第 b 棵决策树, 其中 Θ_b 为第 b 棵决策树中各个节点变量的集合. 值得注意的是, 在确定了森林中的每棵树之后, 由于各个决策树之间性能相近, 故而采用式 (2-12) 的平均法将每棵树结合, 即集成则体现在每棵树相结合这一步. 用来训练每棵树 $T(x;\Theta_b)$ 的训练集采用自助采样法 (boostrap sample) 从总体中随机产生; 而在生成决策树时, 在每个待分裂的结点上, 随机的选取 m 个属性, 从这 m 个属性中选取分类效果最优的属性来划分该结点. 因此, 随机森林的随机性体现在上述两个方面.

在本文中, 我们采用的另一种集成学习的方法为梯度提升回归树 (GBRT). 梯度提升树 (GBRT) 是提升树 (Boosting Tree) 的一种推广, 提升树是典型的序列化集成学习模型, 所谓序列化是指在模型学习中, 后面生成的树是在前面生成树的基础上而来的, 所以每棵树之间有较强的联系. 不过, 提升树的总体模型设定与前述的 RF 相似, 均为若干个单一决策树的加性模型:

$$\hat{f}_{GBRT}^{B}(x) = \sum_{b=1}^{B} T(x;\Theta_b). \tag{2-13}$$

同样地, $T(x;\Theta_b)$ 代表第 b 棵决策树, Θ_b 为第 b 棵决策树的参数. GBRT 中的单个决策树 $T(x;\Theta_b)$ 生成过程采用迭代式的前项分步算法, 在前一棵树的基础上不断的更新式 (2-13) 所示的总体加性模型, 即为:

$$f_{GBRT}^{b}(x) = f_{GBRT}^{b-1}(x) + T(x;\Theta_b) \ , \ b = 1,2,\ldots,B. \tag{2-14}$$

从式 (2-14) 可以看出, 在迭代算法的第 b 步运算中, 已知第 $b-1$ 步所生成的树, 所以, 第 b 步只需求出该步生成树的参数 $\hat{\Theta}_b$. 此处则用平方误差损失函数作为最优化策略, 来求解第 b 棵生成树的参数:

$$\begin{aligned}
\hat{\Theta}_b &= \arg\max_{\Theta_b} \sum_{i=1}^{N} L[y_i, f_{b-1}(x_i) + T(x_i;\Theta_b)] \\
&= \arg\max_{\Theta_b} \sum_{i=1}^{N} [y_i - f_{b-1}(x_i) - T(x_i;\Theta_b)]^2 \\
&= \arg\max_{\Theta_b} \sum_{i=1}^{N} [r - T(x_i;\Theta_b)]^2,
\end{aligned} \tag{2-15}$$

其中 $i = 1,2,\ldots,N$ 表示训练数据集中的第 i 个样本, 记 $r = y_i - f_{b-1}(x_i)$, 表示第 b 步模型拟合的残差. 所以在求解该步的最优化参数时, 只需拟合上述残差 r 即可, 此方法也大大简化了运算, 提高了模型训练的效率. 对于上述平方误差损失函数来讲, 用残差 r 进行最优化时每一步比较容易实现, 但当损失函数为一般函数时,

就需要重新考虑最优化的策略, 一个可行的办法是将式 (2-15) 中的残差用损失函数的负梯度

$$-[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}]_{f(x)=f_{b-1}(x)}, \tag{2-16}$$

近似代替, 式 (2-16) 使得最优化策略中一般损失函数的设定问题就得以有效解决, 这就是随机梯度回归树 (GBRT) 算法.

2.1.5　神经网络模型

神经网络模型优良的模型拟合能力和预测能力相较于其他机器学习模型, 在各个领域应用相当广泛. 如前述所说, 近年来对实证资产定价领域的研究趋向于非线性模型的拟合, 而神经网络模型正是基于其高度的灵活性和复杂的计算开销, 使得样本数据能够充分的被发掘出有效的信息, 进而可以更好的拟合预测模型. 神经网络模型由感知机模型发展而来, 神经网络中多层的学习模型在学习中的表现, 远远优于感知机单层的学习模型.

进一步来讲, 神经网络模型的复杂度取决于其隐藏层的层数和每层中神经元的个数, 神经网络越复杂, 其参数就越多, 计算开销也越大. 但训练数据的大幅增加再加以复杂的训练模型, 会使得训练效果得以大幅提升, 因此基于较强计算机硬件的支持, 深度学习也应用而生. 由于 Gu 等人[2] 在文献中指出, 金融数据受限于数据体量小且白噪声影响较大, 使得过于复杂的神经网络模型或是深度学习模型难以发挥其优势, 所以文中也仅用隐层数为 1～5 的前馈式神经网络模型 (NN1～NN5) 进行模型的拟合.

在本文中, 我们所采用的前馈式多层神经网络模型的输入层即为股票因子数据, 其一个或者多个隐藏层则用于处理这些股票因子的交互, 输出层则输出模型的预测结果. 我们以 NN3 神经网络为例, 其模型设定为如下形式:

$$\hat{f}_{NN3}(z_{i,t}) = \alpha_1 + \boldsymbol{W}_1\sigma(\alpha_2 + \boldsymbol{W}_2\sigma(\alpha_3 + \boldsymbol{W}_3\sigma(\alpha_4 + \boldsymbol{W}_4 z_{i,t}))) + \varepsilon_{i,t+1}, \tag{2-17}$$

其中, $\{\alpha_1,\ldots,\alpha_4\}$ 表示各层的偏差, $\{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_4\}$ 表示各层的权重矩阵, $\sigma(\cdot)$ 为激活函数. 对于激活函数的设定, 我们选取 ReLU 函数作为模型的激活函数, 其函数形式如下:

$$\mathrm{ReLU}(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{otherwise .} \end{cases} \tag{2-18}$$

在模型训练的过程中, 我们采用最小化 l_2 惩罚来作为最优化策略. 最后, 采用随机梯度下降法 (SGD):

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta\frac{\partial L}{\partial \boldsymbol{W}}, \tag{2-19}$$

作为最优化算法来训练模型. SGD 中的一个关键参数为学习率 η, 它控制着梯度下降的步长, 所以设定合适的 η 也对我们的模型训练准确度及速度有着显著的影响.

2.2 总体模型

2.2.1 因子模型

上世纪六十年代, 随着资本资产定价模型 (Capital Asset Pricing Model,CAPM) 的提出, 资本资产的价格在考虑市场风险的基础上得以衡量, 尽管 CAPM 模型的有效性在其后的实际应用中也遭到质疑, 但是其简洁的形式和较强的解释性无疑开创了资本资产定价理论的先河, 我们可以看到在 CAPM 理论中, 资产的超额收益率由一个简洁的一元线性模型决定:

$$\mathbb{E}(r_i) - r_f = \beta_i \left[\mathbb{E}(r_m) - r_f\right]. \tag{2-20}$$

上式中, $\mathbb{E}(r_i)$ 表示资产 i 的预期收益率, 其与无风险利率 r_f 的差值表示资产 i 的超额收益率, 超额收益率与该资产收益对市场收益的敏感程度 β_i 有关, 此处 β_i 也被称为是资产 i 对市场风险的暴露程度, $[\mathbb{E}(r_m) - r_f]$ 则表示市场处于均衡状态时市场组合的风险溢价, 式 (2-20) 这个经典的形式为后来涌现的多因子线性模型提供了基础. 随着市场的不断完善和发展, 研究者认为影响实证资产超额收益率的因素并非是由单一的市场因子所决定的, 于是提出了可以纳入多个影响资产超额收益率因素的模型, 即套利定价理论 (Arbitrage Pricing Theory, APT), 模型如下:

$$\mathbb{E}(r_i^e) = \boldsymbol{\beta}_i^{'}\boldsymbol{\lambda} + \alpha_i, \tag{2-21}$$

其中, $\mathbb{E}(r_i^e)$ 表示资产 i 的预期超额收益率, $\boldsymbol{\lambda}$ 为因子预期收益, $\boldsymbol{\beta}_i^{'}$ 为因子暴露, α_i 为资产 i 的定价误差, 是所选因子不能解释的那部分价格变动.

对于股票市场, 式 (2-21) 反应了股票因子对于股票超额收益率的预测关系, 在式子的右端我们可以加入任意对股票预期收益率产生影响的因子. 虽然在实际的市场之中, 股票的收益率并非全部可以通过因子预期收益及因子暴露的线性组合所准确预测, 但是我们在甄别有效的预测因子以及寻找合适的预测模型时, 通常假定股票资产的定价误差 α_i 显著为零来简化传统的多因子定价模型, 即式 (2-21) 变成如下形式:

$$\mathbb{E}(r_i^e) = \boldsymbol{\beta}_i^{'}\boldsymbol{\lambda}, \tag{2-22}$$

传统的资本资产定价理论通常会从时间序列角度或者时点截面角度去预测资产未来的预期收益率, 常用的方法有组合排序法和回归法等. 在证实 α_i 显著为零的

情况下, 用式 (2-22) 来建模; 相反, 在证实 α_i 显著偏离零的情况下, 即考虑异象资产的模型中, 通常会以式 (2-21) 来建模.

2.2.2 总体预测模型

上一小节中的式 (2-21) 和式 (2-22) 给出了传统资本资产定价领域的主流模型, 在此基础上, 为了克服诸如因子数目繁多、非线性因素影响等问题, 研究者提出了以应用机器学习和深度学方法进行因子建模以及发掘有效因子的更为普适的总体模型, 即对资产的预期收益率做如下的假定:

$$r_{i,t+1} = \mathbb{E}_t\left[r_{i,t+1}\right] + \varepsilon_{i,t+1}, \tag{2-23}$$

其中 $\mathbb{E}_t\left[r_{i,t+1}\right]$ 表示资产 i 在第 $t+1$ 期的预期超额收益率, $\varepsilon_{i,t+1}$ 为误差项且假定其服从正态分布. 此外, 我们更进一步假定 $\mathbb{E}_t\left[r_{i,t+1}\right]$ 在时期 t 给定时, 为预测因子集合的一个常数函数:

$$\mathbb{E}_t(r_{i,t+1}) = f\left(\boldsymbol{z}_{i,t}\right), \tag{2-24}$$

其中, $\boldsymbol{z}_{i,t}$ 是 p 维预测因子向量. 联立式 (2-23) 和式 (2-24), 我们可以得到资本资产定价的任务就是找到如下测算实证资产超额收益率的模型:

$$r_{i,t+1} = f\left(\boldsymbol{z}_{i,t}\right) + \varepsilon_{t+1,i}. \tag{2-25}$$

本文的研究主要集中于中国 A 股市场, 因此 $\boldsymbol{z}_{i,t}$ 表示我们要考察的所有对股票预期收益率产生影响的因素, 也称为股票的特征, $i = 1,\ldots,N$ 表示所选股票的索引且 $t = 1,\ldots,T$ 表示月份. 式 (2-24) 以其简洁的形式, 解释了股票预测因子与股票超额预期收益率之间的关系, 将股票的超额预期收益率看成是一个具有灵活性的函数 $f(\cdot)$, 是为了让总体模型能够适应上述所有机器学习模型的假定, 所以在本文中我们以式 (2-24) 作为总体模型进行建模.

