摘要

订单簿信息是金融市场交易中最重要的信息之一，电子化交易时代的到来驱动了围绕订单簿信息展开的高频交易的发展，并且衍生出了各类基于量价的市场操纵行为，包括但不限于“冰山订单”、“幌骗订单”等。具体到中国的证券交易市场，由于沪深主板的交易规则和主要参与者都与海外有所不同，因此带有市场操纵性质的高频订单信号也会呈现出跟部分文献不一致的性质，因此本文通过对深交所小市值股票的盘口数据进行研究，将观察到的市场情绪、交易逻辑、操盘逻辑等信息归纳总结为了 16 个风险因子，使这类因子贴近本土市场的交易逻辑。

此外不论是在资产定价领域还是其他研究领域，机器学习和深度学习的方法也正在被更广泛地采用。以 lightGBM 模型为代表的树模型在商业决策、信号选股等方面展现了较高的工作效率和准确率，注意力机制则在图像识别、视频生成、自然语言处理等跨专业场景都取得了不俗的效果。本文分别介绍了两类模型的基本原理，并对注意力网络做了加性注意力与多头点积的两种建模，通过自定义的数据输入方式对因子进行训练，并对两类模型的预测表现进行对比，得到了 lightGBM 在简单建模逻辑和小票高频因子的框架下表现通常优于注意力网络模型的实证结论。

关键词：高频信号，注意力机制，决策树模型，投资组合

An Empirical Comparative Study of Tree Models and Network Models Based on High-Frequency Deception Signals

Renzhuo Huang (Master of Finance)
Directed by: Ruixun Zhang

ABSTRACT

Order book information is among the most crucial data in financial market trading. The advent of electronic trading has spurred the development of high-frequency trading based around order book information, and has given rise to various forms of market manipulation based on volume and price. These include, but are not limited to, 'iceberg orders' and 'spoofing orders.' Specifically in China's securities trading market, due to differences in trading rules and main participants compared to overseas markets, the nature of high-frequency trading signals that may be manipulative can also be inconsistent with some literature. Therefore, this paper conducts a study on the market data of small-cap stocks at the Shenzhen Stock Exchange, summarizing observed market sentiments, trading logic, and manipulation tactics into 16 risk factors, aligning these factors closely with the local market trading logic.

Furthermore, both in the domain of asset pricing and other research areas, methods of machine learning and deep learning are being adopted more widely. Tree-based models, represented by the lightGBM model, have demonstrated high efficiency and accuracy in business decision-making and signal-based stock selection, while attention mechanisms have achieved notable results across interdisciplinary scenarios including image recognition, video generation, and natural language processing. This paper introduces the basic principles of both types of models and explores two modeling approaches with attention networks: additive attention and multi-head dot-product attention. By using customized data inputs for training on these factors, and comparing the predictive performance of both model types, the empirical conclusion is that lightGBM generally outperforms attention network models under the framework of simple modeling logic and high-frequency factors for small-cap stocks.

KEY WORDS: High Frequency Signals, Attention Mechanism, Decision Tree Model, Investment Portfolio

第一章　引言

1.1　研究背景

研究订单簿信息对理解金融市场的微观结构及其动态特性至关重要。作为金融资产如股票、债券、期货及外汇等的买卖订单的集合体，订单簿不仅记录了市场参与者的交易意愿，还体现了市场供需关系的即时变化：从微观角度看，订单簿信息提供了买卖双方报价的详细数据，包括价格、数量及其在时间序列上的变化情况，这对于分析市场参与者的策略行为、预测短期价格走势以及评估市场效率具有重要意义；从宏观角度来看，订单簿数据反映了整个市场的流动性分布和价格发现过程，是研究市场冲击反应、信息传递机制及市场稳定性的重要窗口，在极端市场事件下还能揭示市场恐慌或贪婪的集体行为模式。

