基于北向资金流动的量化投资策略研究

——使用 XGBoost 机器学习算法

计金（双）210　王涵　21013000

摘要：本研究旨在探索基于北向资金流动的量化投资策略，采用 XGBoost 机器学习算法构建多因子量化选股模型。首先，我们分析了北向资金流动与沪深 300 指数涨跌的相关性，通过单位根 ADF 检验、格兰杰因果检验和 Pearson 相关性检验验证了北向资金流入与沪深 300ETF 涨跌之间的显著正相关关系。接着，基于北向资金流入构建了指数量化策略，并通过策略模拟验证了其在 2018 年至 2023 年期间的超额收益表现。随后，采用 XGBoost 算法进行多因子选股，通过对 2017 年至 2023 年北向资金投资的股票数据进行滚动训练和预测，构建了收益率高于市场指数的投资组合。研究结果显示，所构建的量化选股策略在多数年份中均实现了超额收益，且具有较高的经济效益。通过夏普比率和 R^2 分数的检验，进一步验证了模型的稳定性和有效性。本研究为境内投资者和机构提供了新的量化投资思路和方法，具有较高的实际应用价值。

第一部分 绪论

1.1 策略研究背景及介绍

1.1.1 北向资金相关介绍

在中国股市中，一般“北”指的是沪深两市的股票，“南”指的是指香港股票。其中北向资金是指通过沪港通和深港通机制，从香港市场流入中国内地股市的资金，主要由外资组成。由于 A 股市场存在管制，不允许外资直接参与，我国分别在 2014 年和 2016 年开通了沪港通和深港通，形成 A 股市场和港股市场中的资金互流，因为内地在香港的北面，就形成了北向资金的说法。近 10 年来，自沪深港通开通以来，北向资金净流入量显著增长，在 A 股市场的成交占比逐年上升，对市场走势和风格的影响也在加大，这也使得北向资金的流入和流出情况成为 A 股市场的风向标。根据 Wind 数据，截至 2024 年，北向资金净流入额已达到 18473.88 亿元[1][2]。

由于“北向资金”大都是国外机构投资，而机构一般都有较为雄厚的资金实力、专门的公司研究团与信息搜集团队，在投资上相比普通的散户更专业，也更有优势的多，故北向资金的变化受到了国内广大投资者和媒体的密切关注。因此北上资金也有“聪明的资金”之称，对于普通价值投资者而言，关注北上资金买

入与卖出以及关注北上资金长期买入的股票，具有一定参考意义，特别是在股市大幅波动时，北向资金的净流入和净流出更是受到了投资者的高度关注[3]。

1.1.2 因子投资理论与机器学习

因子投资理论起源于 20 世纪 60 年代，随着资本资产定价模型（CAPM）的提出，学界开始探索风险与收益之间的关系。然而，CAPM 模型未能完全解释市场的实际表现，这促使研究者寻找其他能够解释股票收益的因子；Fama 和 French 提出了三因子模型，将市场风险、公司规模（SMB）和账面市值比（HML）作为影响股票收益的主要因子；随后，Carhart 在三因子模型的基础上增加了动量因子，形成了四因子模型[4][5]。这些因子的提出为投资者提供了一种新的视角，即通过识别和利用这些因子来构建投资组合，以期获得超额收益。因子投资策略的有效性在很大程度上依赖于资产定价模型的准确性。如果一个因子能够持续地解释股票收益的变动，那么它就可以被用作资产定价模型中的一个风险因子。因此，因子投资策略的实施需要对资产定价模型有深入的理解。

近年来，随着机器学习和大数据分析技术的发展，因子投资策略也在不断地进化，金融领域开始探索如何利用这些技术来提高投资决策的效率和准确性。研究者开始探索非传统的因子，如情绪因子、社交媒体情绪等，这些因子可能对股票收益有预测作用[6][7]。机器学习在因子投资和资产定价中的应用，为金融研究和实践带来了新的机遇和挑战。

在多因子选股中，机器学习模型通常需要大量的训练数据。这些数据可以是时间序列数据，也可以是截面数据，具体使用哪种数据取决于选股策略的具体需求。时间序列训练关注于股票价格或因子随时间的变化趋势，而截面数据训练则关注于某一特定时间点所有股票的截面信息。在《The Journal of Portfolio Management Quantitative Special Issue 2024》发表的论文《Equity Factor Timing: A Two-Stage Machine Learning Approach》中，作者提出了一种两阶段机器学习方法来增强因子预测结果。第一阶段是市场风险体制的识别和预测，第二阶段是因子相对表现的预测。该方法通过整合宏观经济和市场风险因素，提高了模型的预测能力；而在另一篇文献中，学者探讨了使用深度学习模型进行多因子择时的方法，研究选取了 6 个常用风格因子，并利用多任务学习结构降低过拟合风险[8]。这些先例表明，机器学习方法在因子投资领域的应用是多样化的，包括从数据预处理到模型构建和评估的整个流程。通过这些方法，投资者可以更有效地识别和利用影响股票收益的关键因子，从而提高投资决策的质量[9]。

1.2 策略研究意义

策略研究意义从理论和现实两个方面展开阐述

（1）理论方面，在实证检验中首先分析北向资金流动和指数波动率现状，研究北向资金流动与沪深 300 指数涨跌相关关系，丰富指数量化投资策略；之后在实证检验中使用 XGBoost 机器学习算法构建基于北向资金情绪因子的多因子量化选股模型，丰富量化选股理论模型和策略，并对基于北向资金的量化选股模型进行实例验证，丰富量化选股实证研究。

（2）现实方面，基于北向资金流动的指数量化策略和 XGBoost 量化选股策略研究具有较强的实用性，为境内投资者和机构在追求超额收益及控制风险时提供新的思路和方法，进一步挖掘实际选股过程中市场特征因子和资本市场收益的相关关系。

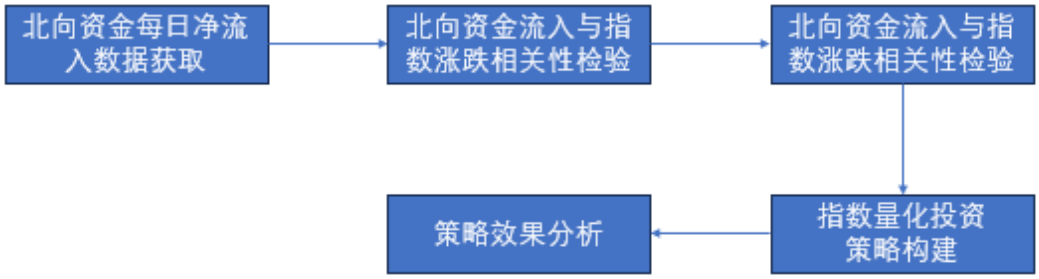
第二部分 策略研究内容

本研究的主要思路是从北向资金流动与内地股票市场的相关性入手分析，从理论层面分析将北向资金流动性作为量化投资依据的可行性，而后设计基于北向资金的指数量化策略和多因子模型选股策略，并通过实证回测分析验证模型的科学性和合理性[10]。