2.2.3 因子模型与总体模型的关系

在前两个小节的描述中, 我们阐述了传统因子模型 (2-22) 和本文所设定总体模型 (2-24) 的各自的表达式, 从表达式中可以看出两种模型都旨在测算出实证资产的超额预期收益率, 只是在其形式上略有不同. 但在股票市场中给定其股票特征的情况下, 总体模型实际上嵌套了因子模型的多种形式. 如果只是考虑简单的情况, 即模型中只包含 n 个股票因子, 计 $\boldsymbol{z}_{i,t}$ 为 $n \times 1$ 维的股票因子向量, 设模型 (2-22) 中的因子暴露满足 $\boldsymbol{\beta}_{i,t} = \theta_1 \boldsymbol{z}_{i,t}$ 且因子收益率向量满足 $\boldsymbol{\lambda}_t = \theta_2$ 时, 则有:

$$f\left(\boldsymbol{z}_{i,t}\right) = \mathbb{E}_t(r_{i,t+1}) = \boldsymbol{\beta}_{i,t}^{'}\boldsymbol{\lambda}_t = \boldsymbol{z}_{i,t}^{'}\theta_1^{'}\theta_2 = \boldsymbol{z}_{i,t}^{'}\theta. \tag{2-26}$$

上式中的 θ_1 和 θ_2 均为常数矩阵. 可以看到, 传统的多因子模型变为简单的线性回归模型, 此间只是对多因子模型的参数做了符合实际的假设. 如果我们所要考虑的股票的特征为个股股票因子和宏观经济变量的交互作用, 且设模型中包含 m 个宏观经济因子 ($\boldsymbol{x}_t$)、n 个股票因子 ($\boldsymbol{c}_{i,t}$), 则有:

$$\boldsymbol{z}_{i,t} = \boldsymbol{x}_t \otimes \boldsymbol{c}_{i,t}, \tag{2-27}$$

其中 $\otimes$ 表示克罗内克积, 此时 $\boldsymbol{z}_{i,t}$ 为 $n \cdot m \times 1$ 维向量. 同样地, 设模型 (2-22) 中的因子暴露满足 $\boldsymbol{\beta}_{i,t} = \theta_1 \boldsymbol{c}_{i,t}$ 且因子收益率向量满足 $\boldsymbol{\lambda}_t = \theta_2 \boldsymbol{x}_t$ 时, 仍有:

$$f(\boldsymbol{z}_{i,t}) = \mathbb{E}_t(r_{i,t+1}) = \boldsymbol{\beta}_{i,t}^{'} \boldsymbol{\lambda}_t = \boldsymbol{c}_{i,t}^{'} \theta_1^{'} \theta_2 \boldsymbol{x}_t = (\boldsymbol{x}_t \otimes \boldsymbol{c}_{i,t})^{'} vec(\theta_1^{'} \theta_2) = \boldsymbol{z}_{i,t}^{'} \theta. \tag{2-28}$$

这里的 $\theta = vec(\theta_1^{'} \theta_2)$ 表示拉直运算. 可以看出, 考虑股票特征间的交互作用时, 总体模型 (2-24) 仍然嵌套了传统多因子模型, 当总体模型不再限制为简单的线性模型时, 其总体模型的形式也更加灵活, 例如考虑广义线性模型或者更为复杂的神经网络模型时总体模型仍能够表现出较强的灵活性和普适性.

通过上述的举例论证, 我们可以将股票的特征集进行进一步扩展, 使所要考虑的股票特征不仅包含个股的股票因子, 还包括宏观经济变量以及行业虚拟变量, 再加入股票因子与宏观经济因子之间的交互作用, 即为, 设模型中包含 n 个股票因子 ($\boldsymbol{c}_{i,t}$)、m 个宏观经济因子 ($\boldsymbol{x}_t$)、j 个行业虚拟变量 ($\boldsymbol{d}_{i,t}$), 则股票的特征向量 $\boldsymbol{z}_{i,t}$ 表示为:

$$\boldsymbol{z}_{i,t} = \begin{pmatrix} \boldsymbol{c}_{i,t} \\ \boldsymbol{x}_t \otimes \boldsymbol{c}_{i,t} \\ \boldsymbol{d}_{i,t} \end{pmatrix}. \tag{2-29}$$

上式中 $\boldsymbol{z}_{i,t}$ 为 $[n(1+m)+m] \times 1$ 维向量. 我们可以看到, 式(2-29)所表示的股票特征包含了股市中理论上决定股票收益率变动的各方面的因素. 至此, 我们仅仅是站在经济金融学的角度假设影响股票收益率的因素由诸如此类的特征构成, 而接下来的研究中, 我们将站在统计学的角度详细考察对股票收益率有显著影响的因素. 在本文中我们选择了 100 个股票因子 (其中 8 个个股因子为虚拟变量), 84 个行业变量, 13 个宏观经济因子, 所以式(2-29)中 $\boldsymbol{z}_{i,t}$ 的总体协变量总数为 $92 \times (13+1) + 84 = 1372$.

第三章　研究设计

3.1　特征选择

我们将影响股票收益率的因素也称作特征, 从前面的叙述中得知, 从经济金融学角度出发, 学者们已提出了成百上千种影响股票收益率的特征, 其中包括个股因子、宏观经济因子、行业因子等等. 在此节中, 我们将介绍在本文中我们所要考察的因素. 尽管已经经过实证检验过的因子种类令人眼花缭乱, 但是这些因子对股票未来收益率的影响程度不尽相同, 不仅如此, 不同的因子对股票收益率产生影响的持续性也各不相同, 从而导致预测股票收益率的因子之间也有着孰优孰劣的对比, 这也是本文所要研究的内容之一, 即为挑选出对股票收益率持续有效的因子. 我们与 Leippold 等人[34] 在文中所选择的因子集基本保持一致, 但是出于某些个股因子的数据集无法获取的缘由, 我们对文献的个股因子集稍作调整以作为本文的个股因子集. 因子集中的每个个股因子都经过了研究者的大量实证检验, 并且这些个股因子在经济金融学的角度都有着合理的解释, 具体的个股因子选取见表3-1所示.

个股因子数据来源于中国经济金融研究 (CSMAR) 数据库和 wind 数据库, 在 CSMAR 数据库中我们获取了截止 2022 年 4 月中旬 A 股市场上市的所有 4932 只股票的基本信息, 以及每只股票对应的自 1997 年 1 月到 2022 年 4 月的月度收益率. 由 4932 只股票构成的股票池作为本文的研究对象, 定量考察股票特征对股票池中所有股票收益率的影响程度.

在文章中, 我们搜集并整理了 100 个个股因子, 其中包含 8 个虚拟变量. 个股因子的选取参考了 Green 等人[1] 和李斌等人[44] 的研究结论, 但是本文中我们并未像文献中那样将个股因子划分诸如动量因子、价值因子等大类, 一方面是考虑我们所要验证的是单个因子的有效性, 另一方面是因子的类别并没有明确的划分规则, 所以因子类别划分在本文的研究中并未有显著的作用. 此外, 为了避免个股因子数据中异常值的影响, 我们参考 Gu 等人[2] 对个股因子数据逐期进行的横截面转化, 此标准化的方法是用 RankGuass 法将数据映射到 $[0,1]$ 区间内. 具体的映射方法为首先计算每个股票因子所对应的股票的秩, 然后将秩除以该股票因子所有非缺失值的数量, 然后再减去 0.5, 便得到了转化后的数据. 根据 Kelly 等人[45] 的研究, 此种标准化方法不仅降低了数据集异常值的影响, 而且在其稳健性分析中发现做此转化后与未做转化前, 分析的结果并无显著差异, 因此经过转化后的数据减少了计算量, 进而提高了模型训练的效率.

表 3-1　个股特征

编号	特征缩写	特征名称	特征英文全称	更新频率
1	absacc	应计项目绝对值	Absolute accruals	季度
2	acc	应计项目	Accruals	季度
3	agr	资产增长率	Asset growth	季度
4	beta	系统性风险	Beta	月
5	betasq	系统风险平方	Beta squared	月
6	bm	账面市值比	Book-to-market	季度
7	bm_ia	行业调整账面市值比	Industry-adjusted book to market	季度
8	cashar	现金资产比	Cash asset ratio	季度
9	cashdebt	债务现金流	Cash flow to debt	季度
10	opind	营运指数	Operating index	季度
11	cfp	现金流价格比率	Cash flow to price ratio	季度
12	cfp_ia	行业调整现金流价格比率	Industry-adjusted Cash flow to price ratio	季度
13	chato	总资产周转率	Asset turnove	季度
14	chato_ia	行业调整资产周转率变化	Industry-adjusted asset turnove	季度
15	chcsho	流通股变动	Change in shares outstanding	季度
16	chemp_ia	行业调整员工人数	Industry-adjusted change in employees	年
17	chinv	存货变化	Change in inventory	季度
18	chmom	动量变化	Change in momentum	月度
19	pm	息税前利润率	Profit margin	季度
20	pm_ia	行业调整息税前利润率	Industry-adjusted Profit margin	季度
21	chtx	税收增长率	Change in tax expense	季度
22	finer	财务费用率	Finance expense ratio	季度
23	currat	流动比	Current ratio	季度
24	depr	折旧率	Depreciation / PP&E	季度
25	divi	派息	Dividend	年
26	abcf	异常经营活动现金流	Abnormal cash flow	年
27	dolvol	交易额	Trading volume	月
28	dy	股利价格比	Dividend to price	年
29	anal	分析师关注度	Analyst attention	年
30	egr	股东权益增长率	Growth in shareholder equity	季度
31	gma	毛利率	Gross protability	季度
32	grCAPX	资本支出变化	Growth in capital expenditures	年
33	herfLn	行业销售集中度	Industry sales concentration	季度
34	hire	员工人数	Number of employees	年
35	idiovol	异质波动	Idiosyncratic return volatility	月
36	ill	流动性	Illiquidity	月
37	invest	资本支出与存货	Capital expenditures and inventory	年
38	lev	杠杆率	Leverage	季度
39	lgr	长期债务增长	Growth in long-term debt	季度
40	maxret	最大日收益率	Maximum daily return	月
41	mom12m	12 个月动量	12-month momentum	月
42	lagretn	短期反转	Short term reversal	月
43	mom6m	6 个月动量	6-month momentum	月
44	mom36m	36 个月动量	36-month momentum	月
45	ms	财务报表得分	Financial statement score	年
46	mve	规模	Size	月
47	mve_ia	行业调整规模	Industry-adjusted size	月
48	nincr	连续上涨日	Sustainable growth rate	季度
49	operprof	经营盈利能力	Operating protability	季度
50	orgcap	组织资本	Organizational capital	季度
51	pchcapx_ia	行业调整资本支出变化率	In.-adjusted change in capital expenditures	年
52	pchcurrat	流动比变化率	% change in current ratio	季度
53	pchdepr	折旧变化率	% change in depreciation	季度
54	pchgm_pchsale	毛利率变化-销售额变化	% change in gross margin - change in sales	季度
55	pchquick	速动比变化率	% change in quick ratio	季度
56	pchsale_pchinvt	销售额变化-库存变化	% change in sales - change in inventory	季度
57	pchsale_pchrect	销售额变化-应收账款变化	% change in sales - change in A/R	季度
58	pchsale_pchxsga	销售额变化-管理费用变化	% change in sales - change in SG&A	季度
59	pchsaleinv	营业收入存货比	% change sales-to-inventory	季度
60	pctacc	应计百分比	Percent accruals	季度
61	susgr	可持续增长率	Sustainable growth ratio	季度

(见下页)