另外一方面，随着机器学习、深度学习的研究热潮席卷世界，机器学习和深度学习在众多领域的应用推动了技术进步和产业变革。决策树作为一种直观且易于理解的机器学习方法，基于数据特征构建决策路径，做出分类或回归预测并为决策提供科学依据，在金融风险评估、客户细分、医疗诊断等场景中发挥着重要作用(Li 等, 2024; Park 等, 2021),；而深度学习中的注意力机制作为一项革命性的技术，在自然语言处理、语义分割、语音识别、图像识别、边缘检测等应用场景中均发挥了重要的作用(Kim 等, 2023)，注意力机制使模型能够在处理信息时自动聚焦于关键部分，提高了数据处理的效率和准确性，甚至近来最热门的大语言模型和 Sora 视频生成模型也是基于其技术底层予以实现。回望资产组合和资产定价领域近年来的文献，相关的研究文献也在不断增加，这说明将机器学习和深度学习引入到金融市场的研究已经成为学界和业界的趋势，本文就是在这样的背景下，希望能够结合决策树模型、注意力网络模型、订单簿信号进行建模，以期取得一些针对中国证券市场交易方面的洞见。

1.2　文献综述

订单簿信息是金融市场交易中最关键的信息之一，量化金融学界很早就对订单簿的相关信息展开过研究。早期的学者受到简陋的交易模式的和不完备的数据留存情况等诸多限制，只能基于简单的假设进行理论构建，比如Roll (1984) 基于线性模型假设对成交时的买卖最优报价进行建模，他通过引入一种衡量隐含交易成本的方法，证明了股票价格变动的自相关性可以用来估计有效的买卖价差。Glosten 和 Lawrence (1987) 则将交易的参与者分类，展示了做市商如何在存在信息不对称的市场中调整买卖价差

去规避知情交易者的套利行为，并基于此去解释买卖价差的来源和交易价格的统计特性。

而随着各个金融市场交易所的交易系统完成电子化，不仅仅交易数据的种类在不断增多，可以被记录的交易数据的频次也在不断提升，学者们也因此得以借助日趋完善的数据库搭建出更加精细化的研究框架，在此基础上更进一步的完成相关实证工作。Lo 等 (1992) 等人通过对连续价格假设下的研究建立了跳价概率预测模型，Easley 等 (1996) 则切入未公开信息的研究视角，研究了信息流和流动性之间的相互作用，此外Lo 等 (2002) 进一步通过对限价单的集中研究，结合生存分析和概率模型等，在涉及“异步交易”的处理方法上构建了新的经济计量模型框架。

千禧年后，随着业界高频交易参与者的增多，以提供流动性作为利润支撑的电子做市商开始崭露头角，这个阶段的研究也开始大量涉足业界实务中的各类难题，Avellaneda 和 Stoikov (2008) 在他们的工作中提出了著名的存货模型，用以支持高频做市商在支持程序化交易的各种金融市场中参与报价，Laruelle 等 (2010) 基于随机过程的理论基础对高流动性市场中的订单进行了最优化分单机制的研究，Bogousslavsky 和 Collin-Dufresne (2023) 通过对订单簿的特征构造了一系列代理变量以衡量其失衡状况，国内学者Shi 等 (2024) 通过对沪深交易所的参与者进行专业度分类，基于该分类刻画不同订单造成的市场冲击和影响，取得了较好的预测效果。

前述研究着重刻画了高频交易领域具有普适性的统计特征，提供的视角也是较为通用的研究视角，但是在市场的实际运行中，“量价”作为价格波动的基础驱动元素是可以被人为操纵的，与前述研究不同，对市场操纵行为的研究在方向上更接近行为金融学的领域，需要对量价特征的来源进行更细致的归因，也是通用视角的衍生分支。在这方面，着眼于量价关系的文献最早见诸Weber 和 Camerer (1998) 对证券交易中的“错置效应”的研究，Weber 等人与同时代的其他研究者一样，受限于交易数据的不足只能给出理论建模和简单实证，但是其对交易者盈亏偏好的不对称性的刻画依旧给后来的研究者带来启迪。随着信息技术的发展，诸如电子邮件的通讯方式改变了人类既有的信息交互行为，这也间接地产生了新的操纵行为，Bohme 和 Holz (2006) 和Hanke 和 Hauser (2008) 都对电子邮件造成的“幌骗订单”进行了深入研究，Bohme 和 Holz 等人主要着眼于场外市场进行研究，他们对幌骗订单在 NBC 的粉单市场和 OTCBB 市场中造成的定价扭曲进行了分析，而 Hanker 和 Hauser 等人则更关注电子邮件驱动的幌骗交易的时序特性、欺诈连续性、交易日特征等性质，并采用面板数据回归的方式对相应的量价特征进行了系统化的实证研究。更直接的“幌骗订单”操纵行为则是由业界的高频交易者带动引发的，Lee 等 (2013) 在对韩国股票市场中的研究中首次对“幌骗交易”的特征进行了详细的刻画和较为严格的定义，此后的学术文献则较少谈及这一

