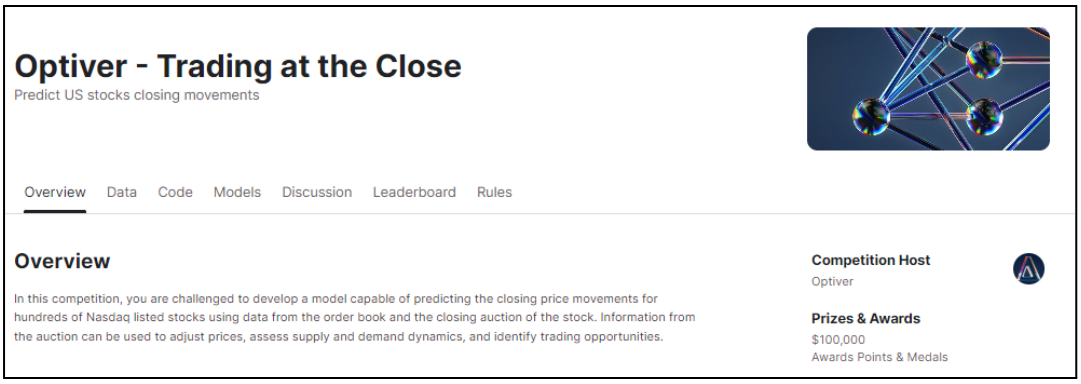
VLG Open Project Report

Optiver Trading At the Close

Aman Behera | Chemical Engineering | 22112013 | aman_b@ch.iitr.ac.in | 9587987718

Introduction:

Implemented a simple LightLGM model capable of predicting the closing price movements for hundreds of NASDAQ-listed stocks using data from the order book and the closing auction of the stock. for Optiver's Trading at the close competition.



Optiver - Trading at the Close

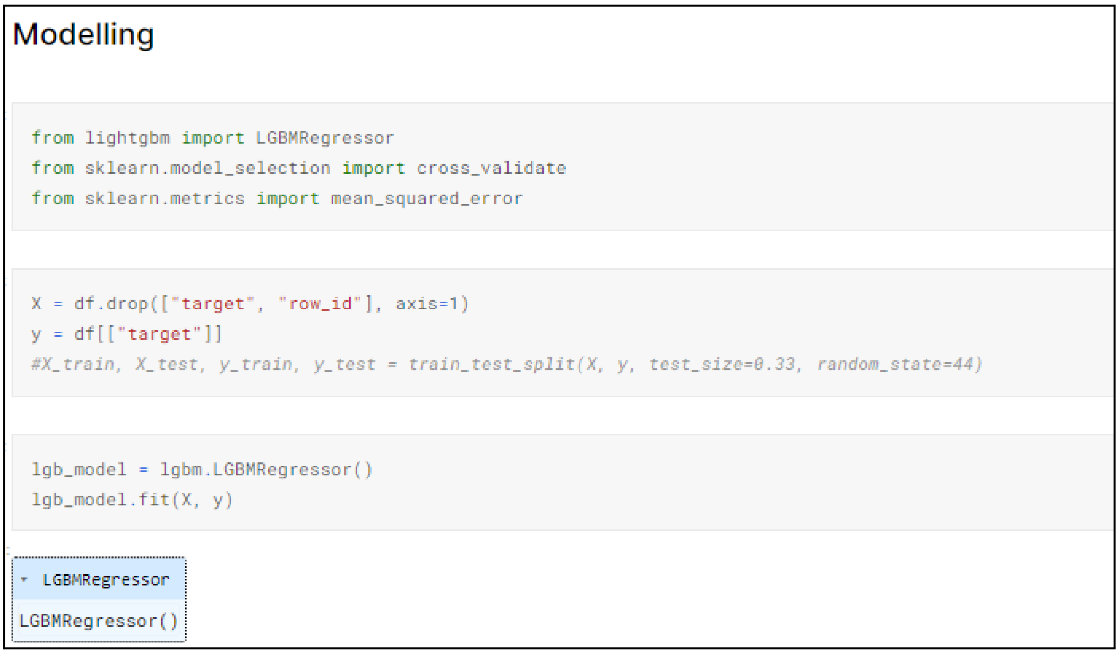
Predict US stocks closing movements

Overview

In this competition, you are challenged to develop a model capable of predicting the closing price movements for hundreds of Nasdaq listed stocks using data from the order book and the closing auction of the stock. Information from the auction can be used to adjust prices, assess supply and demand dynamics, and identify trading opportunities.

