机器学习研究报告

基于集成学习的股票收益率预测和投资组合研究

摘要

本研究基于中国A股市场2010年至2024年的交易数据、财务报表和非结构化文本信息，构建了用于股票收益率预测与投资组合构建的机器学习模型框架。研究采用滚动窗口法整合滞后因子，并引入大语言模型嵌入和自然语言处理技术，将业绩预告、激励计划与公司定期报告等文本内容结构化。在建模方面，本文选取了LightGBM和Stacking两种集成学习方法，对2021年至2024年月度收益率进行预测，并基于预测结果构建月度投资组合策略。

实证结果表明，LightGBM与Stacking模型在样本内外均具有较高的预测精度，显著优于传统线性模型（弹性网）与深度学习模型（CNN）。基于模型预测构建的投资组合在2021—2024年间实现了显著优于沪深300指数的累计收益，并在经过Fama-French五因子调整后仍表现出显著的超额收益（FF5-α显著为正），夏普比率也大幅领先。最后，进一步的特征重要性排序显示，财务滞后变量与结构化的文本因子（如激励比例、净利润预测等）对收益预测具有关键作用。

本研究验证了集成学习方法在量化投资中的应用潜力，强调了多源异质性数据融合与特征工程的重要性，为量化选股提供了方法论参考与实证支持。

作 者

分析师

刘俊淇 骆翰森

参考文献

[1] 张宗新,吴钊颖.媒体情绪传染与分析师乐观偏差——基于机器学习文本分析方法的经验证据[J].管理世界,2021,37(01):170-185+11+20-22.DOI:10.19744/j.cnki.11-1235/f.2021.0011.

[2] 沈艳,陈赟,黄卓.文本大数据分析在经济学和金融学中的应用:一个文献综述[J].经济学(季刊),2019,18(04):1153-1186.DOI:10.13821/j.cnki.ceq.2019.03.01.

[3] 李斌,林彦,唐闻轩.ML-TEA:一套基于机器学习和技术分析的量化投资算法[J].系统工程理论与实践,2017,37(05):1089-1100.

[4] 戴德宝,兰玉森,范体军,等.基于文本挖掘和机器学习的股指预测与决策研究[J].中国软科学,2019,(04):166-175.

[5] 姚加权,张锟澎,罗平.金融学文本大数据挖掘方法与研究进展[J].经济学动态,2020,(04):143-158.

[6] 伊志宏,杨圣之,陈钦源.分析师能降低股价同步性吗——基于研究报告文本分析的实证研究[J].中国工业经济,2019,(01):156-173.DOI:10.19581/j.cnki.ciejournal.2019.01.009.

[7] 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].中国工业经济,2019,(08):61-79.DOI:10.19581/j.cnki.ciejournal.2019.08.004.

[8] 姜富伟,刘雨旻,孟令超.大语言模型、文本情绪与金融市场[J].管理世界,2024,40(08):42-64.DOI:10.19744/j.cnki.11-1235/f.2024.0090.

[9] 刘毅.因子选股模型在中国市场的实证研究[D].复旦大学,2012.

[10] 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.

[11] 黄秋丽,黄柱兴,杨燕.基于递归特征消除和Stacking集成学习的股票预测实证研究[J].南宁师范大学学报(自然科学版),2021,38(03):37-43.DOI:10.16601/j.cnki.issn2096-7330.2021.03.008.

[12] Sigletos G, Paliouras G, Spyropoulos C D, etal. Combining information extraction systems using voting and stacked generalization[J]. Journalof Machine Learning Research, 2005, 6(3): 1751-1782.

目录

图表目录

一、文献综述

李斌等（2019）[7]在对96个异象因子与12种机器学习模型（如Lasso、XGBoost、DNN等）的系统比较中发现，非线性机器学习算法显著优于线性算法，深度学习算法显著优于传统非线性算法，其建立的深度前馈网络预测的多空组合月度收益最高可达3.41%，验证了机器学习在基本面量化投资中的有效性。黄秋丽等（2021）[11]则结合递归特征消除（RFE）和Stacking集成学习（SVM与RF为基础模型，LR为元模型）预测沪深300周收益率，模型在准确率与F1值上均优于单一算法，证明以Stacking为代表的集成学习模型在选股中的稳定性与实用性 。此外，ML-TEA模型（李斌等，2017）[3]将机器学习与技术分析结合，利用多个技术指标预测股价涨跌，三种策略年化收益均超25%，夏普比率显著高于基准与买入持有策略，进一步验证了集成方法的投资价值。

