基于 CatBoost 的 t0 日内交易策略

1 策略概述

本文基于单只股票的日内高频数据，基于 CatBoost 机器学习方法，对短线内（日内）价格走势进行预测，进而捕捉短期盈利机会。

具体而言，本文首先基于某只股票（代码：603000）的单日 3s 频率盘口快照信息构建因子，包括中位价格、订单簿不平衡等高频因子，以及均线等技术指标；其次，基于因子与未来 60s 收益率相关性（RANK-IC）筛选因子，以及基于网格搜索+交叉验证调整参数，以 10 个交易日为单位训练模型，预测中位价格走势；最后，基于万一的买卖佣金费率和万五的印花税，根据预测值设计日内高频交易策略，并设置相应的开仓、平仓、止损的阈值。策略性能表现如第 6 章所示。

2 数据样本

本文所使用的数据样本为代码为 603000 的股票 2023-07-03~2023-12-29 124 个交易日的 3s 频率盘口快照信息，包含：开盘价（open）、最高价（high）、最低价（low）、成交量（volume）、成交额（money）、10 档委托买价（bid_price1~10）、10 档委托买单挂单量（bid_volume1~10）、10 档委托卖价（ask_price1~10）、10 档委托卖单挂单量（ask_volume1~10）。

3 因子构建

本文所采用的因子池构造方式参考 Ntaksris et al.(2020)，因子池包括以下几部分：

第一部分，直接根据盘口快照信息构建高频因子，包括买卖价差、中位价格等因子，频率与盘口快照一致（3s 频率）；参考 Ntaksris et al. (2020), Kercheval & Zhang(2013)两篇文章，第一部分构建的高频因子如下图所示：

Basic Set	Description(i = *level index*, $n = 10$)
$v_1 = \{P_i^{ask},\ V_i^{ask},\ P_i^{bid},\ V_i^{bid}\}_{i=1}^{n}$,	*price and volume (n levels)*

Time-insensitive Set	Description(i = *level index*)
$v_2 = \{(P_i^{ask} - P_i^{bid}), (P_i^{ask} + P_i^{bid})/2\}_{i=1}^{n}$,	*bid-ask spreads and mid-prices*
$v_3 = \{P_n^{ask} - P_1^{ask}, P_1^{bid} - P_n^{bid}, \lvert P_{i+1}^{ask} - P_i^{ask}\rvert, \lvert P_{i+1}^{bid} - P_i^{bid}\rvert\}_{i=1}^{n}$,	*price differences*
$v_4 = \{\frac{1}{n}\sum_{i=1}^{n} P_i^{ask},\ \frac{1}{n}\sum_{i=1}^{n} P_i^{bid},\ \frac{1}{n}\sum_{i=1}^{n} V_i^{ask},\ \frac{1}{n}\sum_{i=1}^{n} V_i^{bid}\}$,	*mean prices and volumes*
$v_5 = \{\sum_{i=1}^{n}(P_i^{ask} - P_i^{bid}),\ \sum_{i=1}^{n}(V_i^{ask} - V_i^{bid})\}$,	*accumulated differences*

Time-sensitive Set	Description(i = *level index*)
$v_6 = \{dP_i^{ask}/dt,\ dP_i^{bid}/dt,\ dV_i^{ask}/dt,\ dV_i^{bid}/dt\}_{i=1}^{n}$,	*price and volume derivatives*
$v_7 = \{\lambda_{\Delta t}^{la},\ \lambda_{\Delta t}^{lb},\ \lambda_{\Delta t}^{ma},\ \lambda_{\Delta t}^{mb},\ \lambda_{\Delta t}^{ca},\ \lambda_{\Delta t}^{cb}\ \}$	*average intensity of each type*
$v_8 = \{\mathbf{1}_{\{\lambda_{\Delta t}^{la} > \lambda_{\Delta T}^{la}\}},\ \mathbf{1}_{\{\lambda_{\Delta t}^{lb} > \lambda_{\Delta T}^{lb}\}},\ \mathbf{1}_{\{\lambda_{\Delta t}^{ma} > \lambda_{\Delta T}^{ma}\}},\ \mathbf{1}_{\{\lambda_{\Delta t}^{mb} > \lambda_{\Delta T}^{mb}\}}\}$,	*relative intensity indicators*
$v_9 = \{d\lambda^{ma}/dt,\ d\lambda^{lb}/dt,\ d\lambda^{mb}/dt,\ d\lambda^{la}/dt\}$,	*accelerations(market/limit)*