Competition Host
Optiver

Prizes & Awards
$100,000
Awards Points & Medals



Modelling

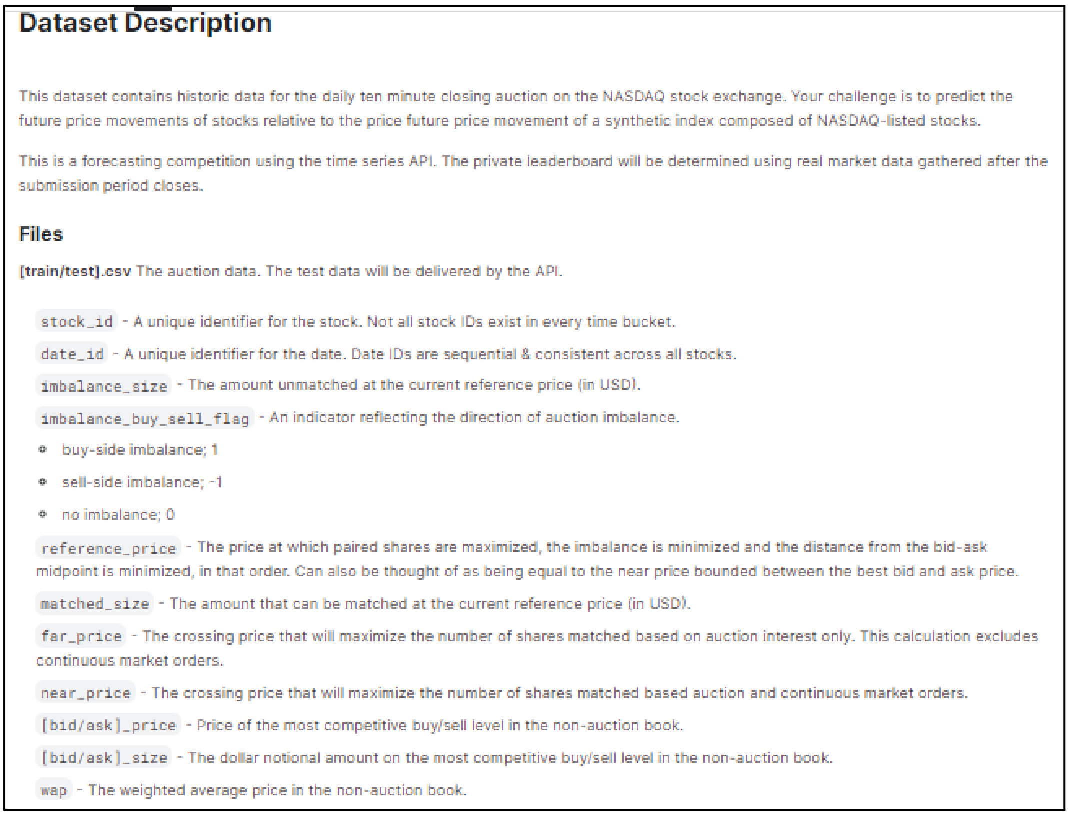
```
from lightgbm import LGBMRegressor
from sklearn.model_selection import cross_validate
from sklearn.metrics import mean_squared_error
```

```
X = df.drop(["target", "row_id"], axis=1)
y = df[["target"]]
#X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=44)
```

```
lgb_model = lgbm.LGBMRegressor()
lgb_model.fit(X, y)
```

```
▾ LGBMRegressor
LGBMRegressor()
```

Understanding the terminologies mentioned in the dataset's description became quite essential for attempting the problem statement. This dataset contains historical data for the daily ten-minute closing auction on the NASDAQ stock exchange. The challenge is to predict the future price movements of stocks relative to the price future price movement of a synthetic index composed of NASDAQ-listed stocks.