编号	特征缩写	特征名称	特征英文全称	更新频率
62	accrub	会计稳健性	Accounting robustness	年
63	quick	速动比率	Quick ratio	季度
64	ldr	长期借款占比	Long-term loan ratio	半年
65	xsga_mve	管理费用市值比	SG to market capitalization	季度
66	xsga	管理费用率	SG to sale	季度
67	realestate	固定资产	Fixed assets	季度
68	volatility	收益率波动	Return volatility	月
69	roa	总资产收益率	Return on assets	季度
70	roavol	盈利波动	Earnings volatility	年
71	roe	净资产收益率	Return on equity	季度
72	roic	投入资本回报率	Return on invested capital	季度
73	rsup	意外收入	Revenue surprise	季度
74	cashsale	营业收入现金含量	Sales to cash	季度
75	invsale	存货与收入比	Inventory to sales	季度
76	revsale	应收账款与收入比	Receivables to sales	季度
77	sgr	营业收入增长率	Sales growth	季度
78	sp	销售价格比	Sales to price	年
79	std_dolvol	交易额波动	Volatility of trading volume	月
80	std_turn	换手率波动	Volatility of turnover	月
81	EVA	经济附加值	Economic value added	季度
82	stdcf	现金流波动	Cash ow volatility	季度
83	tang	偿债能力/总资产	Debt capacity / firm tangibility	季度
84	tb	税收收入与账面收入	Tax income to book income	季度
85	turn	交易换手率	Share turnover	月
86	ea	盈余激进度	Earnings aggressiveness	年
87	es	盈余平滑度	Earnings smoothing	年
88	largestholderrate	最大股东持股率	Largest shareholder ownership	季度
89	top10holderrate	前十大股东持股率	Top 10 shareholders ownership	季度
90	pchEVA	经济附加值变化	Change in EVA	季度
91	fc	融资约束系数	Financing constraints	年
92	R_coe	风险系数	Risk coefficient	年
93	soe	国有企业指标	State owned enterprise indicator	年
94	private	私营企业指标	Private enterprise indicator	年
95	foreign	国外企业指标	Foreign enterprise indicator	年
96	others	其他企业指标	Other enterprise indicator	年
97	MB_mk	主板指标	Main-Board Market	年
98	SM_mk	中小板指标	Small&medium board market	年
99	Next_mk	创业板指标	Second Board	年
100	STAR_mk	科创板指标	STAR Market	年

除了将个股因子作为模型的自变量之外, 我们也加入了股票的行业虚拟变量. 股票的行业分类遵循中国证监会 2012 年发布的行业分类规则, 其具体分类原则与方法见《上市公司行业分类指引 (2012 年修订)》. 每只股票所对应的公司以季度频率更新的行业分类数据来自 CSMAR 数据库, 对每只股票的行业分类结果截止于 2022 年 4 月, 按照行业分类指引, 所有的股票将被分为互不交叉的 90 个大类行业, 但是由于本文中研究的股票池中的股票所对应的公司并非涵盖了《指引》中划定的所有行业, 所以我们将股票池中的股票未涉及的行业剔除, 则剩下的 85 个行业是股票池中的股票均所涉及的. 由于行业变量为分类变量, 所以作为自变量引入回归模型时, 需要将其转化为虚拟变量. 分类变量在设置成虚拟变量的过程中, 为了避免虚拟变量与回归模型的截距项产生多重共线性问题, 虚拟变量的个数一般设置成比类别数少 1, 所以本文中我们总共设置了 84 个虚拟变量, 将这些虚拟变量作为影响股票收益率的特征而加入模型中.

在本文中, 我们加入的第三组特征则是反映宏观经济形势的宏观经济因子. 我们将引入模型的宏观经济因子大致划分为两种类型, 第一类是反映总体股票市场的股市类宏观因子, 例如 Welch 和 Goyal[9] 在其文章中提到的诸如股息率 (dp)、派息率 (de)、收益价格比 (E/P)、账面市值比 (bm)、指数波动 (svar)、净股本扩张 (ntis) 等指标, 以及 Baker 和 Stein[46] 在文章中应用的股票月成交量 (mtr) 指标, 都反映了总体股票市场的利好状况. 另一类宏观经济因子则反映了宏观经济的状况, 一直以来, 股票市场作为市场经济的晴雨表, 反映着市场经济的实时状况和发展趋势, 与此同时, 宏观经济状况也对股票市场有着相同的“反作用力”. 所以许多研究者提出, 一些宏观经济指标会对个股的收益率产生影响, 例如 Welch 和 Goyal[9] 在文章中提到的期限利差 (tms)、通货膨胀 (infl)、国库券利率 (tbl)、违约利差 (dfy) 指标, Chen[47] 在其文章中应用的 M2 增长率 (M2gr) 指标, Rapach 等人[11] 在文章中应用的国际贸易增长额 (itgr) 等指标.

上述文献中所提到的 13 个宏观经济指标, 由于其数据库的权限限制以及数据的时效性无法得到保证, 故我们将部分指标稍加改动, 改动后的指标相较于原文献中的指标, 并未改变其指标性质以及对股票收益率的影响水平. 我们最终搜集整理了股息率 (dp)、派息率 (de)、市盈率 (pe)、账面市值比 (bm)、股票方差 (svar)、股权扩张 (etis)、月成交量 (mtr)、期限利差 (tms)、通货膨胀 (infl)、无风险利率 (rf)、信用利差 (cds)、M2 增长率 (M2gr) 以及进出口额增长率 (eigr) 等 13 个宏观经济指标, 将这些宏观变量与个股因子相结合加入模型中, 具体的宏观经济变量的选取及其构建见表3-2所示. 宏观经济指标的的数据来源于 wind 数据库和国家统计局网站. 这里要指出的是, 宏观经济因子对股票收益率的影响体现在股

表 3-2　宏观经济变量

编号	特征缩写	特征名称	特征英文全称	更新频率	描述
1	m_dp	股息率	Dividend Price Ratio	月度	A 股股票现金分红总额 / A 股总市值
2	m_de	派息率	Dividend Payout Ratio	年度	A 股股票现金分红总额 / A 股总体收益
3	m_pe	市盈率	Price Earnings Ratio	月度	A 股股票总体价格 / A 股总体收益
4	m_bm	账面市值比	Book-to-Market Ratio	月度	A 股股票净资产总额 / A 股总市值
5	m_svar	股票方差	Stock Variance	月度	A 股综合市场每日收益的平方和
6	m_etis	股权扩张	Equity Expansion	月度	A 股股权募集资金总额 / A 股总市值
7	m_mtr	月成交量	Monthly Turnover	月度	A 股月成交金额 / A 股日均市值
8	m_tms	期限利差	Term Spread	月度	10 年期国债利率 - 1 年期国债利率
9	m_infl	通货膨胀	Inflation	月度	月度居民消费价格指数当月值
10	m_rf	无风险利率	Risk-free Rate	月度	月度化无风险利率
11	m_cds	信用利差	Credit Spread	月度	A 股综合市场收益率 - 10 年国债利率
12	m_M2gr	M2 增长率	M2 Growth Rate	月度	M2 月度增长率
13	m_eigr	进出口额增长	Growth in Export-Import	月度	进出口总额月度增长率

市的宏观层面, 也可以看作是对股票指数产生影响. 所以, 要将股票指数层面的趋势变化信息反应至个股中, 就要借助于上一章所提到的克罗内克积来实现. 将每只个股的个股因子与宏观经济变量做张量积, 股票的个股因子因此被扩张成兼具反应宏观经济指标的特征进而被加入模型. 我们将个股因子、行业虚拟变量以及

个股因子与宏观经济指标的克罗内克积均按照时间序列数据的格式进行整理, 并且认为这些特征都对股票的收益率产生影响, 即为, 这些特征通过模型的挑选, 可以直接参与实证资产的定价.

3.2 样本划分

参照李斌等人[44] 的做法, 我们将数据集按时间划分为训练集 (1997-2012) 和测试集 (2013-2022), 两个数据集用来进行模型估计、超参数选择和性能评价. 训练集中的数据主要用于模型的初次拟合, 在指定超参数后将训练数据集输入模型, 得到初次训练的结果；当模型初次拟合后, 我们根据拟合结果适当的调整超参数, 以选取最优拟合效果的超参数. 这里，我们在模型初次拟合时所指定的超参数主要来自于李斌等人[44] 以及 Leippold 等人[34] 的实证研究结论，在其基础上根据模型拟合的情况以及计算的复杂度再进行适当的调整。确定超参数与拟合模型之后我们得到了最终的预测模型, 进而我们将未参与模型训练的测试数据集带入模型, 以评估我们得到的模型的有效性.

除了样本的划分, 我们还考虑模型训练与预测过程中窗口滑动步长的设定. 李斌等人[44] 在文章中指出, 将训练数据的窗口滑动步长分别设置为 3 个月、12 个月和 24 个月时, 其各自拟合的模型之间的预测效果差异并不显著. 所以在本文中出于计算复杂度与预测模型稳健性因素的考虑, 我们将训练窗口滑动步长设置为 12 个月, 即为每 12 个月重新拟合一次模型, 而训练样本在最初的 16 年的基础上每次增加一年, 验证样本和测试样本每次向前滚动一年但实践跨度不变. 值得注意的是, 在训练及预测的过程中, 股票的因子数据要比月收益率数据在时序上滞后一月, 即为第 $t-1$ 月的股票因子数据与第 t 月的股票收益率数据配对, 构成一组“因变量—自变量”组合.

3.3 样本外预测评价

当数据集经过划分, 将训练集拟合出预测模型后, 预留出的测试集用于样本外 (Out-of-sample) 预测评价. 样本外预测的方法就是将未参与模型拟合的测试集数据带入到预测模型中, 然后借助于既定的算式量化的评价预测模型的好坏. 本文中采用的样本外预测评价指标为样本外 R^2, 也记为 R^2_{OOS}. 在具体的应用中, 当模型 S 给定时, 通过测试集数据所预测的未来一月收益率与实际值的差距则通过 R^2_{OOS} 反应出来, 计算公式如下:

$$R^2_{\text{oos},S} = 1 - \frac{\sum_{(i,t)\in\mathscr{T}} \left(r_{i,t} - \hat{r}^{(S)}_{i,t}\right)^2}{\sum_{(i,t)\in\mathscr{T}} r^2_{i,t}}, \tag{3-1}$$

其中, S 表示模型的类别, $\mathscr{T}$ 表示其用于计算的样本仅来自于测试集, $\{\hat{r}_{i,t}\}_{(i,t)\in\mathscr{T}}$ 为月收益率的预测值集合, $t=1,2,\ldots,m$ 为第 S 个模型在第 i 次更新后用于预测的未来一段时间, 是按月为单位. i 表示第 S 个模型的第 i 次更新, $i=1,2,\ldots,10$ 这里 i 最多取到 10 是因为, 我们的数据集中总共包含了 304 个月的数据, 模型初次拟合时以 16 年的数据作为训练集, 则训练集包含了 192 个月, 剩余的 114 个月按照每 12 个月一次的频率更新模型, 则只能更新 10 次.

样本外预测评价指标 R^2_{OOS} 的实质与线性回归的可决系数一致, 反应了可由模型做出解释的股票收益率的离差平方和占股票收益率平方和的比重, 但是此处与可决系的不同点在于, 其分母并未做中心化处理. 在单个股票收益率的预测评价中, 中心化处理存在缺陷, 因为股票超额收益率的历史均值其噪声较大, 并不能准确的反映股票超额收益率的平均水平, 去中心化处理反而损失了对模型评价的稳健型与准确性.

第四章　实证分析

在前三章的叙述中, 我们列出了本文需要探究的股票因子, 引入了机器学习模型、模型评价标准等等. 在接下来的叙述中, 我们依据实验的结果, 给出我们的实证分析.

本文所用到的数据集较大, 具体体现在以下两点. 其一是所搜集的有效股票数量在逐年增长, 如图4-1所示, 从 1997 年统计的 514 只有效股票增长到 2022 年的 4764 只有效股票, 所谓有效股票即为在统计期内股票收益率非缺失的股票; 其二是个股因子数据庞大, 每只个股都有 100 个个股因子, 加之个股因子与宏观因子的克罗内克积以及虚拟变量, 每只股票的因变量维度达到了 1372. 所以训练模型的数据集为高维数据集, 为了避免在拟合模型时出现维数灾难, 我们在遵循可解释性的前提下扩充训练样本, 即模型每更新一次, 训练数据集便增加一年.