图 1 高频因子

数据来源：Kercheval & Zhang (2013)

其中，由于盘口快照信息颗粒度不足，缺乏 Level 2 颗粒度的委托单、主动成交单、撤单信息，因此到达率等指标无法计算（具体而言，图中的高频因子 v7~v9 无法计算）。

此外，本文在高频因子中，引入了两组常用因子：由于其对于盘口多空双方势力的不平衡较为敏感，因此对于短线价格走势预测较为准确。两组因子分别是：订单簿深度不平衡指标（Cao & Hansch, 2009）和交易量订单流不平衡指标（Shen，2015）。

而对于第二部分的因子，本文参考 Ntaksris et al.(2020)，以每 10 个 tick 为一个“子交易日”，计算开盘价、收盘价、最高价、最低价、交易量，并引入常见的技术指标，部分指标

如表 1 所示。

表 1 技术指标池（按字母排序）

技术指标名称	技术指标含义
ABV	主力进出指标
ADL	腾落指标
ADR	涨跌比指标
AR	人气指标
BIAS	乖离率
Bollinger Bands	布林带
BR	意愿指标
DMA	平行线差指标
DMI	趋向指标
KDJ	随机指标
MACD	异同移动平均线
Mike Base	麦克指标
Momentum	动量
PSY	心理线
TAPI	指数点成交值
TRIX	三重指数平滑平均线
VR	成交量变异率

4 模型构建

4.1 CatBoost 模型简介

CatBoost (Categorical Boosting)是一种基于梯度提升决策树（Gradient Boosting Decision Trees, GBDT）的机器学习算法，专门针对类别特征的高效处理进行优化；与传统的 GBDT 算法相比，CatBoost 的突出优势在于其对类别特征的自动化处理能力，避免了传统方法所需要的繁琐编码（如 One-Hot Coding），并支持 GPU 多线程计算。

在股价预测中，CatBoost 具有独特的优势。股价数据具有强烈的非线性特征，且市场动态受到多种因素影响，如宏观经济变化、公司财报、市场情绪等。CatBoost 通过集成决策树捕捉这些复杂的非线性关系，从而提高预测精度。此外，股价预测中往往包含大量类别型特征，如行业分类、技术指标等，CatBoost 能够高效地处理这些类别数据，避免了人工特征工程的繁琐。由于其良好的抗噪声能力，CatBoost 能够有效避免过拟合，并在复杂的股市数据中提取有用信号，提升模型的稳定性和泛化能力。

4.2 因子组合构建

本文主要采用时间序列滑动窗口的策略进行模型构建、训练与预测。参考实盘操作准则，我们选择在每周日训练模型，训练的样本为最近 10 个交易日（即 2 周的样本）[1]，在经过特征工程和参数调优后，对当前周次 5 个交易日的价格走势进行样本外预测。本文的回测期选择为 2023-07-18~2023-12-29.

首先，关于特征工程，本文参考实盘操作准则，按照因子与未来 60s 收益相关性（Rank-IC）进行因子筛选，每次训练模型前，都会筛选出 20 个具有短线预测能力的因子。

此外，我们也编写了迭代特征消除（Recursive feature elimination，RFE）的特征工程函数，由于设备算力有限，最终未能采用。

[1] 滑动窗口长度的取值，已经过投研回测验证。

4.3 模型训练与预测

本文所构建模型的预测目标（输出变量）是下一个 tick 的价格走势，设置为：

$$Y=\begin{cases}1, if\ \dfrac{P_t-P_{t-1}}{P_{t-1}}>0.0001\\ -1, if\ \dfrac{P_t-P_{t-1}}{P_{t-1}}<-0.0006\\ 0, else\end{cases}$$

其中P_t为 t 时刻的中位价格，上述预测目标考虑到了交易成本，Y 取 1 时，做多有盈利机会；Y 取-1 时，做空有盈利机会。因此，模型属于多分类问题。

在选择因子组合、构建预测目标之后，本文对 CatBoost 模型的两个参数：学习率（learning rate）和树深度（depth）进行网格搜索调参。一方面，学习率决定了每次迭代中模型更新的幅度，较大的学习率可能导致模型快速收敛，但也容易陷入局部最优，而较小的学习率虽然可以更为精细的拟合数据，但会增加训练时间；另一方面，树深度控制了每棵决策树的复杂度，树深度过大、复杂度过高，就会导致模型过拟合，泛化能力不足，而树深度和复杂度过低，也会导致模型欠拟合，缺乏预测能力。

