

Kaggle 数据集描述

背景知识

拍卖市场（Auction Market）和连续市场（Continuous Market）是股票交易中两种不同的交易方式，它们在股票市场中有不同的作用和特点。

1. 拍卖市场（Auction Market）：

- 拍卖市场是指定时间段内，市场上的所有买卖订单被收集和匹配在一起，以确定一个统一的交易价格。
- 这个时间段通常被称为开市拍卖（Opening Auction）或收市拍卖（Closing Auction），它们分别在交易日的开市和收市时进行。
- 在拍卖市场中，买卖双方提交订单，而且通常不允许取消或修改这些订单。拍卖结束后，最终的交易价格确定，订单被执行。
- 拍卖市场的目的是确保公平和公正的价格发现，以便股票在市场开市和收市时能够以一个合理的价格交易。

a. 开市拍卖（Opening Auction）：

- 开市拍卖发生在交易日开始时，通常在交易所开市后的一段时间内。
- 这个时间段通常用于确定当天的开市价格，以确保市场在开始交易之前有一个公平的价格水平。
- 在开市拍卖期间，市场参与者可以提交订单，但这些订单不会立即执行，而是在拍卖结束后，形成当天的开市价格。

b. 收市拍卖（Closing Auction）：

- 收市拍卖发生在交易日结束时，通常在交易所收市前的一段时间内。
- 这个时间段用于确定当天的收市价格，以确保市场在结束交易之前有一个公平的价格水平。
- 在收市拍卖期间，市场参与者可以提交订单，这些订单在拍卖结束后形成当天的收市价格。

2. 连续市场 (Continuous Market) :

- 连续市场是股票市场的日常运行方式，它允许股票在整个交易日内以不断变化的价格进行交易。
- 在连续市场中，买卖订单可以随时提交，取消或修改。交易不受特定的开市和收市时间限制。
- 连续市场提供了更大的灵活性，允许投资者根据市场动态进行交易，并在市场开放时持续进行买卖活动。
- 价格在连续市场中会不断变化，取决于市场参与者的需求和供应。

这两种市场具有关联性，因为拍卖市场通常用于开市和收市时，以确保公平的价格发现，而连续市场则用于整个交易日的股票交易。拍卖市场的结果可能会影响连续市场的起始价格或市场参与者对股票的期望价格。

总的来说，拍卖市场和连续市场是股票市场中不同但相互关联的交易机制，用于不同时间段和情境下的交易活动。

数据集描述

1. stock_id – 股票的唯一标识符。不是每个时间桶中都存在所有的股票 ID。
2. date_id – 日期的唯一标识符。日期 ID 在所有股票中都是连续和一致的。
3. imbalance_size – 当前参考价格（以美元计）下的未匹配数量。举个例子，假设在某一时刻，股票 X 的参考价格为 50 美元，而买方和卖方之间存在差距。这意味着有一些买方想以 50 美元买入股票 X，但卖方却要更高的价格出售。这种情况下，就会产生未匹配的订单，其中 "imbalance_size" 表示了这些未匹配订单的总价值
4. imbalance_buy_sell_flag:
 - 该字段指示拍卖失衡的方向。
 - 如果它的值是 1，表示市场存在买方失衡。
 - 如果它的值是-1，表示市场存在卖方失衡。
 - 如果它的值是 0，表示市场没有失衡，即买卖方相对平衡。

例如，如果 "imbalance_buy_sell_flag" 的值是 1，这意味着在当前时刻市场上有更多的买方订单，可能会导致股票价格上升。

5. reference_price:

- 参考价格是在拍卖中匹配股票时的目标价格。
- 它通常等于最佳买价和卖价之间的价格，以确保匹配股票的数量最大化。

举例来说，如果股票 X 的最佳买价是 49.80 美元，最佳卖价是 50.20 美元，那么 "reference_price" 可能会设置为 50 美元，以最大程度地匹配股票。

6. matched_size:

- "matched_size" 表示在当前参考价格下可以成功匹配的股票数量，以美元计算。例如，如果 "matched_size" 的值为 1000 美元，这意味着在当前价格下可以成功匹配 1000 美元的订单。

7. far_price 和 near_price:

- "far_price" 和 "near_price" 都表示可能的交叉价格，以在拍卖和连续市场订单中最大化匹配股票数量。
- "far_price" 通常是在考虑拍卖兴趣的情况下计算的，而 "near_price" 还考虑了连续市场订单。

举个例子，如果 "far_price" 是 50.10 美元，而 "near_price" 是 50.15 美元，这意味着在不同市场情况下，这些价格可能会最大程度地匹配股票。

8. [bid/ask]_price 和 [bid/ask]_size:

- "[bid/ask]_price" 表示非拍卖市场中最具竞争力的买卖水平的价格。
- "[bid/ask]_size" 表示相应的最具竞争力的买卖水平上的美元名义金额。

例如，如果 "bid_price" 是 49.90 美元，而 "bid_size" 是 1000 美元，这意味着在市场上有人愿意以 49.90 美元的价格购买价值 1000 美元的股票。

9. wap (Weighted Average Price):

- "wap" 表示连续市场中的加权平均价格。
- 它是以不同价格级别的交易量加权计算的平均价格。

例如，如果某股票在一个时间段内以不同价格进行了多次交易，那么 "wap" 将是这些交易价格的加权平均值，其中交易量更大的价格对 "wap" 的贡献更大。

$$\frac{BidPrice * AskSize + AskPrice * BidSize}{BidSize + AskSize}$$

10. seconds_in_bucket:

- "seconds_in_bucket" 表示自当天收盘拍卖开始以来的秒数，但它通常用于在交易日结束后分析拍卖活动，而不是市场交易。

11. target:

- "target" 表示在未来 60 秒内，股票 wap（加权平均价格）的价格变动，减去合成指数（synthetic index）的未来 60 秒价格变动。
- "target" 仅在训练数据集中提供，通常用于训练机器学习模型。
- 该字段的单位是基点（basis points），基点是金融市场中的常见价格度量单位，1 个基点等于 0.01% 的价格变动。
- **Synthetic Index:**
 - 合成指数是由 Optiver 为此竞赛构建的一种自定义加权指数，该指数由纳斯达克上市的股票组成。
 - 它用于提供一个市场整体的参考点，以便评估股票 wap 价格的未来变动。
 - "target" 中减去合成指数的原因是为了提供相对于市场整体表现的股票价格变动。通过减去合成指数，可以消除市场整体趋势对股票价格变动的的影响，从而更专注于股票自身的价格波动。这有助于分析和预测股票的相对表现，而不仅仅是受整个市场影响的价格波动。假设股票的价格变动与合成指数高度相关，而且它们的价格走势通常是一致的，那么通过减去合成指数，可以使 "target" 变成一个相对强度或相对弱勢的度量。这有助于更好地理解股票价格相对于市场整体的表现。在金融分析和预测中，相对强度分析经常用于评估资产的相对表现，而不仅仅是绝对价格变动。因此，"target" 中减去合成指数是为了获得更准确的相对价格变动的度量。这种方法有助于投资者更好地了解股票的表现相对于市场整体的情况。