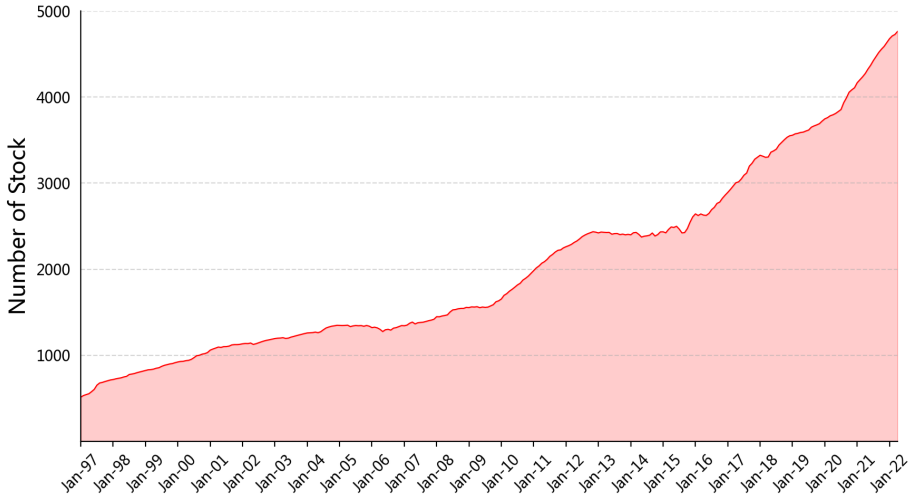

图 4-1　有效股票数量变化

此外, 我们并未剔除 ST 股, 停牌股等股票. 对于停牌股, 只是在训练时未将该股停牌阶段的因子数据纳入训练模型, 这也是一种扩充训练样本的手段. 但是扩充训练样本由此而带来了运算开销增加的问题, 为此申请了兰州大学超算平台的 40 核 CPU 计算节点以进行机器学习模型并行训练与预测, 60 核 CPU+8 卡 GPU 计算节点以进行神经网络模型训练与预测.

4.1　模型预测评价

4.1.1　总体预测评价

在实验阶段, 将所有的数据依次带入训练模型进行模型拟合, 并进行了预测. 当采用样本外 R^2 即 R^2_{OOS} 作为模型的评价指标时, 从表4-1中我们可以看出, 不同的机器学习模型有着不同的预测表现. 同样地, 当股票样本被划分为不同的类别时, 机器学习模型在其不同类别之间的预测结果也不尽相同. 首先我们将所有样本分为五个类别, 运用 13 种机器学习模型分别对五类样本进行训练并测量其拟合效果. OLS 和 OLS-3 模型分别为全部因子参与拟合的线性模型和只有 3 个因子参与拟合的线性模型, 三个因子分别为账面市值比、市值以及 36 个月动量, 可以看出, 全部因子集参与拟合的 OLS 模型的拟合效果甚至不如 OLS-3 模型的拟合效果, 其原因就在于高维因变量所带来的维数灾难和因变量之间的多重共线性的影响. 尽管 OLS 估计的效果不尽人意, 但是可以将其作为基准模型与其余的机器学习模型进行对比. 当所有样本参与模型拟合时, OLS 与 OLS-3 的 R^2_{OOS} 都取到了负值, 而带惩罚的线性模型 LASSO 与 Enet 的预测效果优于 OLS, 从 PCR 与 PLS 模型的预测结果来看, 降维模型的优势在高维数据中的优势凸显了出来, 计算密集型的树形模型的表现也较为出色, 其 R^2_{OOS} 的值都达到了相对较高的水平. 值的注意的是, 对于神经网络模型的拟合, 我们的实验结果与 Leippold 等人[34] 在文章中指出的一致, 当神经网络的层数逐步增加时, 其拟合效果也逐渐向好, 更进一步验证了中国股市相较于 Gu 等人[2] 所研究的美国市场而言, 具有其鲜明的特点. 从所有股票样本参与的模型拟合出发, 我们对比了不同模型对 A 股市场的股

表 4-1　总体样本外预测评价

样本划分	OLS +H**	OLS-3 +H	LASSO +H	Enet +H	PCR	PLS	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
所有样本	-0.26	-0.03	0.97	1.23	2.34	3.34	3.68	2.96	1.01	1.95	2.98	2.01	0.97
市值前 70%	-1.12	-0.11	-1.35	1.43	1.82	1.56	2.25	-0.06	-0.00	1.55	2.01	1.09	0.02
市值后 30%	0.03	1.34	1.92	1.28	2.54	4.01	3.91	2.89	0.00	2.07	4.68	4.99	3.11
股权占比前 60%	-2.90	-1.10	0.13	-1.33	0.95	0.71	1.67	1.88	-0.03	1.19	1.98	2.57	0.10
股权占比后 40%	-0.00*	0.18	-0.09	0.08	2.91	3.89	3.01	1.05	0.74	1.39	2.88	2.59	1.14

** 表示此机器学习模型的损失函数为 Huber 损失

* 表示 R^2_{OOS} 的值较为接近 0, 即第一位有效数字出现在小数点后第三位之后

票收益率的预测差异性, 接下来我们着眼于股票间的差异性, 按照差异将股票分为不同的类别, 进而划分不同的训练集进行模型拟合. 首先, 我们将所有的股票按照其市值的大小, 分成了股票市值在前 70% 的股票与市值处于后 30% 的股票. 从表4-1的实验结果来看, 在 A 股市场中, 市值较低的股票集合更具可预测性. 其中 NN3 与 NN4 模型在小市值股票的预测中均有较好的表现, 尤其是 NN4 模型的预测结果达到了 4.99%, 是一众机器学习模型中预测效果最好的模型. 也就是说, 在 A 股市场, 市值较小的股票相较于市值较大的股票更具有可预测性, 这实际上与

我们主观认为的小市值的公司具有更多的不稳定因素进而股价也相对不稳定的主观观念相悖, 这也是国内股票市场所特有的现象. 为了进一步探究模型拟合的效果与样本的划分有着一定的关系, 我们将表4-1可视化为图4-2所示的分组条形图, 从图4-2 I 中柱状图不同颜色的高低差异可以看出, 小市值股票的预测结果在各个机器学习模型上的表现几乎全部优于以所有股票为训练样本时的预测结果, 而将大市值股票作为训练样本进行拟合时, 其预测效果较差.

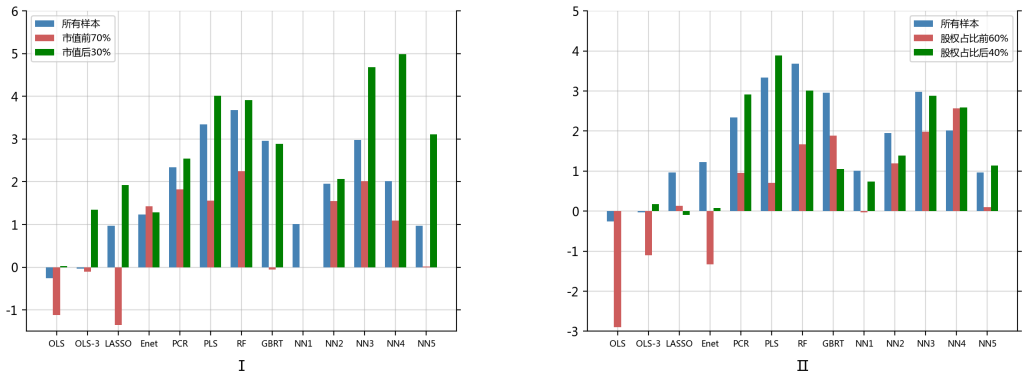

图 4-2　总体样本外预测评价图

国内股票市场中较多的散户投资者倾向于投资小盘股, 在主板市场中, 小盘股相对较低的股价更适合散户投资者进行短线操作, 而购入门槛相对较高的大盘股则多被机构投资者所青睐. 本文为了研究散户投资者是否对小盘股的可预测性做出贡献, 我们按照股票的股权结构将股票划分为两类, 第一类是前十大股东股权占比在前 60% 股票, 第二类是前十大股东股权占比在后 40% 的股票, 前十大股东占比越大则意味着股票的流动性较弱, 更倾向于被机构投资者投资. 从表4-1的实验结果来看, 在按股权占比分类的数据集中训练的结果整体上不如按市值大小分类的数据集中的训练结果, 前十大股东股权占比后 40% 的股票集合中 PLS 模型的训练与预测结果拔得头筹, 而神经网络模型的表现反而落了下风. 从图4-2 II 中可以看出, 用前十大股东股权占比在后 40% 的股票集合所训练的模型的可预测性明显较强, 其柱状图的高度完全高于以前十大股东股权占比在前 60% 的股票所训练的模型. 据此可以得出散户投资者对于小盘股的投资倾向使得 A 股市场中的小市值股票有了较好的可预测性, 并且可以看出, 降维的机器学习方法对前十大股东股权占比在后 40% 的股票有着良好的预测效果, 其预测结果甚至优于计算密集型的树形模型和神经网络模型.

从图4-2中可以看到, 将这几类机器学习方法运用于各个不同类型的数据集中时, 预测的结果虽是有所差别, 但是不同模型之间预测差异的变化趋势相似, 反应在图中可以看到, 除了将所有样本参与模型拟合的柱状图作为基准对比, 在图4-2的两幅图中均有出现外, 其余四类样本的柱状高度轮廓大致相似, 但是其水

平有所差别. 这也说明我们在文中选取的几类机器学习模型在其模型拟合与预测中, 均表现出了较好的稳定性.

4.1.2　分股权性质预测评价

在前一小节的叙述中, 我们探究了 A 股市场出现“异象”的原因之一, 正如 Leippold 等人[34] 所指出的那样, 国内股市中散户对于小盘股的追逐, 使得中国股票市场中小盘股的可预测性强于国外股票市场, 而且综合几类机器学习模型, 总体股票市场的可预测性较之于国外市场也相对较强. 抛除小市值股票的影响, 仅仅集中于对市值占比前 70% 的股票进行拟合及其预测, 所得到的预测结果也是优于 Gu 等人[2] 所研究的美国市场, 尽管 Gu 等人[2] 的研究在股票的样本量上占优. 所以我们更进一步探究出现此种情况的原因, 将国内市场与美国市场稍加对比就会发现 A 股市场中股票的股权性质比较复杂, 在上世纪 90 年代以前, 国内企业基本都是国营企业, 随着市场经济的不断发展与金融改革的深化, 以及民间资本的扩张与外资企业的入驻, 国内企业由原先国营企业一家独大的局面逐渐转变为企业性质多元化的局面. 图4-3所示为 2003 年 12 月至 2022 年 3 月间不同股权性

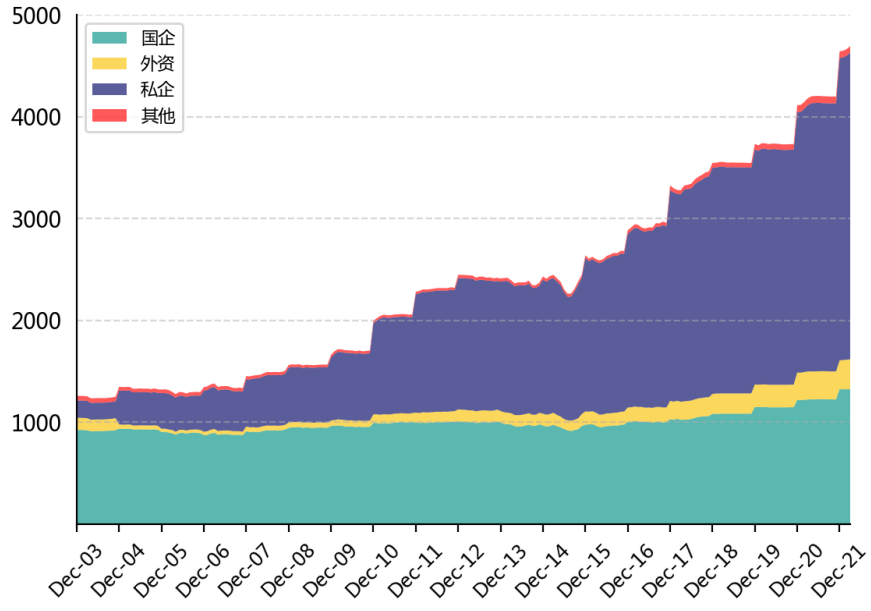

图 4-3　不同股权性质股票数量变化

质股票的数量变化, 之所以从 2003 年开始统计股票的股权性质, 是因为自上世纪末中国证券市场开启以来, 随着市场制度的日趋完善, 直到 2003 年有关公司股东情况披露的相关要求中, 上市公司除了要介绍公司控股股东的情况外, 还必须介绍公司实际控制人的情况. 所以按照公司所披露的实际控制人的性质, 我们将国内企业可以划分为国有企业、私营及民营企业、外资企业以及其他类型企业. 从图4-3中可以看出从 2003 年到 2022 年统计期末, 国有企业的数量基本保持不变,

其他三类企业的数量均保持增长, 但是外资企业与其他性质企业的占比明显少于另外两类企业.