实证检验内容共分为两部分：

（1）基于北向资金流动的指数量化投资策略（第三部分）

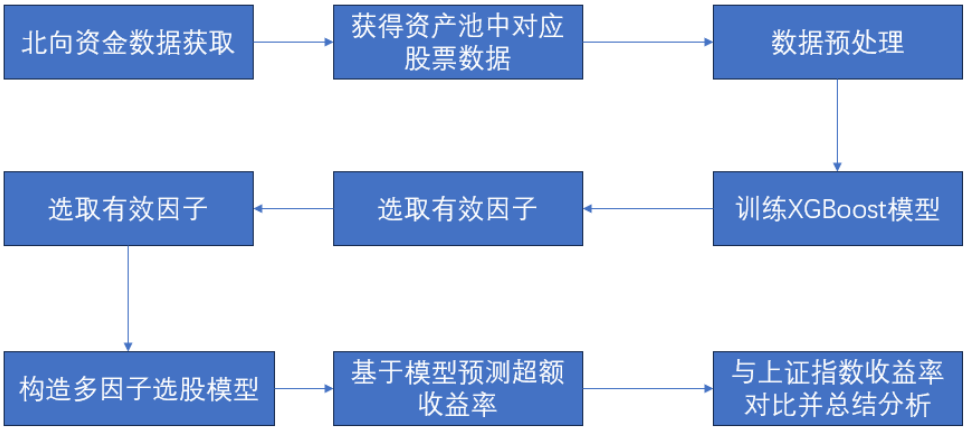
该部分实证研究从北向资金的流动入手，重点分析北向资金流动与沪深 300ET 涨跌的关系，从中获得北向资金判断市场涨跌的信息，依据此关系来进行买卖操作，进而构建指数量化策略。最终将该量化策略的累积收益率与沪深 300ETF 的累积收益率进行对比。



（2）基于北向资金流动使用 XGBoost 构建多因子量化投资策略（第四部分）

该部分研究通过分析比较北向资金所具备的财务、经济、杠杆等特征，构造一系列北向资金研究因子，实证部分使用能够处理非线性问题的 XGBoost 机器学习算法构件多因子选股模型以模拟实现量化交易。其中资产池采用 2017 年至 2023 年北向资金所投资的资产股票，基于多因子选股模型得到的结果构造投资

组合，最终预测得到的超额收益率与上证指数进行对比。该部分研究具体完成过程如下流程图所示：



第三部分 基于北向资金流动的指数量化投资策略

3.1 北向资金净流入流出对指数涨跌的影响分析

3.1.1 数据样本的确定

本实验中通过 Wind 数据库获取到北向资金每日净流入与每日沪深 300ETF 涨跌数据。由于沪港通和深港通开通的时间不同，在实验中我们统一选取在深港通开通后的数据作为实验样本，即数据时间跨度为 2016 年 12 月 5 日至 2024 年 5 月 6 日。

3.1.2 北向资金流入与沪深 300ETF 涨跌的相关性检验

（1）使用单位根 ADF 检验对数据进行平稳性检验

研究中使用的北向资金和沪深 300ETF 数据都为常见的时间序列数据。时间序列数据的波动趋势是我们研究其特征的关键，只有当时间序列数据的特征维持稳定（如均值与方差），那么可以认为在之后的一段时间里，其数据的分布情况与历史数据大概率是保持一致的，基于此我们才可以根据历史数据对未来做一个预测。本文就是否可以依据北向资金来获取超额收益（相比于沪深 300ETF）。所以在进行分析前，首先要确定数据是否能够保证平稳[11]。

本节采用单位根 ADF 检验进行数据的平稳性检验，ADF 检验的原假设$H_0:\sigma=0$，原序列存在单位根，时间序列为非平稳序列。备择假设$H_1:\sigma<0$，原序列不存在单位根，为平稳序列。

对北向资金流入和沪深 300ETF 两个时间序列数据进行 ADF 检验，得到的结果如表 3-1 所示。

表 3-1 ADF 平稳性检验结果

	北向资金流入数据
T 检验值	-13.0528
p_value	$2\times e^{-21}$
1%临界值	-3.4341
5%临界值	-2.8632
10%临界值	-2.5676
	沪深 300 涨跌数据
T 检验值	-14.1825
p_value	0.0000
1%临界值	-3.4341
5%临界值	-2.8632
10%临界值	-2.5676

从上表数据可以看出，北向资金流入数据和沪深 300ETF 数据的 T 检验值分别为-13.0528 和-14.1825，远小于对于的 1%的临界值，p 值都非常接近于 0，均小于 1%，即拒接原假设。所以可以认为北向资金流入和沪深 300ETF 这两组数据是平稳的。两组数据的平稳性结论为后续检验两者的相关性提供了依据。

（2）格兰杰因果检验

格兰杰因果关系检验法是 Granger 于 2003 年提出的一种分析变量之间因果关系的方法。进行格兰杰因果检验的前提是时间序列必须是具有平稳性，上一部分内容已利用 ADF 检验对数据平稳性进行了验证，这里也在此基础上进行分析[12]。

接下来我们对北向资金流入数据和沪深 300ETF 数据进行 Granger 因果检验，验证北向资金流入是否可以预测出沪深 300ETF 的涨跌，得到的结果如表 3-2 所示。

表 3-2 格兰杰因果检验结果

Granger Causality	
Number of lags（no zero）	1
ssr based F test	F=10.4792,p=0.0012,df_deom=1713

ssr based chi2 test	chi2=10.4975,p=0.0012,df=1
likeihood ratio test	chi2=10.4656,p=0.0012,df=1
Parameter F test	F=10.4792,p=0.0012,df_num=1

Granger Causality Number of lags（no zero）	2
ssr based F test	F=6.1939,p=0.0021,df_deom=1710
ssr based chi2 test	chi2=12.4241,p=0.0020,df=2

根据表 3-2 可以看出，p 值小于 0.05，则可以认为北向资金流入对沪深 300ETF 涨跌有影响，两者存在格兰杰因果关系。从近几年数据来看，北向资金的成交量占市场总交易量的比重越来越大，且北向资金的成交量还在不断的上升，所以从北向资金流动来研究沪深 300ETF 和个股有很强的参考性，对获取超额收益有较大的帮助。

（3）Pearson 相关性检验

实验中还使用 Pearson 相关系数对两组数据进行相关性分析。Pearson 相关系数，是由 Karl Pearson 提出用来度量两个变量 X 和 Y 之间的相互关系（线性相关）的指标，被用来描述两个变量线性相关性的强弱。Pearson 相关系数通常用 r 或ρ是表示，其取值范围在-1 与+1 之间，若计算得到的 Pearson 绝对值越大，即越接近于 1 或-1，则相关性越强，若越接近于 0，则相关性越弱[13]。

假设有两个变量 X、Y，那么两变量间的 Pearson 相关系数可以通过以下公式进行计算：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(X)}{\sqrt{E(X^2)-E(Y^2)}\sqrt{E(Y^2)-E(X^2)}} \quad (3-1)$$

我们可以通过$\rho_{X,Y}$的值所在的取值范围来判断两者的相关性强度：取值介于 0.8 和 1.0 之间，可以认为两者之间具有极强相关性；取值介于 0.6 和 0.8 之间，两者之间具有强相关性；取值介于 0.4 与 0.6 之间，两者之间具有中等程度的相关性；取值介于 0.2 与 0.4 之间，两者之间具有弱相关性；取值介于 0 与 0.2 之间，两者之间具有极弱相关或无相关性。通过皮尔逊方法测量出变量间的相关性大小后，还需要进行显著性检验，以确定样本数据计算的相关系数是否能推论总体。原假设$H_0:\rho=0$，备择假设$H_1:\rho\neq 0$。[14]