现有研究表明：Stacking集成学习模型能有效整合多种机器学习模型，提升预测准确性；特征选择（如RFE）与非线性建模（如DNN）结合可优化选股效果；多源数据（基本面、技术面、文本面）融合是提升模型表现的重要路径。同时，李斌等（2019）[7]使用的滚动窗口法给予了本研究灵感，使用滚动窗口来匹配滞后效果。

综上，Stacking集成学习模型在股票收益率预测中展现出强大的优势，结合有效特征选择和多因子输入，为量化投资策略的构建提供了可行、优越的技术路径。

二、研究思路

（一）滚动窗口法

本研究试图通过股票交易数据、财务面数据和文本面数据预测下一月的收益率，显而易见的是，影响下一月股票收益率的一定不止是本月的数据，还受到之前几个月的数据的滞后效应影响。因此在数据处理阶段，我们对数据进行了滚动窗口法处理，将第t期的月收益率与之前6月每月的相关数据匹配，以拟合半年区间内每月的滞后效果，提升模型的预测效果和严谨性。

（二）文本数据处理

同时，本研究扩展了数据获取维度，把文本数据进行了结构化处理。一方面，针对业绩预测、激励计划这类简短、语义明确、主题高度相似的文本，接入豆包大语言模型进行结构化提取，获得标准化的数据指标；另一方面，针对上市公司定期报告等大体量、主题复杂多样的文本，采用潜在语义分析（LSA）方法，提取主题向量特征，丰富了机器学习的特征工程。

（三）模型设计

本研究采用了LightGBM和Stacking集成模型两类集成学习方法对股票收益率预测进行对比分析。LightGBM是一种基于梯度提升框架构建的高效集成学习方法，通过构建多个浅层决策树，逐步拟合残差实现非线性关系的建模；Stacking则通过多层结构组合多个异质模型，其第一层训练若干个基模型，第二层训练一个元模型，把基模型的预测结果作为第二层的输入，学习最优的加权组合方式。在本研究中，选用了弹性网（Elastic Net）、LightGBM、卷积神经网络（CNN）三个模型作为Stacking的基模型，分别作为线性、非线性、深度学习模型的典型代表，保证了基模型的异质性；选用了岭回归（Right Regression）作为元模型。

大量文献指出，在处理诸如股票收益预测等复杂非线性回归问题中，集成学习方法通常优于单一模型，能够显著提升预测精度（如Sigletos et al., 2005）[12]。其中，Stacking相比简单加权或投票策略，能够更充分地挖掘模型间的互补性，因此被广泛应用于金融、医疗、文本等多个预测场景。

以下是本研究采用的两个模型的流程图：

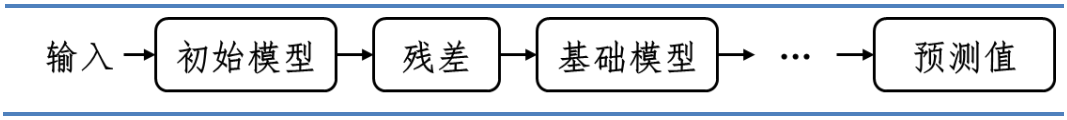

图1 Boosting模型流程图

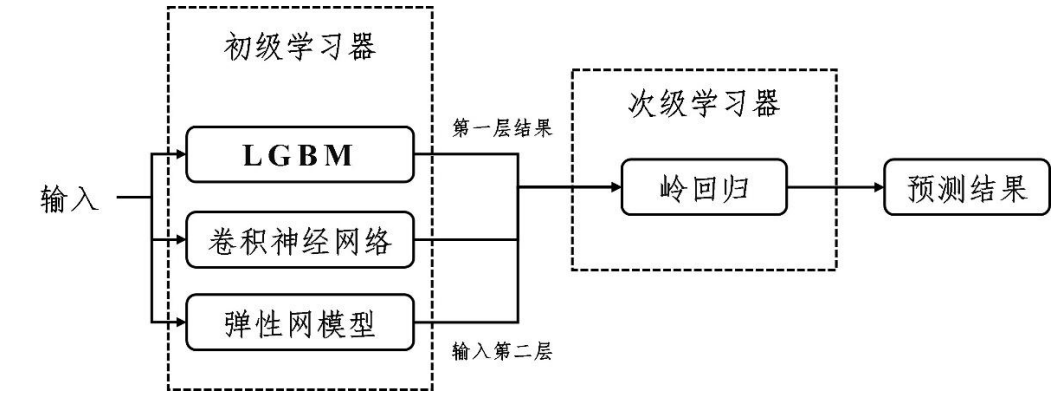

图2 Stacking模型流程图

（四）模型分析

根据模型预测所得的月收益率选定每月收益最高的前20支股票，并计算每月股票组合的实际平均月收益率。根据实际平均月收益率计算FF5-α和夏普比率与沪深300指数进行对比，并设计情景，对比模型所得股票组合与沪深300指数的实际收益曲线，探究模型预测的股票组合是否可以获得比沪深300指数更高的收益。其中，FF5-α是指在