从表4-2中可以看出, 将所有样本集合作为基准与其他集合做对比, 可以直观的看出 R^2_{OOS} 的最大值出现在了以国有企业样本数据为训练集并且机器学习模型为神经网络的情形之中, 由此说明神经网络模型在国有企业集合中的预测性能较为出色. 其次, 国有企业集合在带惩罚的线性模型中效果较好. 但国有企业在树型模型的表现堪忧, 而民营企业与外资企业集合在树型模型中的表现优于国有企业集合.

表 4-2　不同股权性质样本外预测评价

样本划分	OLS +H	OLS-3 +H	LASSO +H	Enet +H	PCR	PLS	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
所有样本	-0.26	-0.03	0.97	1.23	2.34	3.34	3.68	2.96	1.01	1.95	2.98	2.01	0.97
国有企业	-0.08	1.08	1.92	1.87	1.02	2.14	0.13	0.39	1.13	1.08	3.84	3.80	2.01
民营企业	-3.19	1.13	1.01	0.97	1.90	1.39	2.01	2.88	1.17	2.01	2.63	2.19	1.83
外资企业	-3.05	-0.19	1.23	2.09	2.12	1.99	1.90	0.57	0.11	0.13	2.03	2.35	1.18
其他企业	-5.01	-2.18	-2.72	-0.20	0.00	0.11	-1.09	0.01	-1.32	0.57	1.97	0.13	0.04

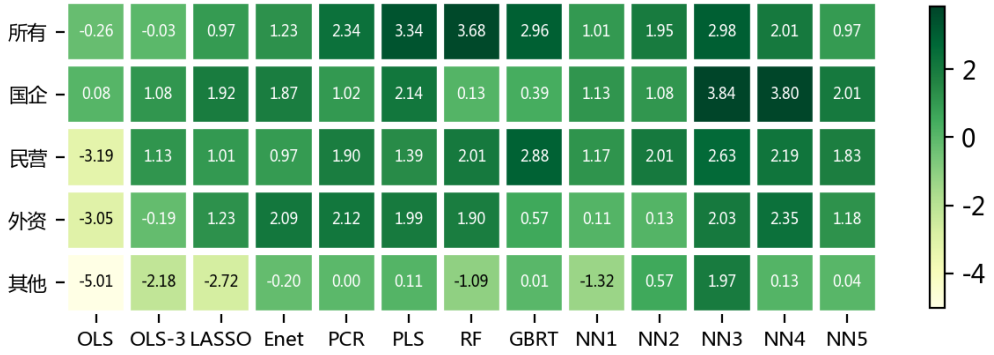

图 4-4　不同股权性质样本外预测热力图

机器学习模型的拟合效果、预测结果与样本量的大小密切相关, 通常来讲, 在其他条件相同的情况下样本量越大, 模型的预测准确性越高, 所以当我们运用 13 类模型对不同股权性质的股票集进行拟合及其样本外预测时, 在不同股票集合上, 模型的预测效果的整体评价能反映出不同股权性质股票集合之间的内在差异. 我们用表4-2中的数据生成热力图来直观的探究不同股票集合中模型预测性能的差异, 从图4-4可以看出, 包含所有样本股票集合与国有企业股票集合在热力图中的颜色明显较深, 因此前两类股票集合在各个模型的预测效果整体上都是优于后面三类股票集合的. 对比所有样本股票集合、国有企业股票集合以及民营企业股票集合, 其中所有样本股票集合在热力图中色差差异较为温和, 即在各个模型的表现相对稳定, 这与所有样本参与模型拟合时的样本量较大有关; 而国企集合与民企集合的样本量差别不大, 但是国企在各个模型中预测的稳健性优于民企. 从热力图的纵向来看, 国企在神经网络中的表现始终是优于其他模型的, 尤其是是在 NN3 与 NN4 中的表现. 这说明, 在国内市场中, 国有企业股票较其余性质的企业

有着更好的可预测性, 正如 Leippold 等人[34] 在文中指出的那样, 这样的预测结果实际上是出乎意料的, 私营企业的 IPO 数量和规模近年来在逐步扩大, 私营企业因其较高的流动性和资产配置的灵活性在其市场中逐渐占据着主导地位, 从我们主观的印象来看, 似乎私营企业的股票无论是在财务数据的披露以及股票价值信息的反映, 较之于国企更加透明和准确, 因此更具可预测性. 但是我们的实证研究表明, 国有企业在可预测性方面, 表现的更好. 这样令人出乎意料的结果显示出了国有企业独特的特性, 是否是国有企业的股权性质使得国有企业在收益率预测中表现出较好的性能? 这一点我们将进行进一步探究.

我们将股票集合划分为国有企业股票集合与非国有企业股票集合, 仍然以所有样本集合作为基准集合参与对比. 非国有企业股票集合包含民营企业、外资企业以及其他企业, 我们从表4-3中看出, 非国有企业集合的可预测性明显低于国有企业的可预测性, 其在某些模型中的表现, 甚至低于将非国有企业集合合并之前的单类集合之中的表现. 这样的结果与我们前面的分析一致, 而且在神经网络模

表 4-3　国企与非国企股票集合样本外预测评价

样本划分	OLS +H	OLS-3 +H	LASSO +H	Enet +H	PCR	PLS	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
所有样本	-0.26	-0.03	0.97	1.23	2.34	3.34	3.68	2.96	1.01	1.95	2.98	2.01	0.97
国有企业	-0.08	1.08	1.92	1.87	1.02	2.14	0.13	0.39	1.13	1.08	3.84	3.80	2.01
非国有企业	-2.90	-0.17	-0.00	1.02	0.40	1.01	1.65	2.79	0.91	1.02	2.05	1.13	0.91

型中表现出来的预测差异尤其显著, NN4 模型在国有企业中的表现我们在上述分析中已经强调, 在表4-3中, 我们可以看到所有的神经网络模型在国企集合中的表现均优于非国企集合. 按照国有企业的性质与特点, 我们认为, 国有企业的股票价格并非完全由市场因素影响, 国家的经济环境、货币政策、财政政策以及其他政府行业政策都会影响到国有企业的股票价格与未来股票价格的走势. 国有企业在复杂机器学习模型中的预测性能表现较好是因为, 上述的种种影响股票价格的非市场性因素需要更为复杂的模型加以拟合, 简单的线性模型难以刻画国有企业集合的预测特性.

4.1.3　分板块预测评价

在前两个小节中, 通过对比 R^2_{OOS}, 分析了国内 A 股市场中出现异象的原因. 从纵向维度, 即股票的集合划分来看, 差异性主要表现在小盘股与国企的集合划分中; 从横向维度看, 即机器学习的不同模型之间, 预测差异性主要来自于神经网络模型与其余的机器学习模型之间. 在本小节, 我们进一步研究国内股市场中不同的板块是否对股票的可预测性产生影响. 为了让上市公司在市场中得到最有利于公司发展的融资渠道与发展条件, 国内 A 股市场划分为主板、中小板、创业板以及科创板四类主要的市场板块, 其不同板块之间在公司上市门槛、交易规则、

监管制度、投资者条件以及投资风险方面都有着较大的区别, 所以研究不同板块股票集合可以进一步探究板块之间的差异是否与 A 股市场所表现出的“异象”有关. 图4-5显示了 A 股市场中各个板块的股票数目及变化情况. 由于 A 股市场的板块机制引入并非是一蹴而就, 期间也经历了种种改革进而逐步完善, 被誉为宏观经济晴雨表的主板市场, 自国内股市设立以来, 就是股票市场的主要板块, 出现的时间最早, 包含的股票也最多; 其次是初设于 2004 年 5 月的中小板市场, 上市于此板块的企业是相对于主板市场而言市值较小且达不到主板市场上市要求的企业, 实际为创业板的一种过渡, 值得注意的是中小板在 2021 年 2 月被批准合并于深交所主板市场; 第三类则是创业板市场, 设立于 2009 年 10 月, 作为创业型公司的孵化器, 其主要目的是为刚成立的高成长、中小型创业公司提供规范的融资环境与长效发展机制, 创业板的企业数量近年来逐步扩大, 但是相比于主板市场, 其

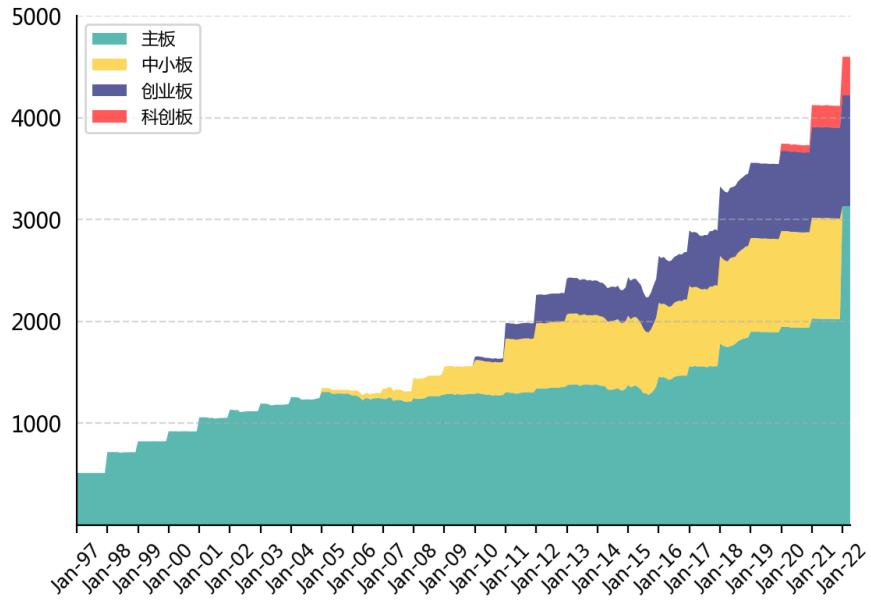

图 4-5　不同市场板块股票数量变化

规模相对较小; 最后一类是于 2018 年 11 月设立的科创板, 其目的在于只持高新技术企业的发展, 但是科创板由于创立时间较晚, 企业的数量有限, 在用机器学习模型进行拟合时, 由于因变量的数量较多而容易出现维数灾难, 使得预测效果不具有可解释性, 所以我们将这部分股票予以剔除. 分别将上述三类板块的股票构成三类股票集合, 运用机器学习模型进行拟合从而测定各个类别在模型中的表现, 以分析不同的板块之间股票的可预测性存在的差别.