利用已有的北向资金流入与沪深 300ETF 涨跌数据求其 Pearson 相关系数并进行显著性检验，结果如表 3-3 所示。

表 3-3 Person 相关系数检验结果

	北向资金净流入	沪深 300 指数
北向资金净流入	1.0000	0.5440
沪深 300 指数	0.5440	1.0000
P 值	$7.06\times e^{-133}$	

表 3-4　隔一天 Pearson 检验结果

	北向资金净流入	沪深 300 指数
北向资金净流入	1.0000	0.0692
沪深 300 指数	0.0692	1.0000
P 值	0.0041	

从表 3-3 中可以看出，在时间对应时，北向资金与沪深 300ETF 的 Pearson 检验结果为两者相关系数为 0.5440，P 值非常接近于 0。其相关系数介于 0.4-0.6 之间，说明两者之间具有较强的相关性，Pearson 相关性成立。

从表 3-4 隔一天的 Pearson 检验结果来看，其相关系数为 0.0692，p 值为 0.0041。随着时间间隔的增大，两者的相关性也越来越小，在间隔大于两天时，其显著性检验已不成立。所以我们可以认为在时间对齐的情况下，北向资金流入与沪深 300ETF 涨跌存在正相关关系。

3.2 基于北向资金流入的指数量化投资策略的构建

根据以上部分的检验分析，北向资金流入与沪深 300ETF 涨跌存在着正相关关系，随着流入量的增大，北向资金对沪深 300ETF 的涨跌越可预见。所以就产生了如下策略思路，根据北向资金的大额流入选取合适的阈值来对沪深 300ETF 进行择时操作，在不考虑手续费和滑点的情况下产生如下策略。

首先获取从起始日（北向资金数据开始日期）到终止日(北向资金数据结束日期)这一段时间内的所有北向资金净买入值；由于北向资金流入高预示着沪深 300ETF 涨，反之为跌，所以接下来对每日的北向资金净买入值的大小进行高低判断，具体判断方法为将当日之前所有北向资金净买入数据进行三等分，并从大到小进行排序，取第 1/3 位置的净买入值作为高点阈值，取第 2/3 位置的净买入值作为低点阈值；随后根据流入值的在哪个阈值范围内进行相应操作。具体操作方法为：若昨日的净流入值大于高点阈值，且目前并未有沪深 300ETF 持仓，则在今日开盘买入沪深 300ETF（若已有沪深 300ETF 持仓，则不做买入操作，继续持仓）。若昨日的净流入值小于低点阈值，且目前有沪深 300ETF 持仓，则在今日开盘卖出沪深 300ETF（若无沪深 300ETF 持仓，则今日不做交易操作）。若昨日的净流入值处于高点阈值和低点阈值之间，则今日不做交易操作；每日重复以上操作得到北向资金大额流入对沪深 300ETF 指数量化策略[15][16]。

在策略构建时需要注意的是，沪港通是 2014 年 11 月开通的，深港通是 2016 年 12 开始才有数据记录，所以该策略的起始时间为 2016 年 12 月，将沪港通和深港通两者数据之和作为北向资金数据。并且由于北向资金数据并不是在每个交易日都要，所有策略中只有在有北向资金数据的交易日才进行调仓操作。按照以上思路我们使用 Python 模拟策略得到结果。

3.3 策略效果分析

图 3-1 中展示了策略的收益情况，基于北向资金流入的指数量化投资策略在 2016 年至 2018 年期间的累计收益率与沪深 300ETF 的收益率基本持平。从 2018 年开始，该策略较普通 ETF 的累计收益率都取得了较大的超额收益。总而言之，该策略取得了较好的效果。

图 3-1 策略模拟结果

第四部分

基于北向资金流动使用 XGBoost 构建多因子投资策略

前文通过对北向资金的流入量分析，构建了指数量化策略。在第四部分内容中，我们考虑将选取出的多因子与北向资金结合，从北向资金中获取哪些股票为“优质股”，哪些股票为“劣质股”，通过持有“优质股”赚取超额收益，构建基于北向资金流动的量化选股策略。

4.1 策略因子的选择

多因子量化选股是一种基于统计和数学模型的投资策略，它通过综合考虑多个影响股票收益的因子，来预测股票未来的市场表现并进行选股。这种策略的核心在于识别和利用那些能够持续产生超额收益的因子，并通过科学的组合和风险管理，实现在不同市场环境下的稳健投资回报[17]。

本文在对之前的学术研究进行总结的基础上，选择并构造了诸多因子，共计 27 个因子作为因子池，并且在后续对因子的有效性检验中，排除对预测收益率无法起作用的因子。因子池如表 4-1 所示。

表 4-1　因子候选池

因子代码	因子名称	因子代码	因子名称
MonTrdTurnR	流通股月换手率	OpPrfPS	每股营业利润
AvgDtrdTurnR	流通股平均日换手率	NAPS	每股净资产
Monret	月收益率	IncomePS	每股营业收入
Monrfret	月无风险收益率	NCFfropePS	每股经营活动现金流量净额
Stamptaxa	A 股印花税	Rmrf	市场溢酬因子
PE	市盈率	Smb	市值因子
PB	市净率	Hml	账面市值比因子
PCF	市现率	Alpha	Alpha
PS	市销率	Beta	风险因子
EPS	每股收益	R2	R 方

ROE	净资产收益率	R2Adj	调整 R 方
AccumFundPS	每股公积金	roll_12_Turn	个股最近 12 个月内日均换手率
Avg_1year_turn		个股最近 12 个月内日均换手率除以最近 1 年内日均换手率	
Avg_2year_turn		个股最近 12 个月内日均换手率除以最近 2 年内日均换手率	
wgt_return_12m		个股最近 12 个月内用每日换手率乘以每日收益率求算术平均值	

4.2 数据预处理

4.2.1 缺失值处理——中位数填充法

在本部分实验中多采用股票多因子的截面数据，由于部分值存在停牌等原因，可能会影响多个因子数据完整性。考虑到包含众多行业，存在某些指标缺失的可能性较大，由于数据不完整或其他原因可能会导致某些因子出现缺失，因此在模型输入之前需要检查是否有缺失值。如果存在少量缺失值，则进行缺失值填充，如果单支股票缺失值较多，则将其剔除样本，以避免模型效果不准确或直接报错。研究中从数据选取上避免了存在较多缺失值的情况，选取区间内线性替换，如果空缺数据连续过多，到可能影响算法的准确性，则考虑更换因子或者数据源。

对于存在的缺失值，实验中采取中位数填充法进行处理。中位数填充法的优点是简单易行，且由于中位数对异常值不敏感，因此在数据分布不对称或者包含异常值时，比均值填充法更稳健。

4.2.2 标准化处理——Z-score

Z-score（标准分数）是一种衡量数据点与数据集平均值（均值）之间差异的统计量。它表示数据点与均值的距离，用标准差的单位来衡量。Z-score 的计算公式如式（4-1）所示：

$$Z = \frac{(X - \mu)}{\sigma} \qquad (4-1)$$

其中X是数据集中的某个具体数据点；μ是数据集的均值；σ是数据集的标准差。

Z-score 的值可以告诉我们数据点相对于数据集的相对位置。它的意义在于，它可以将原始分数标准化，即将其转换为一种相对于平均值的相对位置，这样就可以在不同的数据分布之间进行比较：如果 Z-score 为 0，表示该数据点恰好等