提出市场风格影响后还能带来的“净超额收益”，可以真正衡量“主动选股能力”；夏普比率则衡量每承受一单位风险所带来的超额收益，是一个风险调整后的收益指标，用来比较策略的“风险性价比”。

在进行模型预测、情景分析后，通过分析模型结果和指标来判断哪些因子对于股票收益预测和投资组合成分股的筛选起到了关键性作用。

三、样本和数据来源

本研究选取2010年1月至2024年12月中国A股市场所有上市公司为研究样本，数据为月度频率。为保证模型预测的严谨性和预测能力，样本按时间分为两个部分：2010年1月到2020年12月的数据作为训练和测试样本，2021年1月到2024年12月的数据作为预测样本。

（一）结构化数据

本研究中的结构化数据（如月度股票数据、财务报表数据）从锐思金融数据库（RESSET）中获取，并经过清洗、计算和合并等处理得到了38个特征，如下表所示。

类别	特征名称
交易面特征	收盘价(元)_ClPr 成交量(股)_Trdvol 每股收益(摊薄)(元/股)_EPS 净资产收益率(摊薄)(%)_ROE 每股营业利润(元/股)_OpPrfPS 每股营业收入(元/股)_IncomePS 成交金额(元)_Trdsum 总股数月换手率(%)_MonFulTurnR 流通股月换手率(%)_MonTrdTurnR 总股数平均日换手率()_AvgDFulTurnR 流通股平均日换手率()_AvgDtrdTurnR 市盈率_PE 市净率_PB 市现率_PCF 市销率_PS 收盘价涨幅
财务面特征	营业总收入()_TotOpRev 营业总成本()_TotOpCost 营业利润()_OpProf 利润总额()_TotProf 净利润()_NetProf 基本每股收益()_BasicEPS 稀释每股收益()_DilutEPS 信息发布日期_InfoPubDt 营收增长率 毛利率 流动资产合计(元)_TotCurrAss 非流动资产合计(元)_TotNCurrAss 资产总计(元)_TotAss 流动负债合计(元)_TotCurLia 非流动负债合计(元)_TotNCurLia

类别	特征名称
财务面特征	负债合计(元)_TotLiab 实收资本（或股本）(元)_PaidCap 所有者权益（或股东权益）合计(元)_TotShareEquit 负债率 财务杠杆 经营活动产生的现金流量净额()_NetOpCashFl 滞后天数

表1 结构化特征表

由于本研究选取了月度股票数据，而财务报表是按季度披露的，故而需要进行滞后匹配。本研究采取的合并方法是：每一期财务报表数据仅在其披露日期之后、下一期财报披露日期之前的时间段内生效，在财务报表的生效期与对应的股票交易信息匹配；考虑到财务报表披露后市场反应的动态效应，我们在匹配后还记录了财务报表发布的滞后天数。

考虑到因子数据与收益数据之间的时序依赖性，在获取第t期月度收益率及其对应的股票与财务特征后，本研究采用滑动窗口法构建预测样本集。具体而言，对于每一期t的月度收益率，均匹配其前连续 6 个月（t-6至t-1）的因子数据作为输入特征，用于模型的训练或测试。通过该方法，构建出以“未来收益率”为预测目标、“过去六个月因子”为输入的结构化数据集，确保特征变量具备因果、时间顺序，符合实际预测场景的设定。

在去除数据不完整的观测后，2010年1月至2020年12月共有有效样本257938条，2021年1月至2024年12月共有有效样本180585条。总体来看，每月样本量呈现上升的趋势，符合近年来我国资本市场持续扩容与信息披露制度逐步完善的背景。