经过因子组合选取、预测目标构建、参数调整之后，本文进行滚动窗口的 CatBoost 模型训练。在训练中，参考本文也采用了逐批次训练（Batch-wise Training）和增量训练（Incremental Training）。在每个训练批次中，当前模型的训练不仅依赖于原始数据，还结合了前一批次的训练结果作为基线（baseline），其核心思想在于，通过逐步积累已有模型的预测信息，帮助新模型在训练过程中更好的调整参数、避免过拟合。且分批次训练，可有效缓解内存和计算资源的压力。

在完成模型训练后，本文对回测期内每个 tick 下一个 tick 的价格走势进行预测；对于多分类问题，在评估其样本外预测能力时，通常使用分类准确率（Accuracy）、查准率（Precision）、查全率（Recall）、F1 值（F1-Score）判断。图 1 是回测期内的样本外预测性能，可见，预测准确率最终稳定在 70%~80%左右，可以进行进一步的策略设计。

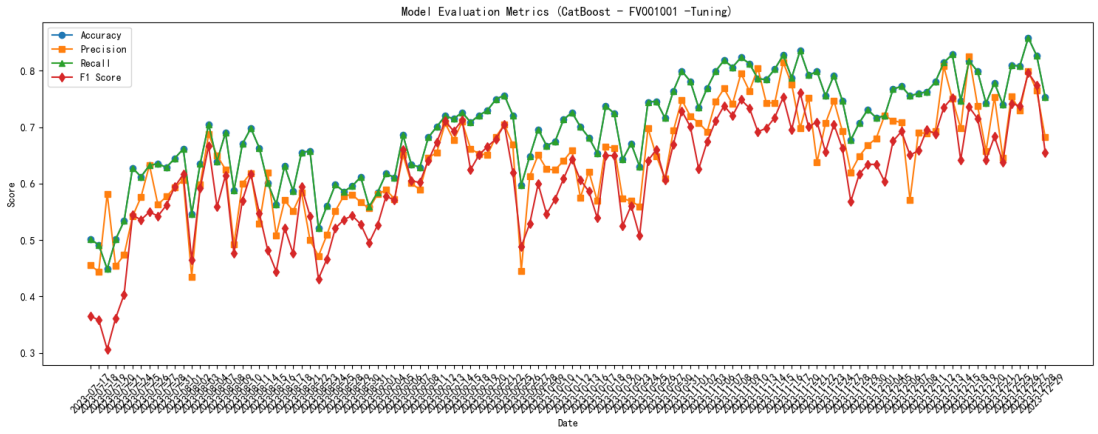

图 2 回测期内样本外预测性能

5 策略原理

本文构建的策略主要参考 2014 年广发证券金融工程专题报告《深度学习之股指期货日内交易策略——大数据深度学习系列之一》。策略逻辑如图 3 所示。

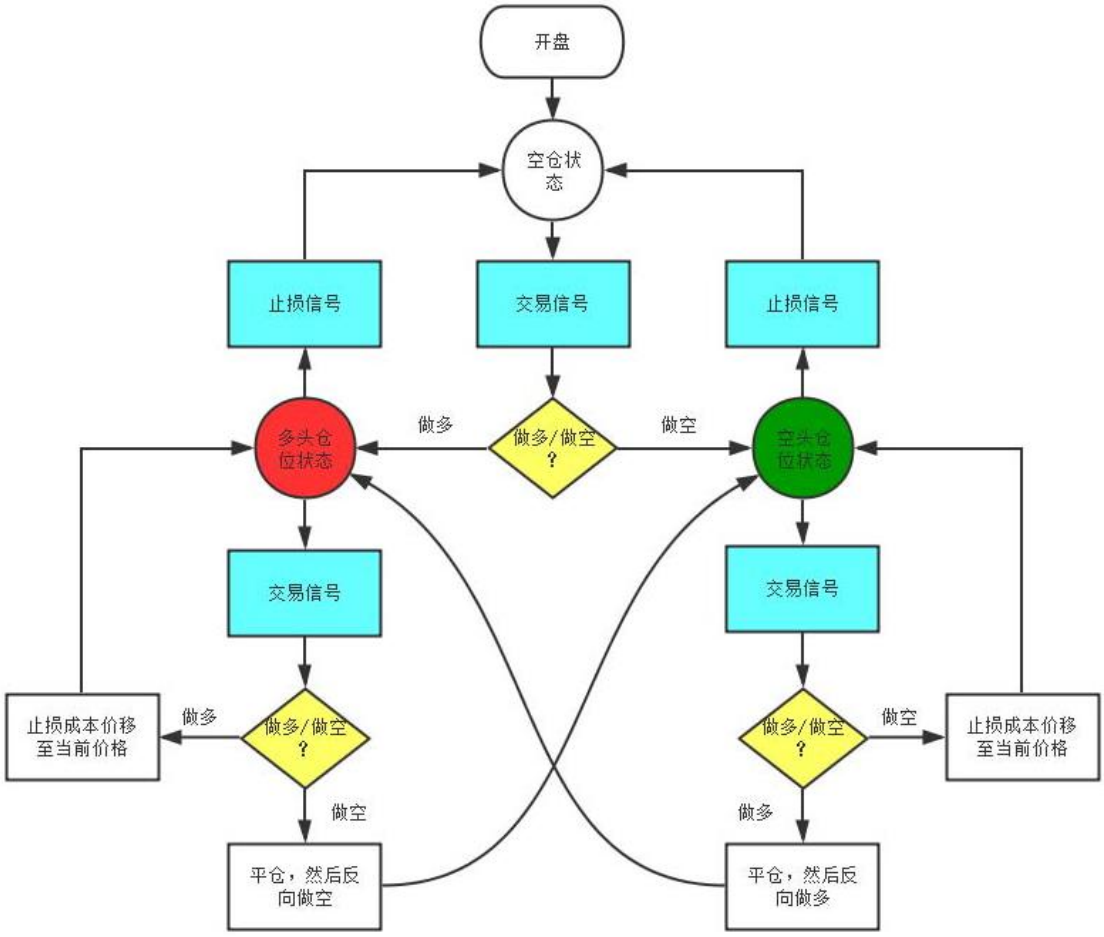

图 3 策略逻辑

数据来源:《深度学习之股指期货日内交易策略——大数据深度学习系列之一》

首先，基于机器学习分类预测模型，我们可以给出样本属于每个类别的得分值，位于[0,1]之间，因此我们得到上涨的预测得分 Score1 和下跌的预测得分 Score2；Score1 越大的样本，未来上涨的几率越大；相应的，Score2 越大的样本，未来下跌的几率也就越大。本文根据一定的做多阈值（BuyTrigger）和做空阈值（SellTrigger），设置相应的交易信号：