于均值；如果 Z-score 为正数，表示该数据点高于均值；如果 Z-score 为负数，表示该数据点低于均值。

4.2.3 因子有效性分析——斯皮尔曼相关系数

斯皮尔曼相关系数，是一种非参数的统计度量，用于评估两个变量之间的相关性。与皮尔逊相关系数不同，斯皮尔曼相关系数是基于变量值的秩次而不是实际值来计算的。其计算公式如式（4-2）所示。

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (4-2)$$

其中r_s是斯皮尔曼相关系数，d_i是排名之间的差异，n是数据对的数量。

斯皮尔曼相关系数不要求数据遵循特定的分布，这使得它适用于不符合正态分布的数据集。它基于数据的秩次而不是实际值，因此对异常值和数据中的非典型观测点具有较好的鲁棒性。但是斯皮尔曼相关系数只能告诉我们变量之间是否存在相关性，以及相关性的强度和方向，但不能告诉我们变量之间关系的确切形式。

4.2.4 因子有效性检验结果

在因子有效性分析中，发现市场溢酬因子 Rmrf、市值因子 Smb、账面市值比因子 Hml、Alpha、风险因子 Beta、R^2和调整R^2的斯皮尔曼相关系数趋近于零，故排除选择这些因子。

4.3 XGBoost 机器学习算法模型

在机器学习领域，提升算法（Boosting）因其在分类和回归问题上的强大性能而受到广泛关注。XGBoost，即 eXtreme Gradient Boosting，是一种高效的梯度提升框架，自 2016 年由陈天奇等人提出以来，已成为业界和学术界广泛使用的算法之一。XGBoost 算法是基于梯度提升决策树（GBDT）的改进版本，它通过引入正则化项来控制模型的复杂度，从而有效避免过拟合问题。算法的核心在于构建一个加权的弱学习器集合，通过迭代的方式逐步优化模型，以最小化预定的损失函数[18]。

4.3.1 XGBoost 核心算法

XGBoost 的核心算法是梯度提升决策树（Gradient Boosting Decision Tree, GBDT）。这是一种集成学习算法，它通过逐步添加弱学习器（通常是决策树）来最小化一个预定的损失函数容。核心算法具体描述如下所示。

（1） 不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数 f(x)，去拟合上次预测的残差。

（2） 当训练完成得到 k 棵树，我们要预测样本的分数，即根据样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数。

（3） 最后将每棵树对应的分数加起来就是该样本的预测值。

为保证预测值和真实值的匹配，XGBoost 将多棵树的得分累加得到最终的预测得分即每一次迭代都在现有树的基础上增加一棵树去拟合前面树的预测结果与真实值之间的残差。XGBoost 模型每轮预测结果如式（4-3）所示。

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ &\cdots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \tag{4-3}$$

4.3.2 XGBoost 模型参数调整

XGBoost 模型中能够实现很好的分类预测效果，但是其前提需要将模型中的参数进行不断的调整优化，才能获得较好的效果，这个过程也称之为超参数优化，常用的参数调优方法有网格搜索、贝叶斯优化、随机搜索等。XGBoost 模型中较为关键的参数如下表 4-2 所示。

表 4-2　XGBoost 模型关键参数

参数名称	参数解释
max_depth	树的最大深度
min_child_weight	子节点的最小权重
gamma	分割节点所需的最小损失
subsample	训练时使用的样本比例
colsample_bytree	每个树中使用的列的比例
learning_rate	学习率
n_estimators	最大迭代次数

XGBoost 模型中有一个很有用的函数“CV”，这个函数可以在每一次迭代中使用交叉验证并返回决策树数量。交叉验证就是重复的使用数据，把得到的样本数据进行切分，组合为不同的训练集和测试集，用训练集来训练模型，用测试集来评估模型预测的好坏。在此基础上可以得到多组不同的训练集和测试集，某次训练集中的某样本在下次可能成为测试集中的样本，即所谓“交叉”。通过评估模型在不同数据子集上的性能，从而选择出最佳的参数组合，同时也可以与网格搜索或随机搜索等参数调优方法结合使用。

本文基于 2017 年至 2023 年的数据集进行交叉验证，最终确定的 XGBoost 模型参数如表 4-3 所示。

表 4-3　XGBoost 模型参数表

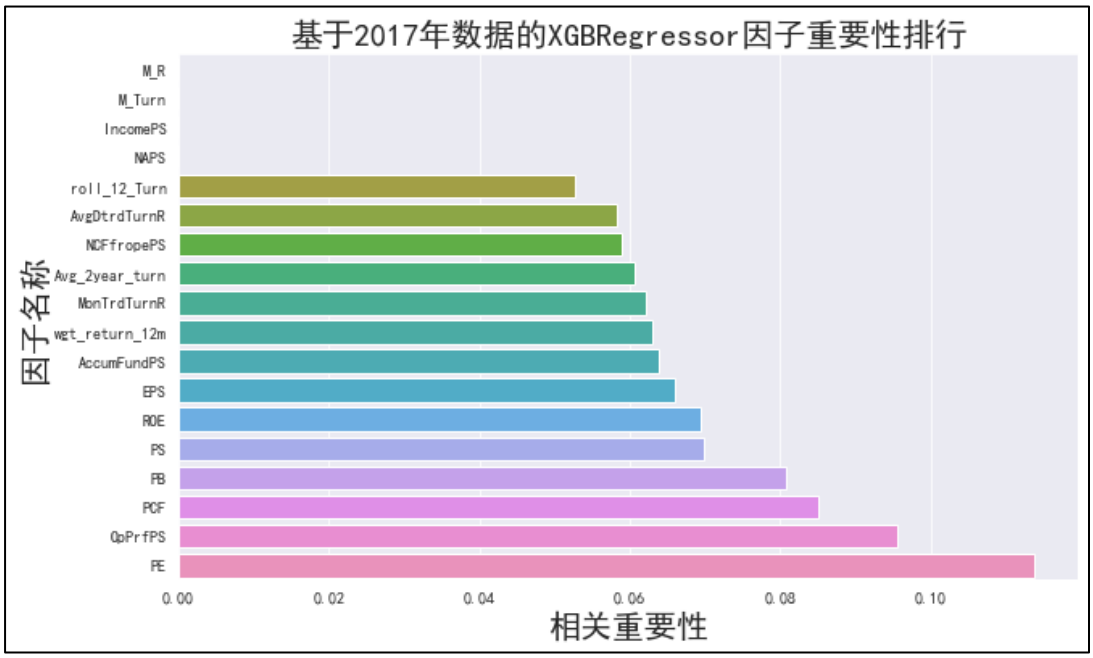
参数名称	参数取值
Learning_rate	0.0999
Max_depth	6
N_estimators	90
Subsample	0.5969
Colsample_bytree	1.0
Gamma	0.1
Min_child_weight	3