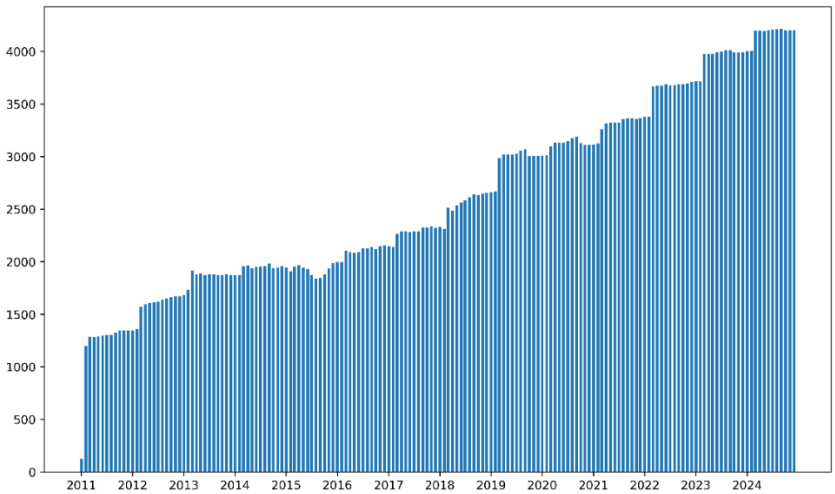

图3 每月样本量趋势图

特征的描述性统计显示，不同特征的取值在数量级和分布上差异较大，会影响模型的效果。因此，本文将训练、测试和预测集的控制数据标准化：

$$\tilde{X} = \frac{X - \bar{X}}{\sigma}$$

其中$\bar{X}$表X示的期望，σ表示X的标准差。

（二）非结构化数据

非结构化数据包括上市公司的业绩预测、激励计划和定期报告。其中，49667份业绩预测和6179份激励计划的原始文本数据来自锐思数据库；60029份定期报告的原始文本数据来自上海证券交易所官网披露的上市公司定期报告。

对于业绩预测和激励计划，本研究接入豆包大语言模型“doubao-1.5-pro-32k-250115”进行文本结构化提取。具体而言，通过构造针对性的 prompt，指导模型从原始自然语言中提取如“净利润预测区间”、“同比变动率”、“是否预盈”等关键财务指标，以及“激励工具类型”、“授予数量”、“激励对象构成”等激励计划核心要素。为提高处理效率与容错能力，整体结构化过程采用多线程方式批量调用 API，并对异常响应进行记录，采取人工处理JSON、修正prompt等方式，确保结构化数据完整性与一致性。模型输出以标准 JSON 格式返回，并通过字段筛选与缺失值统一填充策略，形成统一的结构化财务事件表，用于后续模型训练。

值得一提的是，在业绩预测的结构化处理中，针对原始文本中存在的非标准表达（如“亏损约5亿元至6亿元”），prompt 中加入了明确的字段单位、正负号处理和格式控制逻辑，确保提取数据在数值上具备一致性。

结构化业绩预测和激励计划数据得到特征如下表所示。

类型	特征名称
业绩预测	净利润下限（万元） 净利润上限（万元） 同比变动下限（%） 同比变动上限（%） 每股收益下限（元） 每股收益上限（元） 是否预盈
激励计划	授予数量（万股） 授予比例（%） 授予价格（元） 有效期（月）

表2 结构化文本数据特征表

对于上市公司定期报告，鉴于强制披露带来的报告数量庞大、报告本身的文本体量大，原始PDF文件累计超过百GB，基于BERTopic、LDA的主题建模和基于BERT的情感分析等自然语言处理方法都面临灾难性的内存爆炸、算力不足的问题，所以本研

究采用了潜在语义分析（LSA）作为下位替代方法，在批量提取PDF文字并转存为TXT文件后，执行分词、去停词，使用TF-IDF向量化构建稀疏文本矩阵，并通过SVD将文本特征降维至20维，形成结构化语义特征矩阵，有效捕捉了文本主题信息，作为文本因子输出后续模型。

值得一提的是，由于数据量确实庞大，这一部分设计了分批处理（每批10000份报告）和断点续跑机制，提升了训练的效率和稳定性；然而，分批处理机制决定了稀疏文本矩阵在第一批训练时就已经确定，但是笔者认为占总数1/6的10000份报告得到的文本矩阵已经足够有代表性了。