高频数据领域的话题，反而更多地被记载于各市场监督机构的处罚书中。随着国内主要交易所的数据基础设施建设不断完善，同样也涌现了一批学者从各个角度出发对这一问题进行了较为细致和独到的研究，Kong 和 Wang (2014) 就对上交所和深交所存在的诱单交易进行了分析并形成了一套度量机制，而袁琳 等 (2023) 则基于证监会的处罚信息，对 2015 年到 2019 年的 229 例已被官方定性的幌骗交易进行了计量视角下的实证研究。

前述研究在实证方面多数落脚于传统的计量模型，理论方面则多与随机过程的理论体系有所交叉，因此很少跟机器学习和深度学习产生关联；另一方面，尽管近些年来机器学习已经被应用到金融建模领域之中，在多因子选股方面不论是业界还是学界均有质量较好的文献产出，比如Gu 等 (2020) 使用 Boost 回归树模型对股票数据进行了建模实证研究，Xiang (2022) 使用 lightGBM 模型对股票订单簿价格的隐含波动率进行预测，Tuncer 等 (2022) 则尝试用注意力网络对股指期货的价格序列进行时序预测，但是将订单簿、因子研究和机器学习中的非线性模型组合在一起的研究仍是不常见的，因此这方面的工作仍有较大的完善空间。

1.3 研究意义

我国证券市场的主要参与者与欧美市场的主要参与者存在差异，欧美市场经过多年的发展已经形成了机构参与者常年互相角力的局面，其程序化交易规则也在市场中参与者各类或良性或恶性的竞争中趋于完善。而在中国证券市场中，程序化交易的发展是显著滞后的，散户作为主板证券市场的主要参与者往往在面临机构的研发、交易优势中沦为“韭菜”，这在各类小盘股的挂单量信息中就有体现，具体的形式即为“诱单交易”和“暴力洗盘”。通过人为筛选此类带有套利性质的信号并结合机器学习模型进行实证研究，可以为订单簿数据的深度挖掘提供新的方法论。

此外，尽管学界已有部分工作(Gu 等, 2020; Tuncer 等, 2022; Xiang, 2022) 将机器学习和金融资产定价领域的问题做结合，但是将具体的机器学习模型直接应用到订单簿的盘口数据并针对交易行为进行实证的文章则较为少见，因此本文的工作也为相关的研究提供了一个可供参考的切入视角。总的来说，深入研究订单簿信息对于促进金融市场的透明度、完善市场微观结构理论、指导投资实践，甚至乎完善监管政策都具有较为重要的理论和实践价值。

1.4 本文工作

本文主要做了以下几个工作，目前由于订单簿的挂单五档信息在价格插入时是存在位移的，因此日内需要适当地重构订单簿来生成更稳定的信号，本人通过对市场情

绪的观察与分析，将盘口数据中的交易价格、挂单、撤单、成交单等特征归纳总结为16 个主要的统计类风险因子，这些因子涵盖了个股交易中的散户买卖逻辑、专业投资者的操盘逻辑和极端市场下的情绪信息，可以被划分为基础逻辑信号、描述性订单失衡信号、描述性幌骗订单信号、互相关信号、自相关信号等 5 个类别。

模型对比实证阶段，本文基于生成的 16 个信号进行了 lightGBM 建模和注意力网络建模。本文在注意力网络的建模阶段自定义了一个将特征和标签划分为键、值、查询的数据传入与训练机制，把全数据集切分成合适的批量大小，以滚动的方式分别训练模型。然后，结合 4 个非信号的基础特征（股票代码、交易日期、可交易状态、隔日开盘十分钟时间加权价格），在验证集上通过已经训练好的模型预测资产组合权重，形成隔日交易的组合并进行滚动回测。最终对线性模型、lightGBM 模型、注意力网络模型形成的资产组合进行对比，完成实证并得出结论。