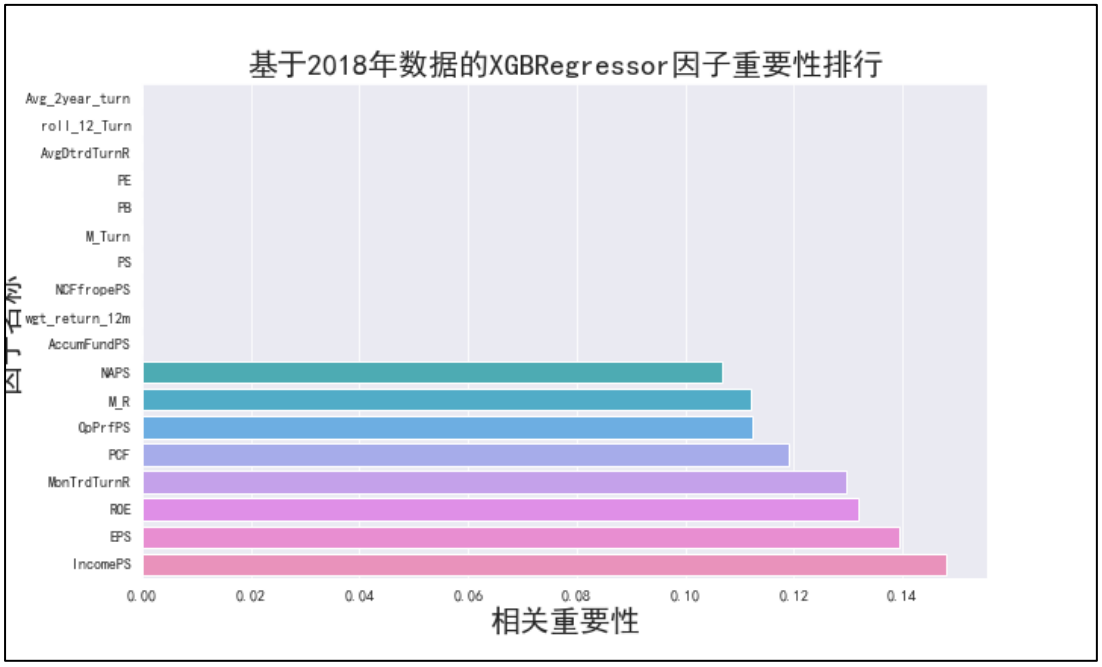
4.4 使用 XGBoost 模型进行拟合预测

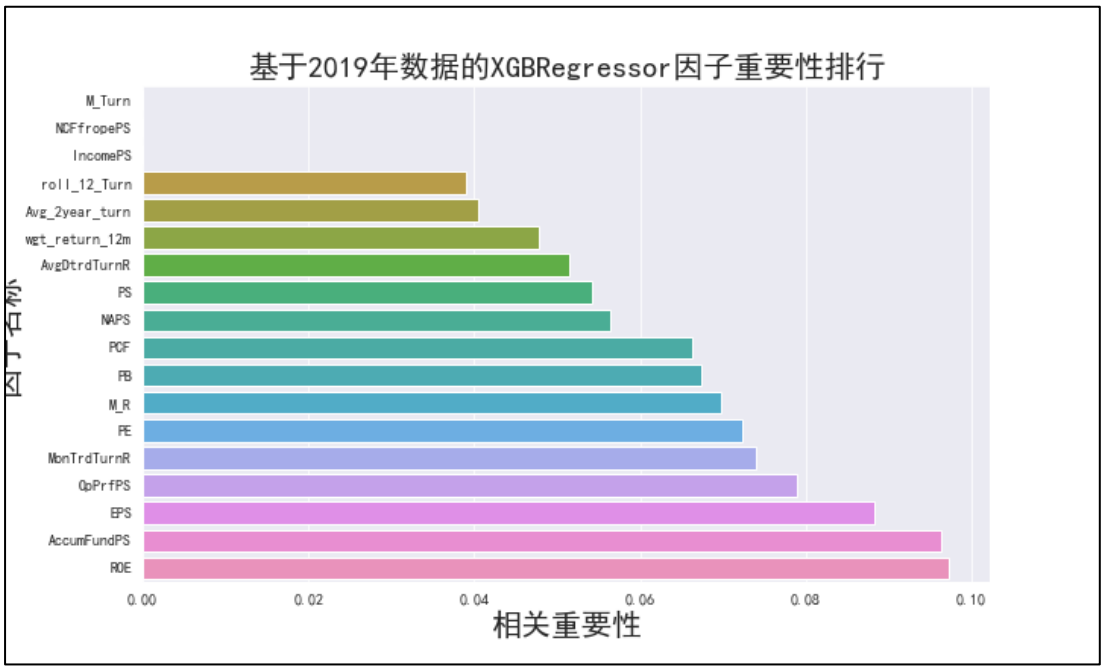
本文基于 2017 年至 2023 年这 6 年间的数据对 XGBoost 模型进行滚动训练，最终得到的拟合和预测结果如下所示。