SVD特征对应的关键词如下表所示（切片）。

特征	关键词1	关键词2	关键词3	关键词4	关键词5	
SVD特征1	0	公司	现金	0	投资	…
SVD特征2	现金	活动	收到	支付	第三季度	…
SVD特征3	适用	2018	2017	2019	价值	…
SVD特征4	适用	0	主营业务	利润	0	…
SVD特征5	0	0	半年度	价值	金融资产	…
SVD特征6	2019	2018	2002	综合	收益	…
SVD特征7	2015	2016	2017	2014	2002	…
SVD特征8	2017	2018	适用	0	公司	…
SVD特征9	2013	2014	2012	集团	本行	…
…	…	…	…	…	…	

表3 SVD特征关键词对应表（切片）

数据结构化后，按股票代码和年月与结构化数据进行匹配，得到包含文本数据的最终结构化数据。

（三）Fama-French五因子数据

除此之外，为了评估模型选股的优劣，本研究还引入了FF5-α和夏普比率。为了计算这两个指标，本研究从中央财经大学金融学院官网获取了Fama-French五因子数据。获得如下特征：

特征名称	解释
trdmn	交易月份
mkt_rf	市场风险因子
smb	规模风险因子
hml	账面市值比风险因子
rmw	盈利能力因子
cma	投资模式因子
rf	无风险利率

表4 Fama-French五因子特征表

将五因子特征与通过模型选股得到的实际平均月收益率以及根据沪深300指数计算的月收益率通过年月合并即可求出FF5-α和夏普比率，判断模型选股的优劣。

夏普比率计算公式如下：

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p} \quad (\times\sqrt{12})$$

其中R_p是投资组合的平均收益率，R_f（rf）是无风险收益率，σ_p是投资组合收益率的标准差。

FF5-α则是将投资组合的收益率和Fama-French五因子进行线性回归得到截距，其中截距即FF5-α＞0且显著则说明该投资组合在去除市场因素干扰后依然有正收益。

四、模型构建与调参

（一）模型构建过程

根据上述研究思路，在完成样本选择和数据处理后，我们将结构化的2010-2020年的数据随机划分为训练集与测试集两个部分，其中训练集占比70%，测试集占比30%。由于本研究只处理了上交所的上市公司定期报告数据，所以接下来的模型构建不基于定期报告的语义特征，后文将针对上交所公司单独建模，以保证分析结果的一致性与可比性。

构建模型的总体流程如图所示。

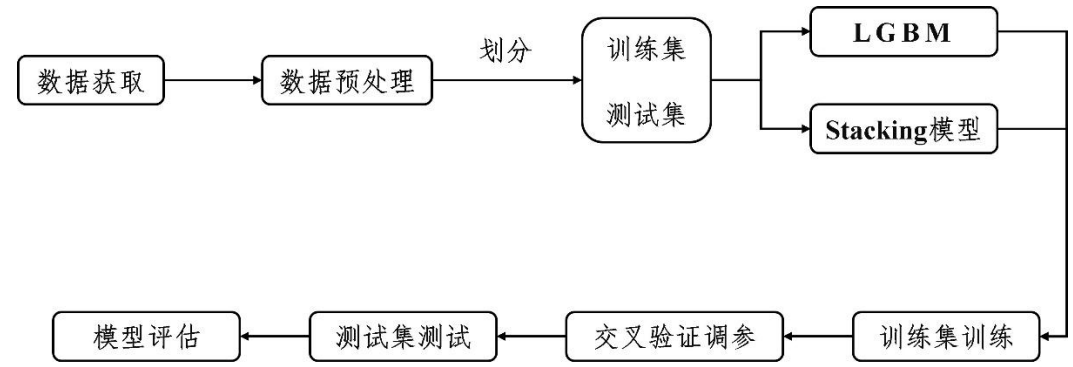

图4 建模流程图

在本研究建模过程中，首先按前述的模型设计训练两个模型（LightGBM、Stacking模型），在训练过程中使用K折交叉验证进行调参，最后使用测试集进行测试，评估模型在样本内（训练集）和样本外（测试集）的有效性。

（二）调参过程

首先我们构建了LightGBM模型，并采用 K 折交叉验证进行超参数调优。鉴于LightGBM模型在训练效率和资源消耗方面表现优越，选用了5 折交叉验证以平衡模型稳定性与计算开销，并采用网格搜索方法调节叶子接节点数、最大深度、学习率、总共迭代的树数四个核心超参数。

在此基础上，进一步构建了Stacking集成学习模型。该模型的基模型选取了LightGBM、卷积神经网络（CNN）与弹性网（Elastic Net）；元模型采用岭回归（Ridge Regression）。其中，针对 CNN 模型的高计算复杂度，采用 3 折交叉验证并调节学习率、最大训练轮数、批处理大小及卷积核数量等超参数；弹性网模型采用 5 折交叉验证，优化正则化强度（alpha）和 L1 比例（l1_ratio）两个关键参数；在岭回归中同样采用 5 折交叉验证，仅对正则化强度参数进行调节。