$$Signal = \begin{cases} 1, if\ Score1 > BuyTrigger \\ -1, if\ Score2 > SellTriger \\ 0, else \end{cases}$$

其中，Signal 取值为 1 代表做多信号，取值为-1 代表做空信号。关于 BuyTrigger 和 SellTrigger 的取值，本文设定：

$$BuyTrigger := P(Score1 > BuyTrigger) = \alpha$$

$$SellTrigger := P(Score2 > SellTrigger) = \alpha$$

也即，当预测得分超过$(1-\alpha)$的滚动分位数，则触发相应的做多和做空信号。关于α的取值，我们发现，在整个样本内，有大约 8%的样本，下一秒股价变化超过 0.0007，股价流动性较好，因此我们设置$\alpha = 0.04$.

其次，在设置信号之后，本文具体的交易逻辑如下：t0 交易要求每日开盘/收盘时均为空仓，每日交易由做多/做空信号触发，持仓时间不定。如果在持有多仓时，有空头信号触发，则立即平仓并反向建立空仓；若继续触发多头信号，则不改变持有头寸。相反，如果在持有空仓时，有多头信号触发，则立即平仓并反向建立多仓；若继续触发空头信号，则不改变持有头寸。由于市场噪音和突变的情况，预测模型只能准确预测一部分股价变化，故本文

设置 5%的价格止损和收盘前 10min 的时间止损。在持仓且亏损超过 5%的情况下，需要立即平仓止损；止损成本价以最近一次发出交易信号的股票价格为基准进行计算。在收盘前 10min，若持有头寸，需立即平仓，并不再对之后的交易信号产生反应。简化起见，每次多空操作所依据的价格均为对手盘 1 价。不考虑时间止损的收益。

本文根据上述策略逻辑生成回测区间内每日的交易单，示例如表 2 所示；entry_price 和 exit_price 分别代表入场、出场价格，entry_time 和 exit_time 分别代表入场、出场时刻，entry_type 和 exit_type 分别代表入场、出场方式，return 代表单次收益。