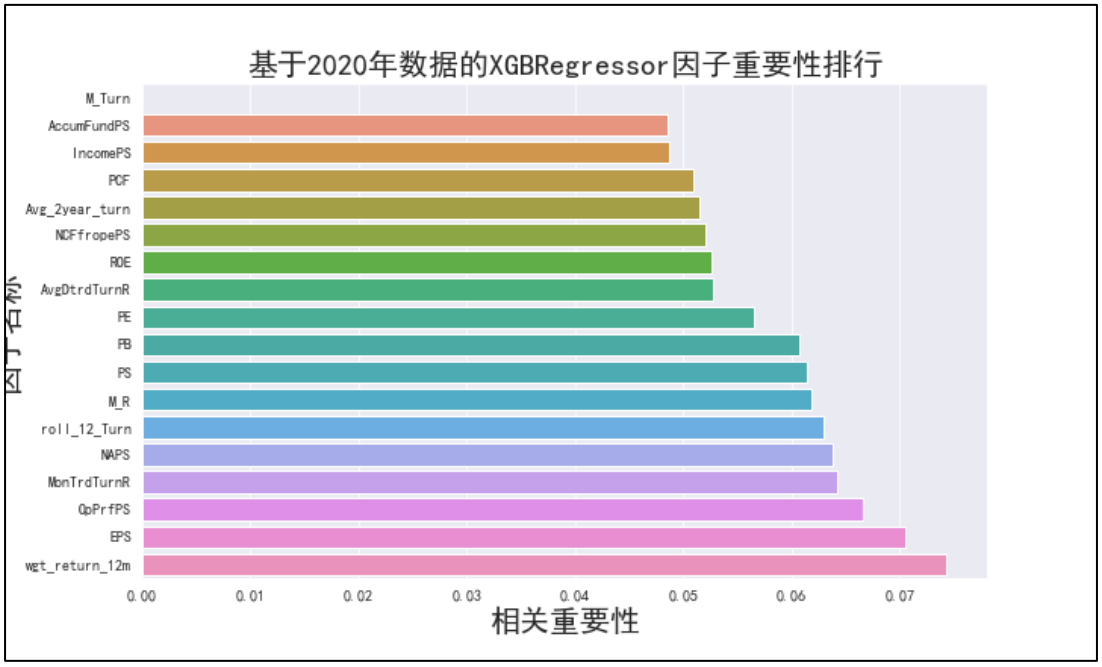
4.4.2 使用 XGBoost 模型进行因子重要性排序

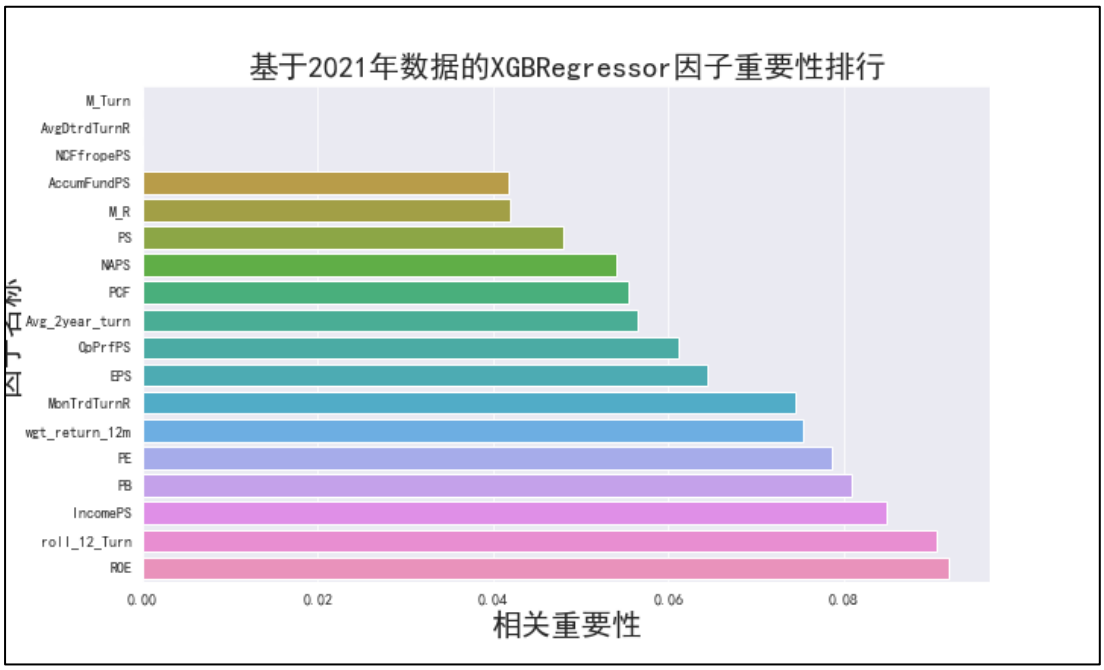
XGBoost 模型在每轮训练过程中，会将各项因子的权重进行计算以了解每个因子对结果的解释作用。实验中通过 get_score 函数对训练期内各指标的重要性数值进行获取，并进行排序，最终得到了 2017 年至 2023 年每年的各指标的重要性占比。重要性占比如图 4-1 所示。

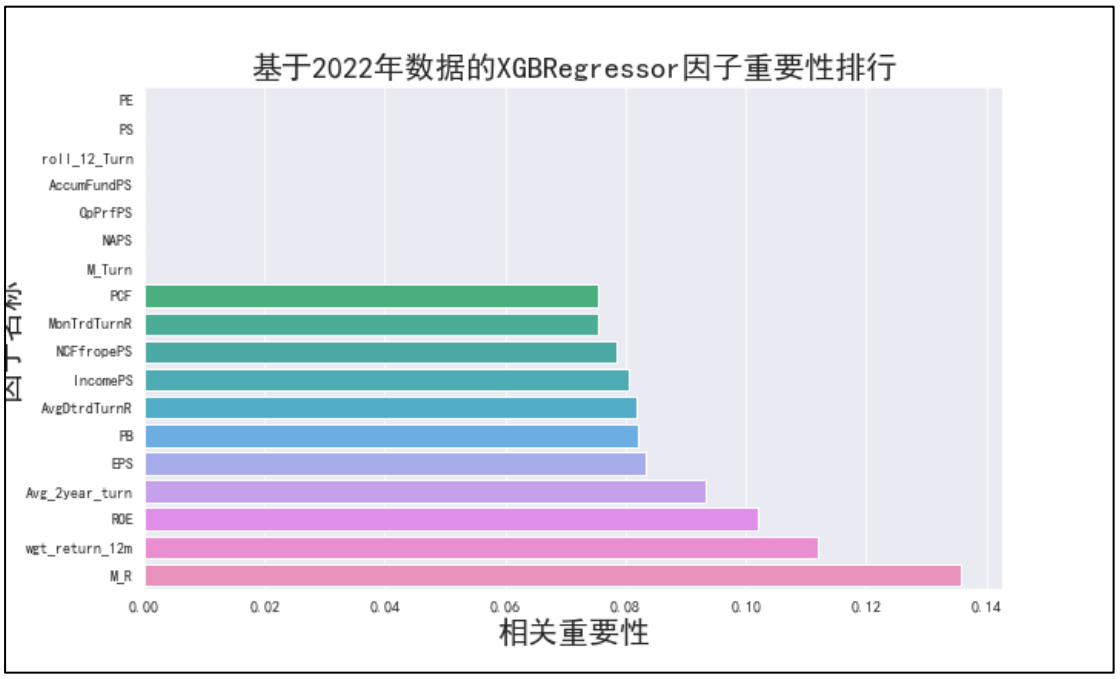












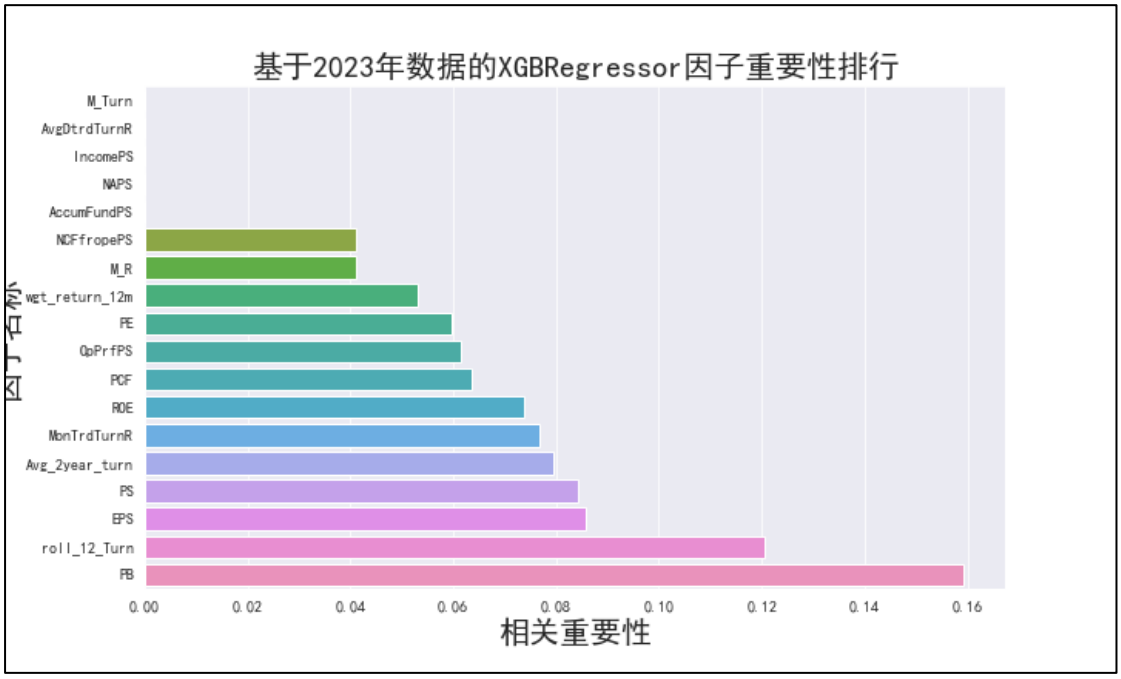


图 4-1 2017 年至 2023 年每年的各指标的重要性占比

从图中可以看到 PE、OpPrfPs、EPS 和 PB 指标在每年的特征重要性图中排名均靠前，反映出实验中选取的指标具有较高的稳定性。

4.4.3 基于训练后的模型预测超额收益率

经过 2017 年至 2023 年数据的训练后，可以通过输入相应的因子变量对超额收益率进行预测。研究中使用训练后的模型，对每月北向资金投资的股票进行

收益率预测，依据模型每月预测的超额收益率序列进行排序，从中选取超额收益率最高的前十只股票，采用等权重方法构造投资组合。同时本文为对比所预测的超额收益率是否有着较好的经济效益，选用沪深 300 作为基准指数与预测的超额收益率进行对比。具体对比图如图 4-2 所示。