各模型的具体超参数及其搜索范围详见下表。

模型	参数名称	参数池	最优参数
LightGBM	叶子节点数	[31, 50]	50
	最大深度	[-1, 10, 20]	10
	学习率	[0.1, 0.01]	0.1
	迭代的树数目	[100, 200]	200
卷积神经网络	学习率	[0.01, 0.001]	0.01
	最大训练轮数	[10, 20]	20
	每个批次样本数	[16, 32]	16
	CNN卷积核数	[8, 16]	16
弹性网	正则化强度	[0.1, 1.0]	1.0
	Lasso正则化比率	[0.3, 0.7]	0.3
岭回归	正则化强度	[0.1, 1.0, 10.0]	1.0

表5 模型参数调整表

五、实证结果

（一）模型评估指标分析

完成模型构建和调参之后，我们得到了各个模型在最优参数配置下的预测性能表现，如下图所示。

模型	R^2 In Sample	R^2 Out of Sample	MSE In Sample	MSE Out of Sample
LightGBM	0.464	0.334	0.011	0.014
弹性网模型	0.000	-6.107	0.020	0.021
卷积神经网络	0.010	0.004	0.020	0.021
Stacking模型	0.482	0.330	0.011	0.014

表6 模型评估指标对比

从R^2和MSE等评估指标来看，集成学习模型（包括LightGBM和Stacking）在预测精度上明显优于弹性网模型与卷积神经网络模型，验证了集成方法在处理金融时间序列非线性特征时的优势。因此基于LightGBM和Stacking算法的选股模型更可能战胜市场，获得更高的收益。

（二）股票评估指标对比

在获得模型预测的月度平均收益率后，本研究分别基于各模型的预测结果进行排序，选出每月预测收益率排名前20的股票构建投资组合，并计算出对应月份的实际平均收益率。我们将模型得出的收益率走势与沪深300指数进行对比，并在假设初始投资资金为1元的情景下，绘制了2021-2024年期间各投资策略收益情况与沪深300指数的对比曲线图，结果如下图所示。