我们将股票集合分为主板市场股票集合、中小板市场股票集合以及创业板市场股票集合, 按照类似于上一小节的分析方法, 分别对这三类板块的股票集合进行模型拟合及预测. 由于不同的板块市场成立的时间不尽相同, 因此我们按照各自的成立时间并以此为起始时段, 按照年度跨度滚动拟合模型. 尽管 A 股市场中,

每只股票所对应的企业之间存在着种种差异, 甚至是所处同一个板块的股票在其股权性质、市值大小、股权占比情况等等方面都有差异, 但是同一板块市场内的企业都符合其板块的限定条件, 这多多少少体现了板块内部各个股票之间的“同质化”, 因此不同板块市场的股票集合, 其预测的差异性能够很好的体现出不同板块市场之间的特征. 从下表4-4中我们可以看出, 不同板块市场的股票集合, 经由不同机器学习模型分别拟合并预测, 得出的样本外预测评价指标 R^2_{OOS} 值的差异较为明显. 同样地, 我们仍然将所有股票集合数据集作为基准, 将其与其他股票集合

表 4-4　不同板块市场样本外预测评价

样本划分	OLS +H	OLS-3 +H	LASSO +H	Enet +H	PCR	PLS	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
所有样本	-0.26	-0.03	0.97	1.23	2.34	3.34	3.68	2.96	1.01	1.95	2.98	2.01	0.97
主板市场	0.00	0.05	2.08	2.00	1.83	2.97	2.95	3.08	0.97	1.34	3.01	2.19	1.71
中小板	-2.35	1.04	2.12	1.77	3.19	2.81	2.01	2.82	0.91	0.93	3.06	3.94	2.08
创业板	-1.07	1.91	2.09	2.88	0.05	0.91	1.33	1.76	-0.10	0.92	0.12	1.10	0.73

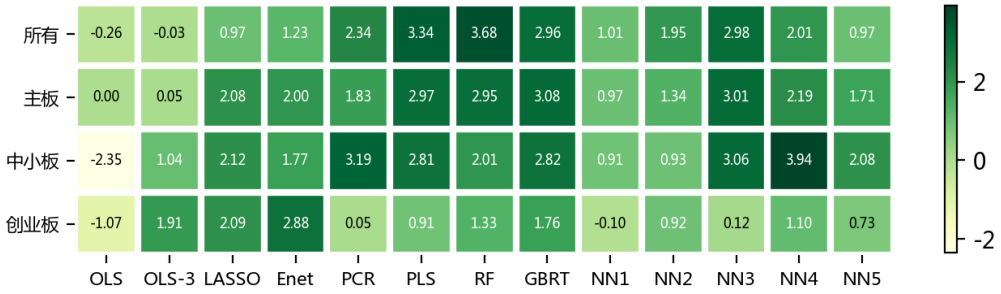

图 4-6　不同板块市样场样本外预测热力图

作对比. 从表4-4中我们看出, 主板市场中虽是国有企业居多, 但是其神经网络模型在主板市场股票集合中的预测结果并非全局最高, 这与上一小节中以股权性质为分类标准而划分股票数据集后所进行拟合而产生的预测结果有所不同, 尽管如此, NN3 与 NN4 模型在主板市场股票集合中的表现, 相较于其他模型仍是最优的. 这里我们值得关注的则是中小板市场在 NN4 模型中的表现, NN4 模型不仅在中小板市场中 R^2_{OOS} 的值达到最高, 而且在表4-4所示的全局范围内 R^2_{OOS} 的值也是最高的, 这说明中小板市场有其特殊性, 我们在接下来将着重探讨这一发现.

从热力图4-6直观的观察各个集合在不同模型中预测的差异程度, 所有样本数据集与主板市场股票集合在各个模型之中的 R^2_{OOS} 差异水平不大并且轮廓相似, 两个集合在降维模型 PCR 和 PLS 以及 RF 模型中的表现不分伯仲, 但是有一点值得注意, 主板市场集合在各个模型中的 R^2_{OOS} 均大于 0, 从这里可以看出, 主板市场股票集合的数据集较之于其他集合, 具有模型预测上的稳定性. 主板市场中的企业绝大多数是处于发展成熟期的企业, 公司具有稳定的业绩表现, 其包含了国有企业与非国有企业的大部分公司, 多数是行业的头部企业, 所以在各个机器学习模型中的拟合及预测的表现则相对较为稳定. 同时, 以主板市场股票集合做拟合及预测的复杂度和难度相对国有企业集合而言较小, 所以我们可以看出, 其

在相对较为简洁的训练模型中反而显现出良好的预测性能, 如在带惩罚的线性模型和降维线性模型中都有较好的预测表现.

中小板市场股票集合在模型中的可预测性则出现较大的波动, 在 OLS 模型中的预测表现欠佳, 在 NN4 模型中的表现最好. 但从总体来看, 中小板市场股票集合的可预测性较高, 除了在 OLS 模型中的表现较差之外, 在其余的机器学习模型中都表现出相对较好的预测结果. 我们知道, 中小板市场中企业的流通盘大约都在 1 亿元以下, 其市值规模上还是与主板市场中的企业有着较大的差距, 所以中小板市场中的绝大多企业单从市值规模上看, 实际上是处于将所有股票样本按市值大小排布后的尾端. 加之中小板市场对于股票投资者没有任何限制, 所以中小板市场股票集合的特征就类似于前述所分析的股票市值在后 30% 的企业集合, 但与之不同的是, 中小板市场股票集合在不同模型中表现的波动性要明显高于市值占比后 30% 的股票集合.

最后则是创业板市场股票集合, 从图4-6中可以看到, 创业板市场股票集合在线性模型中的表现较好, 但是从总体来看, 其在各个机器学习模型中预测效果均处于低位. 创业板市场中企业的地位仅次于主板市场, 以美股的纳斯达克市场为例, 板块内的企业大多是处在成长阶段的高成长性企业. 开立于深交所的创业板市场由于股价高、交易制度限制、投资风险较大等因素, 其流动性较之于主板市场与中小板市场而言处于较低水平, 交易者进入创业板交易的门槛较高, 阻挡了大部分散户投资者, 所以根据前述的分析, 散户投资者参与较少的股票可预测性相对较弱. 在本小节中, 我们的分析进一步证实了前两个小节中所陈述的两方面的研究内容. 其一是市值小、流通性高的股票可预测性较强; 其二是国有企业股票集合在绝大多数机器学习模型中的预测效果要优于非国有企业集合. 在日常的投资活动中, 统计个股的市值在整个股市中的分位、个股的股权性质以及个股的股本结构信息是较为繁琐的, 而我们通过实证分析发现, 代替上述复杂流程的方法就是直接按照股票所属的板块市场进行个股的研究. 这个方法简便易行, 通过股票代码判定股票所处的板块市场, 进而可以判定该板块市场是否具有可预测性, 以及该板块中适用的机器学习模型为哪些, 这样的判定有助于我们更加准确选取出适应于该股票的预测模型. 也可以选出对该股票股价影响最大的个股因子与宏观经济因子, 这是我们在下文中所要探究的内容.

4.2　个股特征变量对比

4.2.1　国有—非国有企业个股因子预测能力对比

在 4.1 节中, 我们将股票数据集按照股票不同的特征分成不同的集合类别, 运用机器学习模型进行样本外预测评价, 分析了各个集合在机器学习模型中的样本外可预测性的差异, 找出了国内 A 股市场出现“异象”的原因. 而对于影响股票预期收益率的个股因子而言, 并非所有的个股因子在预测收益方面都表现的同样重要, 这不仅取决于个股因子所含有的信息量的不同, 而且还取决于预测模型以及股票集合的不同. 现在我们讨论哪些个股因子在相应的股票集合中所含的信息量较大. 我们将参与模型拟合的股票因子逐一设置为 0, 计算 R^2_{OOS} 的减少量来衡量个股因子的影响力. 在图4-7和图4-8中, 我们展示了国有企业样本与非国有企业样本中, 运用机器学习方法进行拟合和预测时的前十个最重要的个股因子水平, 这里所列的前十个个股因子种类并非在所有机器学习模型中都是一致的, 即每个机器学习模型在相同的股票集合中所得出的排在前十的个股因子不完全相同, 因此这里我们将每个个股因子在不同机器学习模型中的贡献率测量值求平均数, 然后按照其平均数的大小, 分别选出了在特定的股票集合中排在前十位的个股因子. 这样做的原因是, 平均数对每个个股因子在所有不同机器学习模型中的贡献率做了综合的考量. 通过这样的综合考量, 使得排在前十的个股因子满足了不同机器学习模型的挑选, 更能体现出这些因子的重要程度.

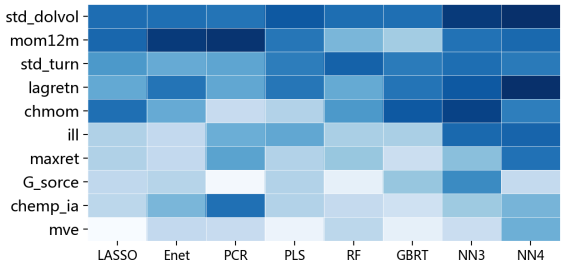

图 4-7　国企前十个股因子

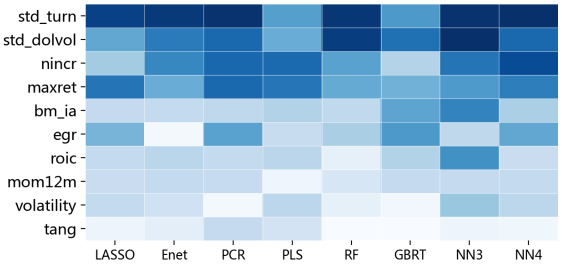

图 4-8　非国企前十个股因子

我们将国有企业与非国有企业分成不相交的两类股票集合来研究其各自影响力较大的个股因子, 并且在选出了所有模型中综合影响力最强的十种因子后, 再按照其各自的预测值减少量的测定值画出热力图, 这样做, 其一是为了分析国有企业与非国有企业在模型的可预测性之中的差异是由哪些个股因子所影响; 其二是从热力图中可以更加直观的看到哪些因子对于模型的适应性较好, 以便精准的选择有效因子, 这样的分析结果有助于投资者在实际的投资中, 根据所要投资股票的股权性质, 有针对性的选择准确的个股因子来帮助其进行投资决策; 其三是分析不同因子在不同模型中表现的差异性, 深入考察不同模型预测性能的驱动因素.

图4-7和图4-8中所示的个股因子在各个模型中的表现, 通过颜色的强弱来表达. 在国有企业与非国有企业股票集合中均排入前十的个股因子中, 有反应市场流动性、动量以及企业基本信息的三大类指标, 反映股票动量的因子如 12 个月动量 (mom12m)、动量变化 (chmom)、最大日收益率 (maxret) 以及短期反转 (lagretn), 这些因子在国有企业集合中对模型预测有较大的贡献, 这也说明在国有企业中动量因子对股票的价格变化起到重要的作用. 值的注意的是, 在国有企业集合中, 综合考量排名前十的因子在神经网络模型 (NN3、NN4) 中均有着较好的表现, 再一次证明了神经网络模型可以兼顾较多的因子, 从多个因子发掘出影响股票收益率的因素. 再观察图4-8, 在非国有企业集合中排在第一、第二位的交易额波动率 (std_dolvol) 和交易换手率波动性 (std_turn) 因子都反映了市场的流动性, 对非国有企业来说, 反应股票流动性的因子应该被受到重视. 此外, 流动性因子无论是在国有企业还是非国有企业, 都对股票的收益率有较大的影响, 这一点, 从图4-7和图4-8中都可以观察到, 而且这类因子在不同的机器学习模型中表现的也相对稳定. 与此同时, 连续上涨日 (nincr) 和最大期收益率 (maxret) 这两个动量因子对非国有企业股票集合的影响仅次于上述流动性因子, 所以动量因子的重要程度也不言而喻. 因此, 流动性因子与动量因子都是我们在衡量股票风险溢价时的重要参考因素.