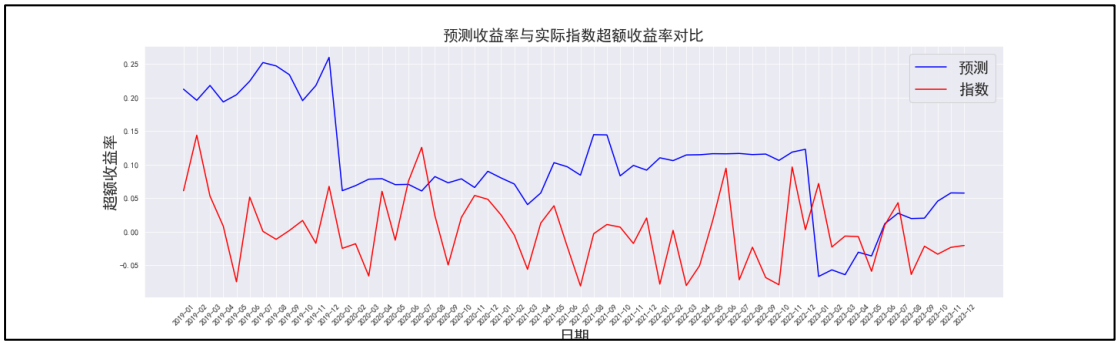


图 4-2 预测收益率与实际指数超额收益率对比图

从上图中可以看到预测的超额收益率的波动形状与沪深 300 指数变动大致相同，并且具有很好的经济效益，长期高于沪深 300 指数，说明具有很好的超额收益率表现。

4.5 XGBoost 模型效果检验

4.5.1 模型预测效果检验

R^2分数也称为决定系数，是用于评估回归模型拟合优度的一种统计指标。它表示自变量对于因变量的解释程度，即自变量能够解释因变量变异的比例。由于实验中使用逐年训练，每年训练均从 1 月开始，故每月的R^2分数会有所波动，但整体趋势较好。具体R^2分数图如图 4-3 所示。

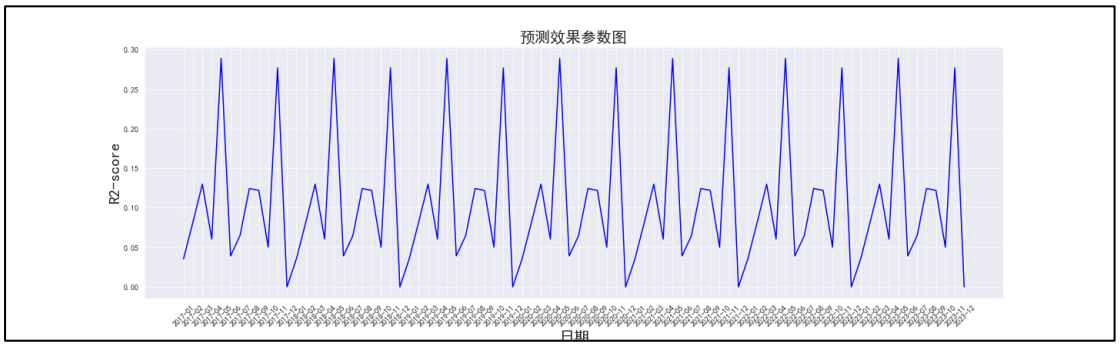


图 4-3 预测效果参数图

4.5.2 投资组合经济效益检验

实验中为验证所构造的投资组合是否具有良好的效果，采用夏普比率进行验证。夏普比率是由诺贝尔经济学奖得主威廉·夏普于 1966 年提出的一个金融指

标，用于衡量投资的绩效。夏普比率的核心思想是，投资者应该寻求在每单位风险上获得尽可能高的超额回报。因此，一个较高的夏普比率意味着投资在承担每单位总风险时产生了较高的超额回报。它特别适用于比较和评估具有不同风险水平的投资组合或金融工具的表现。具体计算公式如式（4-4）所示。

$$SHARPRATIO = \frac{R_p - R_f}{\sigma_p} \qquad (4-4)$$

其中是R_p投资组合的预期回报率；R_f是无风险利率；σ_p是投资组合的回报率的标准差。如果夏普比率为正，表示投资组合的回报超过了无风险利率，投资者获得的回报高于他们可以免费获得的无风险回报。

本文最终预测的夏普比率如下所示。从中可以看到仅在 2023 年夏普比率具有大幅下滑，但总体数值为正且数值较大，说明具有较高的超额回报。

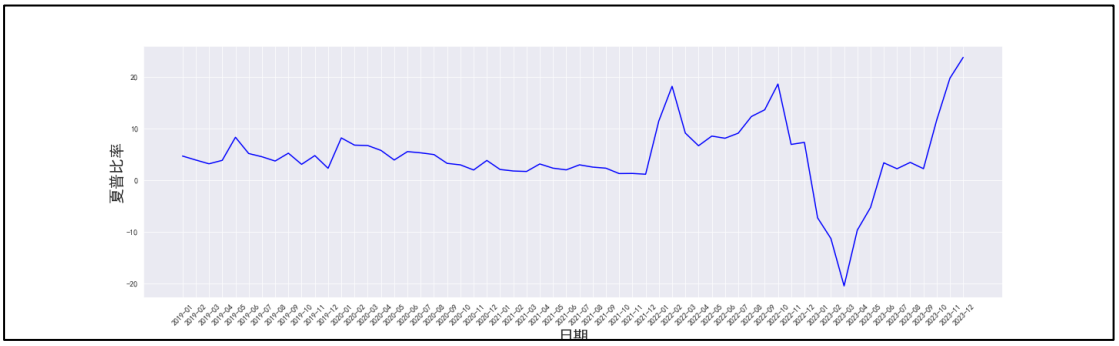


图 4-4 每月夏普比率折线图

4.5.3 超额收益率对比

将每年基于预测超额收益率构造的投资组合与沪深 300 进行收益率对比，不难看出投资组合具有高于市场指数的收益，说明基于该策略可以取得略高于基准的收益，实现正向超额收益的获取。具体收益率对比如图 4-5 和表 4-4 所示。