Dataset Description

This dataset contains historic data for the daily ten minute closing auction on the NASDAQ stock exchange. Your challenge is to predict the future price movements of stocks relative to the price future price movement of a synthetic index composed of NASDAQ-listed stocks.

This is a forecasting competition using the time series API. The private leaderboard will be determined using real market data gathered after the submission period closes.

Files

[train/test].csv The auction data. The test data will be delivered by the API.

- `stock_id` - A unique identifier for the stock. Not all stock IDs exist in every time bucket.
- `date_id` - A unique identifier for the date. Date IDs are sequential & consistent across all stocks.
- `imbalance_size` - The amount unmatched at the current reference price (in USD).
- `imbalance_buy_sell_flag` - An indicator reflecting the direction of auction imbalance.
  - buy-side imbalance; 1
  - sell-side imbalance; -1
  - no imbalance; 0
- `reference_price` - The price at which paired shares are maximized, the imbalance is minimized and the distance from the bid-ask midpoint is minimized, in that order. Can also be thought of as being equal to the near price bounded between the best bid and ask price.
- `matched_size` - The amount that can be matched at the current reference price (in USD).
- `far_price` - The crossing price that will maximize the number of shares matched based on auction interest only. This calculation excludes continuous market orders.
- `near_price` - The crossing price that will maximize the number of shares matched based auction and continuous market orders.
- `[bid/ask]_price` - Price of the most competitive buy/sell level in the non-auction book.
- `[bid/ask]_size` - The dollar notional amount on the most competitive buy/sell level in the non-auction book.
- `wap` - The weighted average price in the non-auction book.

Here, is what I understood about the different features of the column data:

- **imbalance_size:**

Imagine you're at a party and there's a game where everyone has to grab a balloon. The imbalance_size is like the number of balloons left over after everyone has grabbed one. If there are more balloons than kids, that's like a sell imbalance in the stock market. If there are more kids than balloons, that's like a buy imbalance. So, imbalance_size helps us understand how many items (like stocks or balloons) are left over after everyone has taken what they want.

- **reference_price**

Let's imagine you're at a school fair and there's a stall selling your favorite cookies. The reference price is like the price tag on the cookies. It's the price that everyone agrees is fair for the cookies.

Now, imagine there are more kids who want to buy the cookies than there are cookies available. This is like a buy imbalance in the stock market. On the other hand, if there are more cookies than kids who want to buy them, this is like a sell imbalance.

- **imbalance_buy_sell_flag**

The imbalance_buy_sell_flag is like a flag that the stall owner raises to let everyone know whether there are more buyers or sellers. If there are more buyers, they raise a 'buy' flag. If there are more sellers, they raise a 'sell' flag.

So in simple terms, these terms help us understand how many people want to buy or sell something (like stocks or cookies), and at what price they agree to do so.

- **matched_size**

Imagine you’re playing a game of musical chairs at a birthday party. The matched size is like the number of kids who have found a chair to sit on when the music stops. It’s a measure of how many kids (buyers or sellers) have successfully found a chair (shares) at the party (market). So, if there are 10 chairs and all 10 are taken when the music stops, the matched_size is 10.

- **far_price and near_price** *(Edited)*

Imagine you are at a school cafeteria, and you want to buy a slice of pizza. There are two lines: one for the students who have pre-ordered their pizza, and one for the students who are ordering their pizza on the spot.

The pre-ordered pizza line is shorter, so you decide to wait in that line. The students who pre-ordered their pizza can get their pizza for $2.00 per slice. This is the far price for the pizza.

The students who are ordering their pizza on the spot have to wait longer, but they can get their pizza for $2.50 per slice. This is the near price for the pizza.

The far price is lower than the near price because the students who are willing to wait are taking on less risk. They know that they will definitely be able to get their pizza for $2.00 per slice. The students who are ordering their pizza on the spot are taking on more risk, because they don't know if there will be enough pizza left when they finally get to the front of the line.