4.2.2　不同板块市场个股因子预测能力对比

在前面的叙述中, 我们分析了机器学习模型在不同板块市场中的可预测性存在差别, 并且在前一小节证实了预测能力的差别源于不同样本集合中股票的个股因子对于股票未来收益率的贡献不同, 所以不同板块市场股票集合对股票收益率影响最大的前十个个股因子也有所不同. 图4-9至图4-12分别展示了主板市场、中小板市场、创业板市场以及总体市场中综合评价处于前十位的个股因子.

从图中首先我们可以看到, 在三类不同的板块市场中, 排在首位的个股因子仍然是与股票动量和市场流动性有关的因子, 这进一步说明了在本文所研究的所有个股因子中, 对股票的预测性影响最大是动量及流动性因子. 图4-9中显示了主板市场中的前十个因子, 由于主板市场中的股票以国有企业与大型民营企业居多, 所以其个股的特征类似于上一小节中提到的国有企业股票集合, 除了交易额波动率 (std_dolvol) 因子以外, 主要是以动量变化 (chmom)、12 个月动量 (mom12m) 和连续上涨日 (nincr) 等动量因素为主. 其中, 销售价格比 (sp) 作为主板市场中特有的预测因子出现在前十中, 与其相类似的反应股票价值的股利价格比 (dy) 也排入前十. 相对于主板市场, 中小板市场以小盘股居多, 所以从图4-10中可以看出, 中小板市场中交易额波动率 (std_dolvol) 和交易换手率波动性 (std_turn) 因子显现出

其重要性, 这与非国有企业集合中的情形相似. 此外, 经营盈利能力 (operprof)、资本支出变化 (grCAPX) 等反应企业经营情况的指标在中小板市场中的表现相对较好, 这一观察结果也符合 A 股市场中的散户投资者对于小盘股的追逐, 投资者对企业经营情况的关注使得这类因子表现出较好的预测性能. 与此同时, 在中小板市场中表现突出的因子还包括了收益率波动和盈利波动率这两个交易摩擦类因子, 这类因子主要由股票的收益数据所构造而来, 所以这类时效性较高的因子包含了更多的市场交易信息, 因此在中小板市场中被机器学习算法挖掘出来, 从图 4-9 中我们可以看到收益率波动因子在主板市场中也被重视. 众所周知, 创业板市场相对来说缺乏流动性, 所以在创业板市场中, 除了 12 个月动量 (mom12m) 因子之外, 预测效果显著的个股因子则集中于反应企业基本情况的个股因子之中, 如员工人数变化 (hire)、杠杆率 (lev) 和行业调整规模 (mve_ia) 等因子. 同时, 营业收入增长率 (sgr)、派息额 (divi) 以及资产增长率 (agr) 等成长因子, 反应着创业板市场中企业未来一段时间的发展前景与发展潜力, 因此这类因子对于创业板市场的影响也比较大.

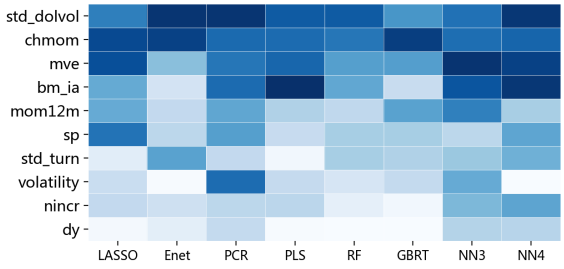

图 4-9　主板市场前十个股因子

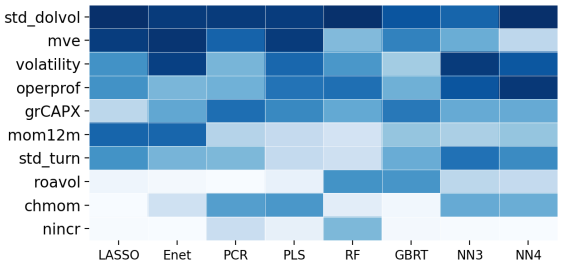

图 4-10　中小板市场前十个股因子

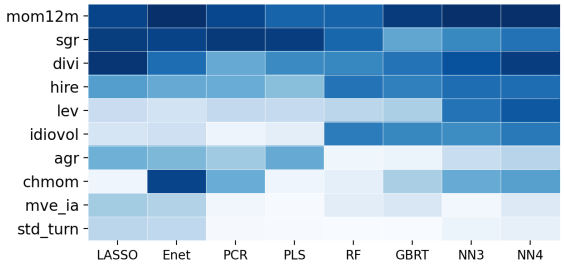

图 4-11　创业板市场前十个股因子

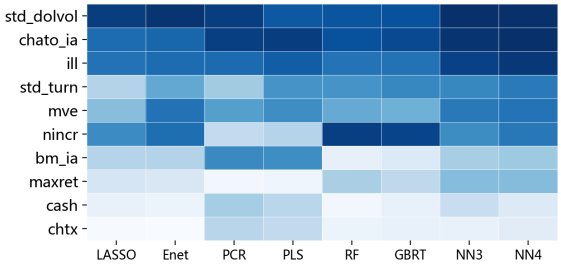

图 4-12　总体样本前十个股因子

图4-12则列出了在所有样本集合中所测算出的前十位个股因子, 即为反映了整个 A 股市场中综合影响力较强的十个个股因子, 可以看出, 所有样本集得出的有效因子与主板市场股票集中的有效因子种类大致相同, 这是因为主板市场的股票数量占据了所有 A 股市场股票的大部, 且以国有企业为主. 此外, 从图4-9至图4-12中可以明显的看出, 降维线性模型以及神经网络模型在四类不同集合的因子挖掘中, 均能识别出更多的对股票收益率贡献较大的因子, 所以 4.1 节中模型的样本外预测评价结果可知, 以 PLS、PCR 为代表的降维线性模型和以 NN3、NN4

为代表的神经网络模型在本文的实证资产定价的研究中表现出了良好的性能. 值得指出的是, 尽管树形模型的模型复杂度更高, 运算开销更大, 但是其在个股特征变量挖掘中的表现稍逊于降维线性模型, 出现这种的情况的原因之一可能是本文受限于树形算法中的计算复杂度, 将树的最大深度和叶节点最小样本数设置为能够节省计算成本的较优值, 这导致了回归树模型拟合不充分, 因此预测误差较大.

4.3　宏观预测变量对比

从前面的叙述中可以看出, 股票收益率的起伏是由个股因子、行业虚拟特征变量以及宏观经济因子的共同作用而产生, 行业虚拟变量的引入使得我们的预测过程更接近真实的市场行为, 其更多的是对应行业影响其对应的股票, 所以并不具有随机性. 在 4.2 节中我们着重讨论了影响股票收益率的个股因子, 在接下来的叙述中, 我们探究 13 个宏观经济因子在股票收益率的预测中是否有着不同的表现.

宏观经济因子在模型中是以克罗内克积的形式与个股因子相结合而被引入模型, 所以考察每个宏观经济因子重要性的方法和 4.2 节中考察个股因子重要性的方法相似, 即为逐个将宏观因子的值设置为 0 带入指定模型从而测量 R^2_{OOS} 的减少量, 并将减少量之和做归一化处理, 然后将每个宏观经济变量因子的重要性转化为百分比的形式, 这样做的目的是可以直观的看出每个宏观经济因子在股票收益率预测中的相对重要程度. 下图显示了宏观经济变量重要程度测量的箱线图, 其样本集合是所有 A 股样本股票集合. 我们在试验中将样本集合按照不同板块进行划分, 以测量不同板块间宏观经济因子的不同表现, 但是试验结果并未显示出不同板块市场间宏观经济变量有显著差异, 这说明宏观经济变量对于个股的影响仅仅是在股市的宏观指数层面.

图4-13是 13 个宏观经济因子在机器学习预测模型中的重要性占比箱线图, 图中显示了每个宏观经济因子的重要性在 11 种机器学习模型 (LASSO、Enet、PCR、PLS、RF、GBRT、NN1、NN2、NN3、NN4、NN5) 中的分布情况与总体情况, 上图中绿色三角形表示宏观经济因子在 11 种机器学习模型中的重要性百分比的平均值, 可以看出, 重要性百分比平均值大于 10% 的四个宏观经济因子分别是账面市值比 (m_bm)、股权扩张 (m_etis)、通货膨胀 (m_infl) 以及信用利差 (m_cds), 说明这四类宏观经济因子对于股票未来收益率的预测有着较大的贡献. 同时, 图中的红色加号表示宏观经济因子在 11 种机器学习模型中的重要性百分比异常值, 即为宏观经济因子在某种机器学习模型中的重要性占比较大或极小, 从图中我们可以看出, 除了通货膨胀因子 (m_infl) 的整体表现较好没有异常值以外, 以上其他三类宏观经济因子均存在异常值, 即为在某种机器学习模型中的表现出色. 对股

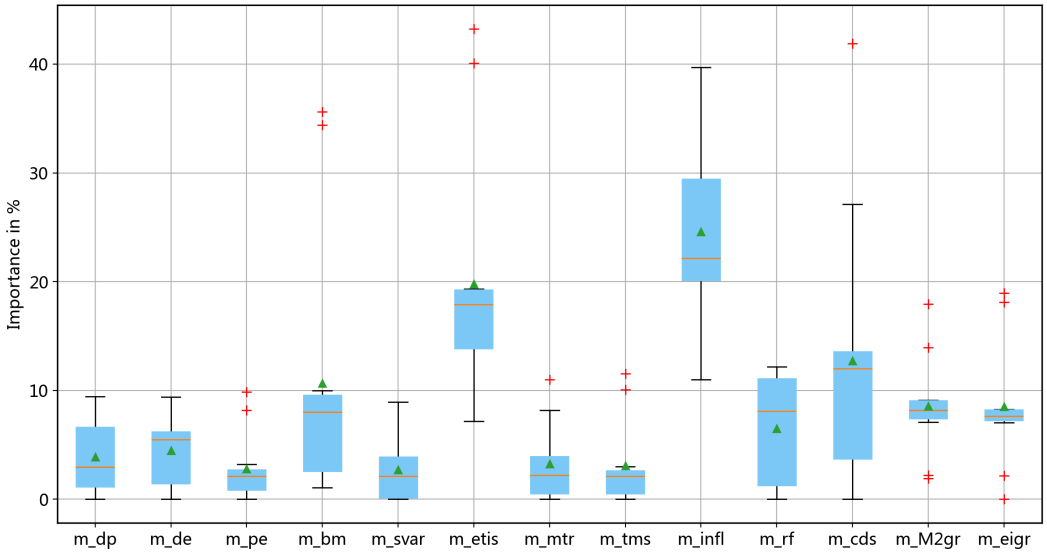

图 4-13　宏观经济变量重要性箱线图

票未来收益率预测整体贡献最大的通货膨胀 (m_infl) 因子, 不仅反映反应了市场的经济活动状况, 也反映着资本市场的整体情况, 所以通货膨胀因子在股票收益率的预测中起到重要的作用. 股权扩张 (m_etis) 因子作为另一个重要的宏观经济因子, 其在 PCR 模型中的重要性占比达到了 43.26%, 如 Leippold 等人[34] 在文中提到的那样, 国内 A 股市场的 IPO 机制导致了股权扩张因子在预测股票收益率时发挥着重要的作用. 本文中的信用利差 (m_cds) 因子的选取参考了 Gu 等人[2] 在文中所引用的违约利差 (default spread) 因子, 信用利差作为可以补偿投资者基础资产高于无风险收益的利差, 是总体经济情况的一种反应, 在市场中反应着投资者对经济前景的预期, 所以信用利差因子在股票收益率预测中的表现应该受到重视. 对于账面市值比因子, 箱线图中的极值出现在 LASSO 与 Enet 两个模型之中, 其重要性占比分别为 34.4% 与 35.6%, 说明这两个模型强烈的支持账面市值比 (m_bm) 因子.