表 2 交易单示例

entry_price	entry_time	entry_type	exit_price	exit_time	exit_type	return
33.13331407	2023-07-19 09:30:00	开多	32.67038463	2023-07-19 09:31:15	平多	-0.01397
32.67038463	2023-07-19 09:31:15	开空	32.59325915	2023-07-19 09:31:18	平空	0.002361
32.59325915	2023-07-19 09:31:18	开多	30.88145962	2023-07-19 09:40:09	止损平多	-0.05252
32.11321161	2023-07-19 13:55:00	开多	32.11321161	2023-07-19 14:46:57	时间止损	0

6 策略评估

上述策略的各种性能指标如表 3 所示：年化收益率为 82.77%，与研报《深度学习之股指期货日内交易策略——大数据深度学习系列之一》的回测结果相近；最大回撤 15.74%，略大于研报中的回测结果，策略鲁棒性需进一步加强。回测期内部分 PNL 净值曲线如图 4 所示。

表 3 策略回测结果

年化收益率	年化波动率	夏普比率	最大回撤	最大回撤率
82.77%	0.5758	1.438	0.2189	15.74%

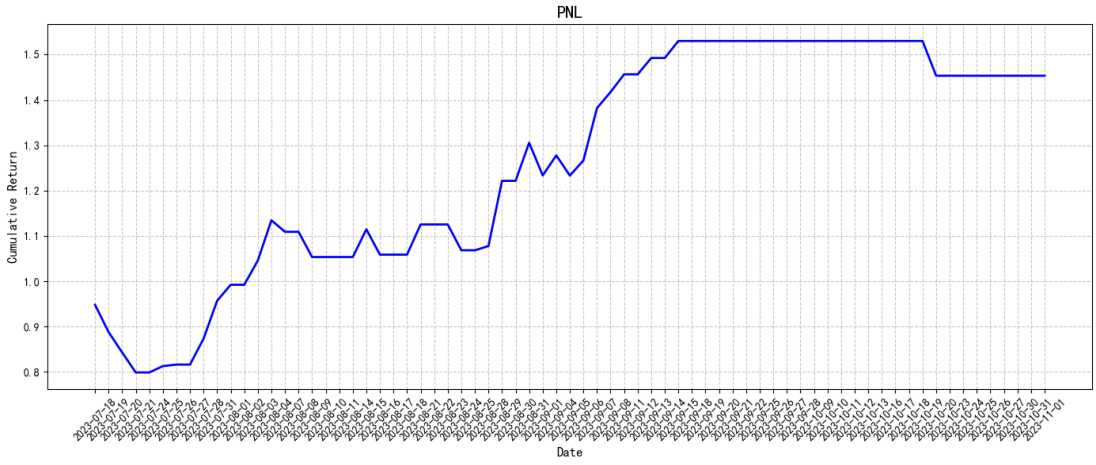

图 4 净值曲线

由于本文未考虑到对手盘挂单量、实盘发单笔数等，因此本文所计算的策略年化收益往往偏大。

此外，本文设置以下策略改进思路：（1）根据未来 60s 收益率的预测值设置交易信号，提升鲁棒性；（2）若多空信号较为强烈，引入二档抢单入场机制和分层出场机制；（3）引入

止盈机制，降低持仓风险、锁定收益；（4）在持有头寸且收到反向交易信号时，放弃上文所述的直接反向建仓策略，而改用“先平仓、再建仓”的策略，及时平仓。

参考文献

[1]. Cao C, Hansch O, Wang X. The information content of an open limit-order book[J]. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 2009, 29(1): 16-41.

[2]. Colby R W, Meyers T A. The encyclopedia of technical market indicators[J]. (No Title), 1988.

[3]. Kercheval A N, Zhang Y. Modelling high-frequency limit order book dynamics with support vector machines[J]. Quantitative Finance, 2015, 15(8): 1315-1329.

[4]. Ntakaris A, Kanniainen J, Gabbouj M, et al. Mid-price prediction based on machine learning methods with technical and quantitative indicators[J]. Plos one, 2020, 15(6): e0234107.

[5]. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.

[6]. Shen J. A pre-trade algorithmic trading model under given volume measures and generic price dynamics[J]. Applied Mathematics Research eXpress, 2015, 2015(1): 64-98.

[7]. 广发证券金融工程专题报告《深度学习之股指期货日内交易策略——大数据深度学习系列之一》，2014 年 6 月 18 日