- **Bid/Ask Price**

Imagine you're at a garage sale, and you see a cool toy. The "bid price" is how much you're willing to pay for it, like saying, "I'll give you $5." The "ask price" is how much the seller wants for it, like saying, "I want $8 for this toy."

- **Bid/Ask Size**

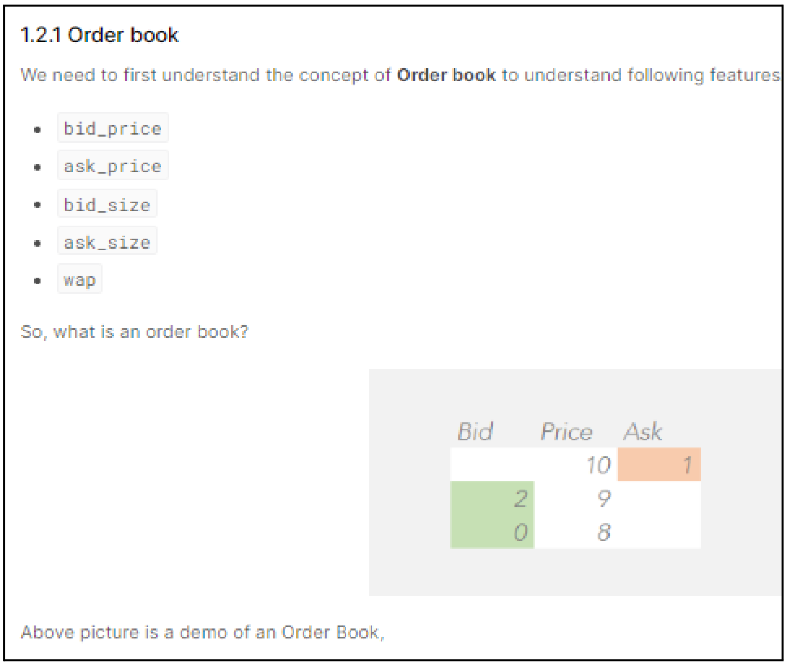
This is like seeing how many kids at the schoolyard want to buy or sell the same toy. If 5 kids want to buy the toy for $5 each, that's a "bid size" of 5. If 3 kids want to sell it for $8 each, that's an "ask size" of 3.

- **WAP (Weighted Average Price)**

Imagine you're buying candy from different friends, and each friend charges a different price. You buy 2 candies for $1 each and 3 candies for $2 each. The weighted average price is like asking, "On average, how much did I pay for each candy?" It's like adding up the total money spent on candy and dividing it by the total number of candies.

Explanation of some essential terminologies:

Many different terminologies like order book, auction book, market maker (like Optiver, Jane Street), and WAP required explanations. Although theoretical explanations can be accessed from Google or any other source, the following images serve to give a working insight into how trading at close happens. Lastly, since I had used quartiles for outlier detection an image explaining the same has also been added.