除此之外, 我们发现其他反应市场总体情况的宏观经济因子, 如无风险利率 (m_rf)、M2 增长率 (m_M2gr) 以及进出口增长额 (m_eigr) 因子, 在收益率预测中的综合表现普遍高于反应 A 股股票整体情况的股票类宏观经济因子, 例如股息率 (m_dp)、派息率 (m_de)、市盈率 (m_pe)、股票方差 (m_svar) 等因子, 这些股票类宏观因子在股票收益率预测中被大多数机器学习模型所忽略.

第五章　结论与展望

随着影响实证资产价格的因素越来越多元化, 传统的实证资产定价理论在因子的选取与模型的选择方面都受到了挑战. 通过前面几个章节的叙述, 我们分析了传统实证资产定价模型在准确测量股票风险溢价时所面临的挑战, 结合国内外研究文献给出了运用机器学习方法来解决传统实证资产定价领域问题的途径, 随后我们以国内 A 股市场股票为研究对象, 实证的检验了机器学习模型在资本资产定价领域所发挥的优势.

我们以准确衡量股票风险溢价为目标, 通过将个股因子与宏观经济因子作为回归的自变量代入机器学习模型中, 来探究机器学习模型对股票未来收益率的可预测性, 验证了传统的线性模型在股票收益率预测的中表现不如非线性机器学习模型, 其中以 NN3、NN4 为代表的神经网络模型和以 RF 为代表的树形模型在股票收益率的预测中表现良好. 其次, 以 PCR、PLS 为代表的降维线性模型相较于传统的实证资产定价模型也有着不错的表现. 与此同时, 我们通过比较机器学习模型在不同股票集合中的预测结果, 证实了国内 A 股市场不同于国外股票市场而出现“异象”的原因是 A 股市场中存在的国有企业与大量的散户投资者, 指出国有企业的特殊性使得国有企业并非像我们想象的那样不具有可预测性, 反而国有企业的可预测性强于非国有企业. 在将股票集合按照板块市场划分后, 我们证实了中小板市场因其所包含的股票大都为小盘股, 所以可预测性略高于主板市场, 而创业板市场因为流动性较弱的原因使得可预测性相对较弱.

在对比了机器学习模型对市场的可预测性之后, 我们进一步运用这些机器学习模型挖掘出了对股票收益率影响较大的个股因子和宏观经济因子. 选出了不同股票集合中对股票收益率的综合影响排在前十的个股因子, 并通过热力图分别揭示了每种机器学习模型中的重要个股因子. 通过分析得出了流动性因子与动量因子在收益率预测中起到了重要的作用. 最后, 我们通过验证, 得出了宏观经济因子对于股票收益率的影响只在股票指数层面, 并且揭示了通货膨胀 (m_infl)、股权扩张 (m_etis) 以及信用利差 (m_cds) 这三类宏观经济因子对股票预期收益率有着较大的影响.

本文在得出上述结论的同时还有些不足之处. 本文的研究主题是机器学习模型在实证资产定价中的应用, 只将机器学习模型用于单一股票的风险溢价的衡量中来, 并未采用资本投资中常用的组合投资方式进行风险溢价的测算. 在实际的投资实践中, 投资者一般通过多空组合的方式来配置资产, 所以作为本文的后续研究方向, 会将机器学习模型引入多空投资组合的收益率测算中来.

此外由于本文所用的数据量较大, 导致计算开销较大, 所以并未引入更为复杂的机器学习算法和深度学习算法进行探究, 所选的几类机器学习模型相对来说是具有代表性的, 但是如果想要得到更好的预测效果以更准确的衡量股票的风险溢价, 可以引入其他新兴的机器学习算法. 不同的机器学习算法对不同市场的适应性也是不同的, 这也可以作为进一步发掘股票影响因素的研究方向.

参考文献

[1] GREEN J, HAND J R, ZHANG X F. The characteristics that provide independent information about average U.S. monthly stock returns[J]. The Review of Financial Studies, 2017, 30(12): 4389-4436.

[2] GU S, KELLY B, XIU D. Empirical asset pricing via machine learning[J]. The Review of Financial Studies, 2020, 33(5): 2223-2273.

[3] FAMA E F, FRENCH K R. The equity premium[J]. The Journal of Finance, 2002, 57(2): 637-659.

[4] BANSAL R, YARON A. Risks for the long run: A potential resolution of asset pricing puzzles[J]. The Journal of Finance, 2004, 59(4): 1481-1509.

[5] FAMA E F, MACBETH J D. Risk, return, and equilibrium: Empirical tests[J]. Journal of Political Economy, 1973, 81(3): 607-636.

[6] FAMA E F, FRENCH K R. Common risk factors in the returns on stocks and bonds[J]. Journal of Financial Economics, 1993, 33(1): 3-56.

[7] FAMA E F, FRENCH K R. A five-factor asset pricing model[J]. Journal of Financial Economics, 2015, 116(1): 1-22.

[8] GIGLIO S, XIU D. Asset pricing with omitted factors[J]. Journal of Political Economy, 2021, 129(7): 1947-1990.

[9] WELCH I, GOYAL A. A comprehensive look at the empirical performance of equity premium prediction[J]. The Review of Financial Studies, 2008, 21(4): 1455-1508.

[10] BAI J, NG S. Determining the number of factors in approximate factor models[J]. Econometrica, 2002, 70(1): 191-221.

[11] RAPACH D E, STRAUSS J K, ZHOU G. International stock return predictability: What is the role of the United States?[J]. The Journal of Finance, 2013, 68(4): 1633-1662.

[12] FAN J, LI Q, WANG Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017, 79(1): 247-265.

[13] CHEN J, TANG G, YAO J, et al. Investor attention and stock returns[J]. Journal of Financial and Quantitative Analysis, 2022, 57(2): 455-484.

[14] RAPACH D E, STRAUSS J K, ZHOU G. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy[J]. The Review of Financial Studies, 2010, 23 (2): 821-862.

[15] LEUNG R C, TAM Y M. Statistical Arbitrage Risk Premium by Machine Learning[J]. arXiv preprint arXiv:2103.09987, 2021.

[16] OZBAYOGLU A M, GUDELEK M U, SEZER O B. Deep learning for financial applications: A survey[J]. Applied Soft Computing, 2020, 93: 106384.

[17] KHANDANI A E, KIM A J, LO A W. Consumer credit-risk models via machine-learning algorithms[J]. Journal of Banking & Finance, 2010, 34(11): 2767-2787.

[18] MORITZ B, ZIMMERMANN T. Tree-based conditional portfolio sorts: The relation between past and future stock returns[J]. Social Science Electronic Publishing, 2016.

[19] HUTCHINSON J M, LO A W, POGGIO T. A nonparametric approach to pricing and hedging derivative securities via learning networks[J]. The Journal of Finance, 1994, 49(3): 851-889.

[20] YAO J, LI Y, TAN C L. Option price forecasting using neural networks[J]. Omega, 2000, 28 (4): 455-466.

[21] HEATON J, POLSON N, WITTE J. Deep learning for finance: Deep portfolios[J]. Applied Stochastic Models in Business and Industry, 2016, 33(1): 3-12.

[22] GU S, KELLY B, XIU D. Autoencoder asset pricing models[J]. Journal of Econometrics, 2021, 222(1): 429-450.

[23] CHEN L, PELGER M, ZHU J. Deep learning in asset pricing[J]. Management Science, 2023.

[24] FAMA E F, FRENCH K R. Value versus growth: The international evidence[J]. The Journal of Finance, 1998, 53(6): 1975-1999.

[25] ROUWENHORST K G. International momentum strategies[J]. The Journal of Finance, 1998, 53(1): 267-284.

[26] WANG F, XU Y. What determines Chinese stock returns?[J]. Financial Analysts Journal, 2004, 60(6): 65-77.

[27] DREW M E, NAUGHTON T, VEERARAGHAVAN M. Firm size, book-to-market equity and security returns: Evidence from the Shanghai Stock Exchange[J]. Australian Journal of Management, 2003, 28(2): 119.

[28] WANG C, CHIN S. Profitability of return and volume-based investment strategies in China's stock market[J]. Pacific-Basin Finance Journal, 2004, 12(5): 541-564.

[29] CHEN X, KIM K A, YAO T, et al. On the predictability of Chinese stock returns[J]. Pacific-Basin Finance Journal, 2010, 18(4): 403-425.

[30] CAKICI N, CHAN K, TOPYAN K. Cross-sectional stock return predictability in China[J]. The European Journal of Finance, 2017, 23(7-9): 581-605.

[31] PAN L, TANG Y, XU J. Speculative trading and stock returns[J]. Review of Finance, 2016, 20(5): 1835-1865.

[32] LIU J, STAMBAUGH R F, YUAN Y. Size and value in China[J]. Journal of Financial Economics, 2019, 134(1): 48-69.

[33] LIU Y, ZHOU G, ZHU Y. Trend factor in china: The role of large individual trading[J]. Social Science Electronic Publishing, 2021.

[34] LEIPPOLD M, WANG Q, ZHOU W. Machine learning in the chinese stock market[J]. Journal of Financial Economics, 2022, 145(2): 64-82.

[35] CAO Q, PARRY M E, LEGGIO K B. The three-factor model and artificial neural networks: predicting stock price movement in China[J]. Annals of Operations Research, 2011, 185(1): 25-44.

[36] WANG J J, WANG J Z, ZHANG Z G, et al. Stock index forecasting based on a hybrid model[J]. Omega, 2012, 40(6): 758-766.

[37] CHEN K, ZHOU Y, DAI F. A LSTM-based method for stock returns prediction: A case study of China stock market[C]//2015 IEEE International Conference on Big Data (Big Data). [S.l.]: IEEE Computer Society, 2015: 2823-2824.

[38] ZHANG X, HU Y, XIE K, et al. A causal feature selection algorithm for stock prediction modeling[J]. Neurocomputing, 2014, 142: 48-59.

[39] YUAN X, YUAN J, JIANG T, et al. Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market[J]. IEEE Access, 2020, 8: 22672-22685.

[40] HUBER P J. Robust estimation of a location parameter[J]. Breakthroughs in statistics: Methodology and distribution, 1992: 492-518.

[41] CARHART M M. On persistence in mutual fund performance[J]. The Journal of Finance, 1997, 52(1): 57-82.

[42] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

[43] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[44] 李斌, 邵新月, 李玥阳. 机器学习驱动的基本面量化投资研究 [J/OL]. 中国工业经济, 2019 (08): 61-79 https://kns.cnki.net/kcms/detail/11.3536.f.20190820.1041.008.html.

[45] KELLY B T, PRUITT S, SU Y. Characteristics are covariances: A unified model of risk and return[J]. Journal of Financial Economics, 2019, 134(3): 501-524.

[46] BAKER M, STEIN J C. Market liquidity as a sentiment indicator[J]. Journal of Financial Markets, 2004, 7(3): 271-299.

[47] CHEN S S. Predicting the bear stock market: Macroeconomic variables as leading indicators[J]. Journal of Banking & Finance, 2009, 33(2): 211-223.