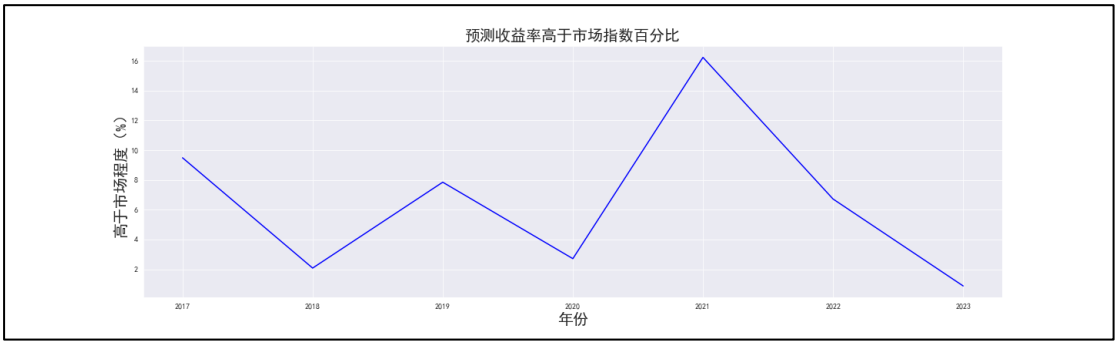


图 4-5 预测收益率高于市场指数百分比折线图

表 4-4　量化投资组合收益率高于市场知识百分比数据

年份	超额收益百分比（%）
2017	9.491346
2018	2.083134
2019	7.851312
2020	2.715911
2021	16.234667
2022	6.721212
2023	0.882541

第五部分 总结

在本研究中，我们通过对北向资金流动与 A 股市场相关性进行了深入分析，并构建了基于北向资金流动的量化投资策略，主要包括指数量化策略和多因子量化选股策略。研究结果表明，这些策略在实际应用中具有较高的实用性和有效性。

首先，通过对北向资金净流入和沪深 300 指数涨跌的相关性分析，我们验证了北向资金流动可以作为 A 股市场的有效预测指标。通过单位根 ADF 检验、格兰杰因果检验和 Pearson 相关性检验，研究发现北向资金流入与沪深 300ETF 涨跌之间存在显著的正相关关系。这为基于北向资金流动的指数量化策略提供了理论依据和数据支持。

之后，在构建指数量化策略时，我们通过对每日北向资金净流入数据进行高低判断，并基于此进行买卖操作。策略的模拟结果显示，从 2018 年开始，基于北向资金流入的指数量化投资策略取得了明显的超额收益，优于沪深 300ETF 的表现。

第四部分中，我们利用 XGBoost 机器学习算法构建了基于北向资金流动的多因子量化选股模型。通过对 2017 年至 2023 年北向资金投资的股票数据进行滚动训练，模型能够有效预测超额收益率，并构建高于市场指数的投资组合。实验结果显示，所构建的量化选股策略在多数年份中均实现了超额收益，且具有较高的经济效益。

最后，通过对策略效果的夏普比率和 R2 分数进行检验，我们进一步验证了模型的稳定性和有效性。结果显示，除 2023 年外，预测的夏普比率总体较高，说明投资组合具有较好的风险调整回报。

根据本文基于北向资金流动的研究思路，投资者可以利用港交所每日更新的

北向资金流动数据，洞察海外投资者的持仓动向，从而做出更为理性的投资选择。此外，通过分析不同托管机构的持股变化，投资者能够深入挖掘陆股通交易行为背后的信息，拓展研究视野。在构建投资组合时，投资者需实施风险控制措施，包括对投资结构、资产配置和市场风险的有效管理，以及定期的投资组合审查和调整，以适应市场变化，实现收益最大化。本文提出的选股策略可以结合市场风险控制和止损机制，或利用北向资金的流入流出作为短期择时信号，以提升投资策略的胜率。

总的来说，本研究基于北向资金流动构建的量化投资策略，通过理论分析和实证验证，证明了其在实际应用中的可行性和有效性。该研究为境内投资者和机构在追求超额收益及控制风险方面提供了新的思路和方法，也为量化投资领域的进一步研究提供了参考和借鉴。

第六部分 参考文献

[1] 王天屹. 基于机器学习的北向资金因子交易策略设计[D]. 沈阳工业大学,2023.

[2] 沈立, 王小雅, 李蕴霏, 等. 北向资金持股比例变化和股票交易的交互影响研究——基于面板 VAR 模型的实证分析[J]. 金融理论与教学, 2024,42(01):83-88.

[3] 陈健, 曾世强. 北向资金是 A 股的风向标吗？[J]. 金融市场研究,2024,(01):71-79.

[4] 杨何灿, 吴隽豪, 杨咸月. 北向资金与境内股票市场流动性——基于高频数据的传导机制[J]. 经济研究, 2023,58(05):190-208.

[5] 吴迪. 基于北向资金流动的量化投资策略研究[D]. 山西财经大学,2022.

[6] 李梦圆. 基于决策树的多因子选股模型研究[J]. 生产力研究,2024,(02):145-149.

[7] 欧阳资生, 唐伯聪. 基于 VMD-Bi LSTM-ATT 预测模型的碳中和指数量化投资研究[J]. 金融经济, 2023, (10):75-90.

[8] 周方召, 石祥翔, 贺志芳, 等. 量化投资基金和股票市场稳定性[J]. 金融经济学研究, 2023, 38(02):81-96.

[9] 朱睿. 基于LSTM 神经网络的股票指数价格预测与量化投资策略研究[D]. 云南财经大学, 2023.

[10] 孙博文. 基于机器学习算法的选股与择时量化投资策略研究[D].山东大学,2023.

[11]曹文康, 徐哈宁, 肖慧, 等. 基于贝叶斯优化 XGBoost 的多元输入模型对滑坡位移预测效果研究[J]. 测绘工程, 2024, 33(02):49-55.
[12]孙娜, 周绍伟, 潘姿宇. 基于 XGBoost-LSTM 模型的多特征股票价格预测研究[J]. 数学建模及其应用, 2023,12(04):32-39.
[13]Ayyoub F , Mohamed M , Mehdi M , et al. Harnessing LSTM and XGBoost algorithms for storm prediction[J]. Scientific Reports,2024,14(1):11381-11381.
[14]Sun Z , Wang X , Huang H , et al. Predicting compressive strength of fiber-reinforced coral aggregate concrete: Interpretable optimized XGBoost model and experimental validation[J]. Structures,2024,64106516-.
[15]Kalyani P , Manasa Y , Ahammad H S , et al.Retraction Note: Prediction of patient's neurological recovery from cervical spinal cord injury through XGBoost learning approach. [J]. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 2024.
[16]Liu M , Sun W , Chen J , et al. An automated quantitative investment model of stock selection and market timing based on industry information[J]. Egyptian Informatics Journal, 2024, 26100471-.
[17]Xu J .Fundamental Quantitative investment research based on Machine learning[C] //University of évora.Proceedings of 2023 International Conference on Digital Economy and Management Science(CDEMS 2023). Soochow University, Economics Department;,2023:5.
[18]Mao J , Wang K , Zhao W , et al.Research on Quantitative Trading Investment Strategy Based on LSTM and Dynamic Programming[C] //Wuhan Zhicheng Times Cultural Development Co. , Ltd..Proceedings of 2022 International Conference on Software, Data Processing and Information Technology (SDPIT 2022). College of Science,Chongqing University of Technology;Chongqing Key Laboratory of Public Big Data Security Technology; College of materials science and Engineering, Hebei University of Science And Technology; IT service center,Heibei Open University; College of Computer Science and Engineering, Chongqing University of Technology;,2022:7.