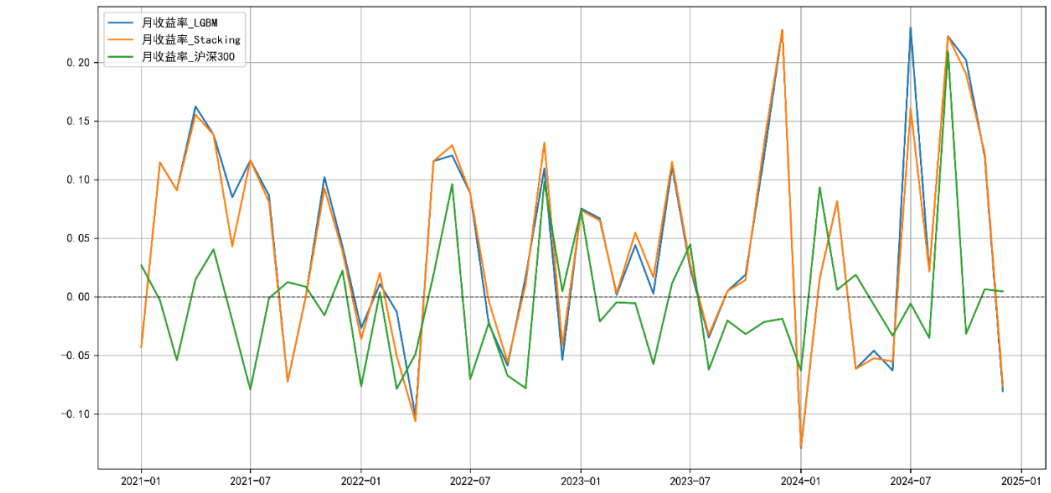

图5 投资组合月收益率曲线图

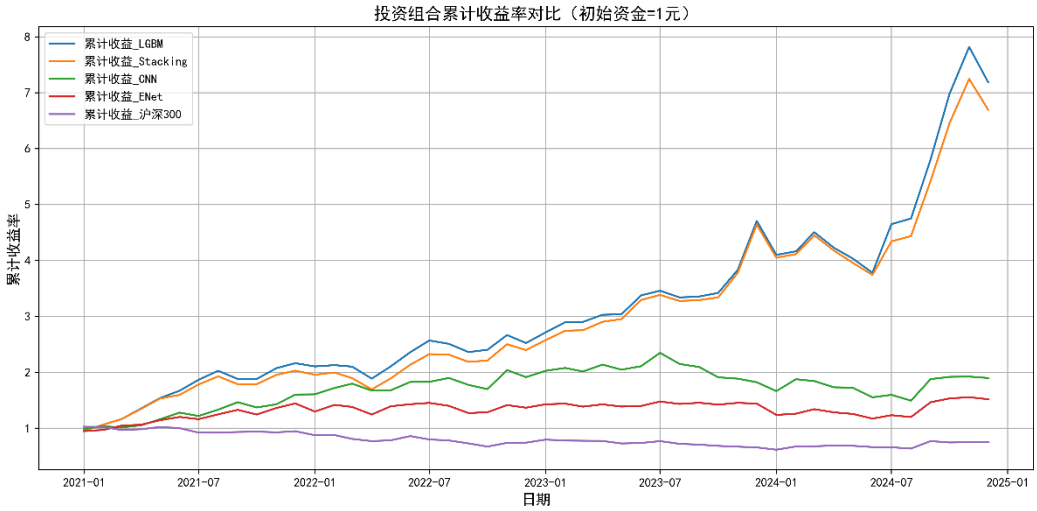

图6 投资组合累积收益曲线图

在直观对比模型与市场的收益走势的基础上，本研究引入了FF5-α和夏普比率作为评估股票组合风险调整后收益水平的专业指标。具体而言，我们各投资策略的月度实际收益率与Fama-French五因子联立，并估计得到每种策略的FF5-α值，然后计算对应的夏普比率，结果如下表所示。

模型	FF5-α	夏普比率
LightGBM	0.0382**	1.727
Stacking模型	0.0366**	1.710
卷积神经网络模型	0.0088	0.674
弹性网模型	0.0009	0.506
沪深300指数	-0.0026	-0.362

**指FF5-α在1%的水平下显著

表7 各投资组合股票评估指标对比

从表中我们我们可以看出，LightGBM模型和Sracking模型对应投资组合的FF5-α指标均为正且显著，表明在控制市场及其他风格因子后，这两种模型仍能够实现显著的超额收益；同时，LightGBM和Stacking模型的夏普比率也远高于其他模型及沪深300指数，说明这两个模型推荐的投资组合“收益风险性价比”较高。

结合收益走势的直观对比及两项专业评估指标，我们可以认为基于LightGBM模型和Stacking模型的预测结果构建选股策略是具有跑赢市场的能力的。然而，值得注意的是，尽管Stacking模型在集成过程中融合了LightGBM与其他模型的预测结果，其整体的表现并不如LightGBM，这说明在本研究样本和设定下，LightGBM的预测能力更为稳健。

（三）重要特征排序

在完成模型训练、预测以及情景分析后，我们进一步对表现最优的LightGBM中的各个特征进行了重要性排序，并提取了其中排名前二十的关键特征。由于数据处理使用了滚动窗口法，最终模型中共包含239个特征，因此我们可以认为位列前20的特征对于股票收益预测及投资组合成分股的筛选起到了关键性的作用。结果如下图所示。

序号	特征
1	业绩预测有效期（月）
2	激励授予比例（%）
3	激励授予价格（元）
4	激励授予数量（万股）
5	财务报表滞后天数_前1年
6	预测每股收益下限（元）
7	是否预盈
8	收盘价变化_前1年
9	预测同比变动下限（%）
10	预测净利润下限（万元）
11	成交金额(元)_前1年
12	预测每股收益上限（元）
13	收盘价变化_前2年

序号	特征
14	财务报表滞后天数_前2年
15	每股收益(摊薄)(元/股)_前1年
16	财务报表滞后天数_前6年
17	预测净利润上限（万元）
18	预测同比变动上限（%）
19	财务报表滞后天数_前5年
20	财务报表滞后天数_前3年

表8 股票收益预测中最具贡献的前20个特征变量

（四）引入定期报告语义特征的实证增益分析

为评估定期报告的语义信息是否有助于股票收益率预测，笔者基于上文所述的LightGBM和Stacking集成学习模型，对上交所的1476家公司在2010年1月至2020年12月的38501条月度观测数据构建了两组对照模型：一组引入了通过TF-IDF+SVD方法提取的20维语义特征，另一组未引入并保持股票交易数据、市场数据、其他文本数据一致