1.2.1 Order book

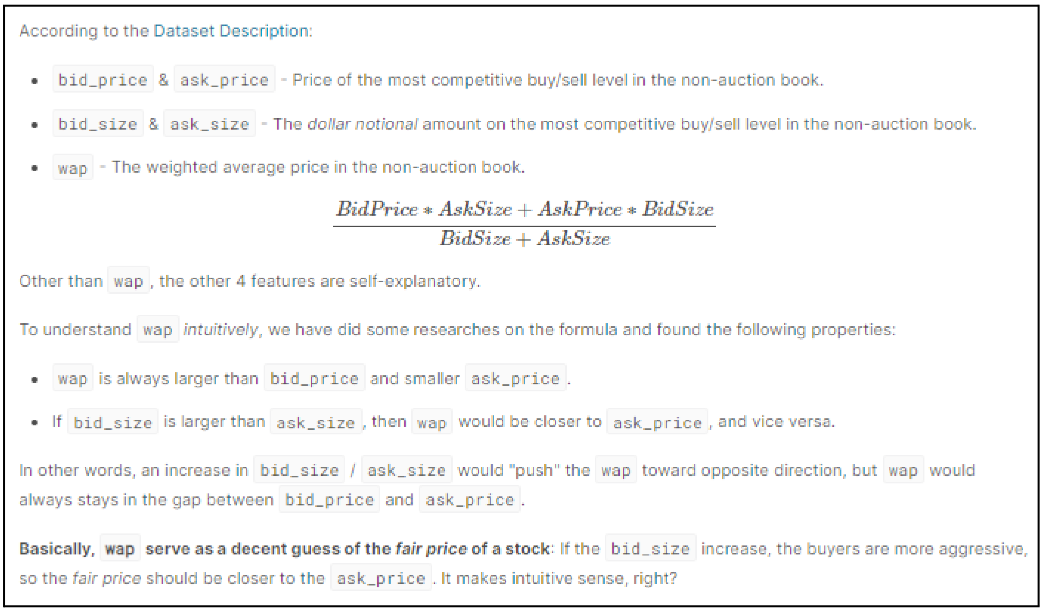
We need to first understand the concept of **Order book** to understand following features

- `bid_price`
- `ask_price`
- `bid_size`
- `ask_size`
- `wap`

So, what is an order book?

Bid	Price	Ask
	10	1
2	9	
0	8	

Above picture is a demo of an Order Book,



According to the Dataset Description:

- `bid_price` & `ask_price` - Price of the most competitive buy/sell level in the non-auction book.
- `bid_size` & `ask_size` - The *dollar notional* amount on the most competitive buy/sell level in the non-auction book.
- `wap` - The weighted average price in the non-auction book.

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

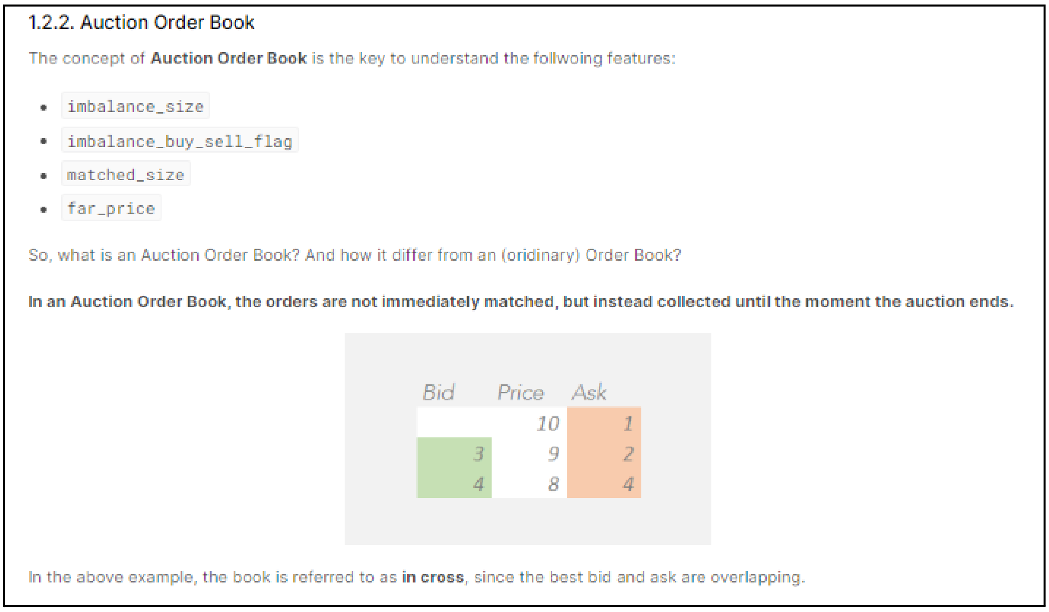
Other than `wap`, the other 4 features are self-explanatory.

To understand `wap` *intuitively*, we have did some researches on the formula and found the following properties:

- `wap` is always larger than `bid_price` and smaller `ask_price`.
- If `bid_size` is larger than `ask_size`, then `wap` would be closer to `ask_price`, and vice versa.