实证结果显示，以Stacking集成学习模型为例，引入语义特征的模型在样本外预测表现更优，说明文本信息有助于增强模型的泛化能力。尽管该模型在FF5-alpha和夏普系数上的表现不如基准模型，但是在累计收益率上表现更为强劲，表明该模型用于长期选股更有优势。

评价指标	已使用语义特征	未使用语义特征
R^2 In Sample	**0.657**	**0.659**
R^2 Out of Sample	**0.345**	**0.321**
FF5-α	**0.014**	**0.014***
夏普比率	**0.600**	**0.703**
累计收益	**1.993**	**1.823**

*在5%水平下显著

表9 引入语义特征前后模型性能与投资表现对比

综上，定期报告中的文本语义特征作为非结构化信息补充，能够改善模型的预测表现与累积收益，在长期选股中展现出一定增益效果，但对风险调整收益的提升仍需进一步挖掘与优化。

六、研究结论

基于上述实证结果，LightGBM和Stacking模型在样本内外均表现出较高的预测准确性，明显优于传统的弹性网模型与卷积神经网络模型。并且，在构建以模型预测收益率排序的投资组合后，我们发现这两种模型均能实现显著的超额收益，其FF5-α指标为正且在1%的显著性水平下成立，夏普比率也远高于沪深300指数，体现出较强的“风险性价比”。特别是LightGBM模型在各项指标中均优于Stacking模型，展现出更稳健的

预测性能。总而言之，使用LightGBM和Stacking模型筛选投资组合从理论上具有打败市场的能力。

进一步的特征重要性分析表明，滚动窗口引入的滞后财务变量、文本结构化指标（如激励比例、预测净利润）在模型中占据关键地位，说明多源异质性数据的融合与时序动态建模对于提升选股效果具有显著作用。

综上所述，本研究验证了基于集成学习的机器学习方法在股票收益预测与投资组合构建中的有效性与可行性，为量化投资策略的设计与因子选择提供了方法论的参考与实证支持。

七、思考与启发

在本次研究中遇到的最大的困难就是数据量庞大，且特征很多，导致正常的集成学习模型没办法运行完成。所以，在构建模型的过程中，我们一直在寻找可以节省算力，更加轻量化的模型进行预测，这对我们进行模型设计提出了比较大的考验。我们原本在Stacking中准备使用随机森林和支持向量机模型作为初级学习器，但是由于算力不足，只能更换。在换成卷积神经网络模型后，有因为使用CPU运行导致模型训练了一整晚都没有跑出结果，只能转而使用GPU运行，提升算力。

除此之外，对于上百GB的公司定期报告的语义分析也是一大难题，在我们的个人电脑上执行传统自然语言处理方法诸如LDA主题建模、BERT情感分析都面临灾难性的性能瓶颈，最后采用了潜在语义分析方法（LSA）作为下位替代，有效提高了股票收益预测准确性；但是我们相信公司定期报告能挖掘的信息不止LSA、能提升的收益预测准确率也不止于现有模型，希望有机会找到更优方法、更强算力进一步研究。

同时，在对于数据合并的处理中，由于出现了很多月度、季度无法对齐以及部分数据只有部分股票的部分日期有的情况，在这种情况下必须脱离一一对应去合并。例如，在合并股票数据和财务报表时，以财务报表发布日期和下一期财务报表发布日期作为起点和终点来确定匹配区间进行匹配。对于数据合并方面的实践，这样的操作需要一定的逻辑能力。

另外，接入大模型也是一个新奇的体验，相比于极度耗时的人工提取、繁杂且效果较差的正则表达式、适应性受限的第三方库，基于大语言模型的文本信息结构化不仅大幅提升了提取信息的效率和准确性，还能灵活应对不同的文本结构和表达方式，在实际应用中展现出极强的鲁棒性。这也让笔者意识到大模型在批量处理文本信息方面的巨大应用潜力。

当然，在研究中，我们还学习了一些和股票、证券市场相关的知识。为了评估模型推荐的股票组合的优劣，我们了解到了FF5-α和夏普比率作为投资组合的评估指标，

进而了解到他们的计算需要找到Fama-French五因子数据。通过Ai辅助，我们在中央财经大学金融学院官网找到了免费的数据资源。在进行本研究之前，我们基本上没有什么股票的基础知识，但是，在研究过程中，我们逐渐掌握了股票投资的知识和逻辑。