In other words, an increase in `bid_size` / `ask_size` would "push" the `wap` toward opposite direction, but `wap` would always stays in the gap between `bid_price` and `ask_price`.

Basically, `wap` serve as a decent guess of the *fair price* of a stock: If the `bid_size` increase, the buyers are more aggressive, so the *fair price* should be closer to the `ask_price`. It makes intuitive sense, right?



1.2.2. Auction Order Book

The concept of **Auction Order Book** is the key to understand the follwoing features:

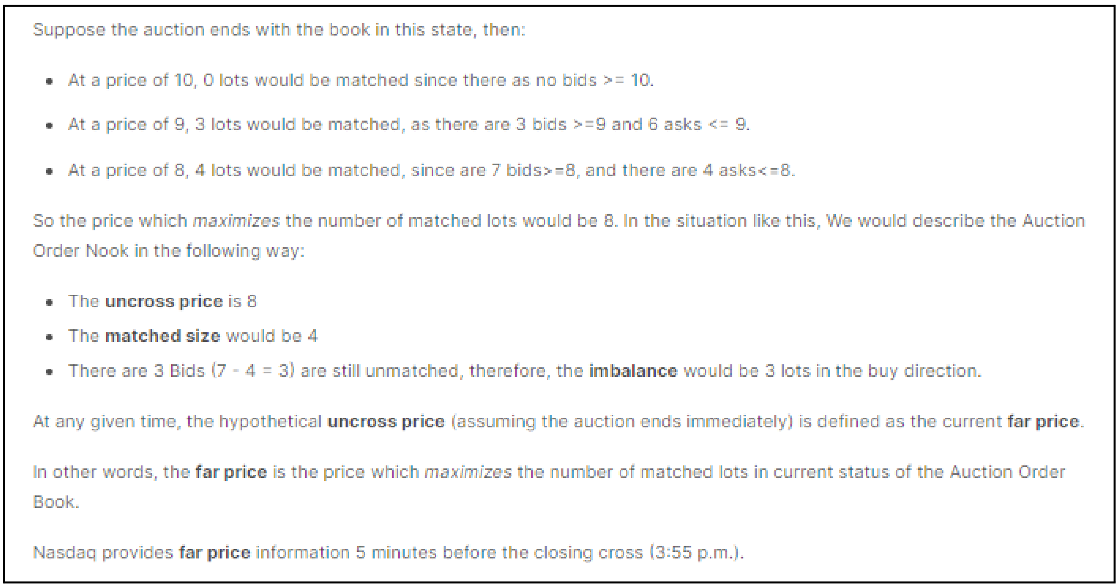
- `imbalance_size`
- `imbalance_buy_sell_flag`
- `matched_size`
- `far_price`

So, what is an Auction Order Book? And how it differ from an (oridinary) Order Book?

In an Auction Order Book, the orders are not immediately matched, but instead collected until the moment the auction ends.

Bid	Price	Ask
	10	1
3	9	2
4	8	4

In the above example, the book is referred to as **in cross**, since the best bid and ask are overlapping.



Suppose the auction ends with the book in this state, then:

- At a price of 10, 0 lots would be matched since there as no bids >= 10.
- At a price of 9, 3 lots would be matched, as there are 3 bids >=9 and 6 asks <= 9.
- At a price of 8, 4 lots would be matched, since are 7 bids>=8, and there are 4 asks<=8.

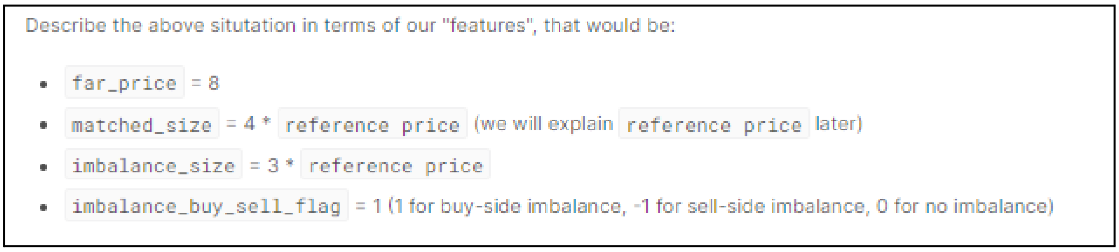
So the price which *maximizes* the number of matched lots would be 8. In the situation like this, We would describe the Auction Order Nook in the following way:

- The **uncross price** is 8
- The **matched size** would be 4
- There are 3 Bids (7 - 4 = 3) are still unmatched, therefore, the **imbalance** would be 3 lots in the buy direction.

At any given time, the hypothetical **uncross price** (assuming the auction ends immediately) is defined as the current **far price**.

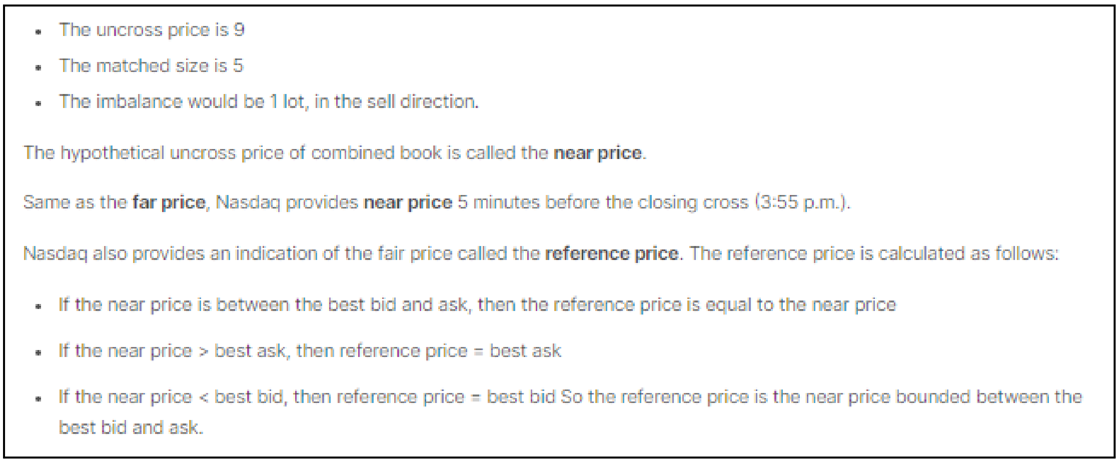
In other words, the **far price** is the price which *maximizes* the number of matched lots in current status of the Auction Order Book.

Nasdaq provides **far price** information 5 minutes before the closing cross (3:55 p.m.).



Describe the above situtation in terms of our "features", that would be:

- `far_price` = 8
- `matched_size` = 4 * `reference price` (we will explain `reference price` later)
- `imbalance_size` = 3 * `reference price`
- `imbalance_buy_sell_flag` = 1 (1 for buy-side imbalance, -1 for sell-side imbalance, 0 for no imbalance)



- The uncross price is 9
- The matched size is 5
- The imbalance would be 1 lot, in the sell direction.

The hypothetical uncross price of combined book is called the **near price**.

Same as the **far price**, Nasdaq provides **near price** 5 minutes before the closing cross (3:55 p.m.).

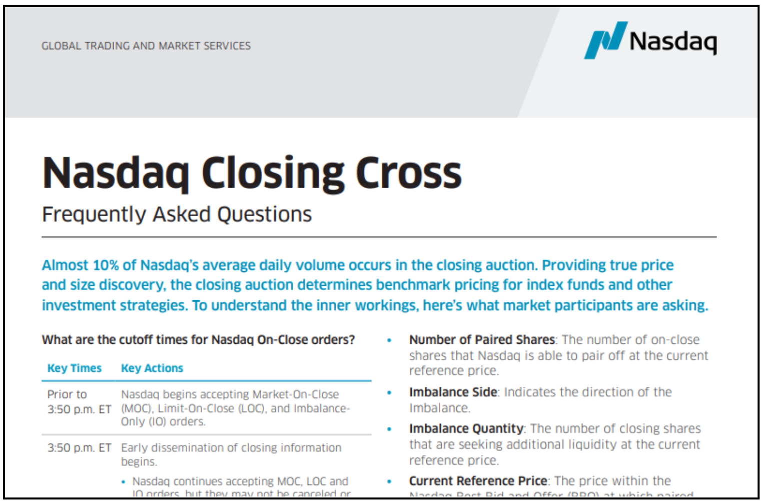
Nasdaq also provides an indication of the fair price called the **reference price**. The reference price is calculated as follows:

- If the near price is between the best bid and ask, then the reference price is equal to the near price
- If the near price > best ask, then reference price = best ask
- If the near price < best bid, then reference price = best bid So the reference price is the near price bounded between the best bid and ask.

Essential Resources and References Used:

Here are some of the resources I used to get myself familiar with the terminologies and to understand the workings of trading at the close.

1. Market Making
    - https://blog.quantinsti.com/market-making/
    - https://www.investopedia.com/ask/answers/06/brokerandmarketmaker.asp
2. Order types and order books
    - https://www.investopedia.com/terms/o/order-book.asp
    - https://www.investopedia.com/investing/basics-trading-stock-know-your-orders/
    - https://www.schwab.com/learn/story/3-order-types-market-limit-and-stop-orders
3. Auction price decision framework
    - https://www.investopedia.com/terms/a/auctionmarket.asp
    - https://www.nyse.com/article/parity-priority-explainer
4. Trading at close terminologies
    - https://www.nasdaqtrader.com/content/productsservices/Trading/ClosingCrossfaq.pdf
    -
    -



GLOBAL TRADING AND MARKET SERVICES

Nasdaq

Nasdaq Closing Cross

Frequently Asked Questions

Almost 10% of Nasdaq's average daily volume occurs in the closing auction. Providing true price and size discovery, the closing auction determines benchmark pricing for index funds and other investment strategies. To understand the inner workings, here's what market participants are asking.

What are the cutoff times for Nasdaq On-Close orders?

Key Times	Key Actions
Prior to 3:50 p.m. ET	Nasdaq begins accepting Market-On-Close (MOC), Limit-On-Close (LOC), and Imbalance-Only (IO) orders.
3:50 p.m. ET	Early dissemination of closing information begins. • Nasdaq continues accepting MOC, LOC and IO orders, but they may not be canceled or

- **Number of Paired Shares**: The number of on-close shares that Nasdaq is able to pair off at the current reference price.
- **Imbalance Side**: Indicates the direction of the Imbalance.
- **Imbalance Quantity**: The number of closing shares that are seeking additional liquidity at the current reference price.
- **Current Reference Price**: The price within the Nasdaq Best Bid and Offer (BBO) at which paired

Refer to the official documentation of Nasdaq to understand the terminologies associated with close

- Like what is close?
- Why doesn't trading happen 24x7?
- Why do market makers bring liquidity to the table?
- What is buy-sell disparity?
- And other questionable terminologies

To understand how the stock market works refer to the following resources:

https://youtu.be/p7HKvqRI_Bo?si=o_11VZY7VvDQ0_F2
https://www.youtube.com/watch?v=ZCFkWDdmXG8
https://youtu.be/Vb8JA3Y7aoE?si=_0ZtPnfJvIrKMohp
https://youtu.be/WZlApjlzrfY?si=-wt21al0o8NttKsA

Scope of Improvements:

I would next intend to tune the hyperparameters of my model, tried to implement it before mid-evaluation but wasn’t able to. Still got a decent score of 5.41 using the simple LGM model and outlier detection using quartiles percentiles. I might wish to seek heavy models like LSTM or NN, but since the score from my present model is good enough, I would hopefully stick to improving it, adding more visualization in the form of graphs/diagrams to make the code more engaging.

Experimenting with variations in the model/pipeline is another thing I am looking forward to, changes like increasing the folds of cross-validation, using other methods for outlier removals, using mean of the column to fill our NA values, etc.