# Order Book Analysis

## Artagan Malsagov

## January 17, 2024

# 1 Data Description

The data consist of 5 levels of both sides of the order book, for 5 different days. Each days spans roughly 10 hours worth of data (36 billion micros, see table below)

| date | min timestamp | max timestamp | avg bbo mid | avg 5 level order volume |
|---|---|---|---|---|
| 20190610 | 0 | 36000000000 | 10064 | 673 |
| 20190611 | 0 | 36000000000 | 10127 | 866 |
| 20190612 | 0 | 36000000000 | 9999 | 955 |
| 20190613 | 0 | 36000000000 | 10065 | 908 |
| 20190614 | 0 | 35999621354 | 9894 | 797 |

## 1.1 Data resampling

The order book is of the form below:

| timestamp | bp0 | bq0 | ap0 | aq0 |
|---|---|---|---|---|
| 110 | 10045 | 62 | 10055 | 98 |
| 175 | 10065 | 46 | 10075 | 42 |
| 220 | 10075 | 9 | 10080 | 25 |

where the timestamps are in microseconds. Plotting a histogram of the frequency of the timestamps, we see that the updates aren't uniformly distributed:
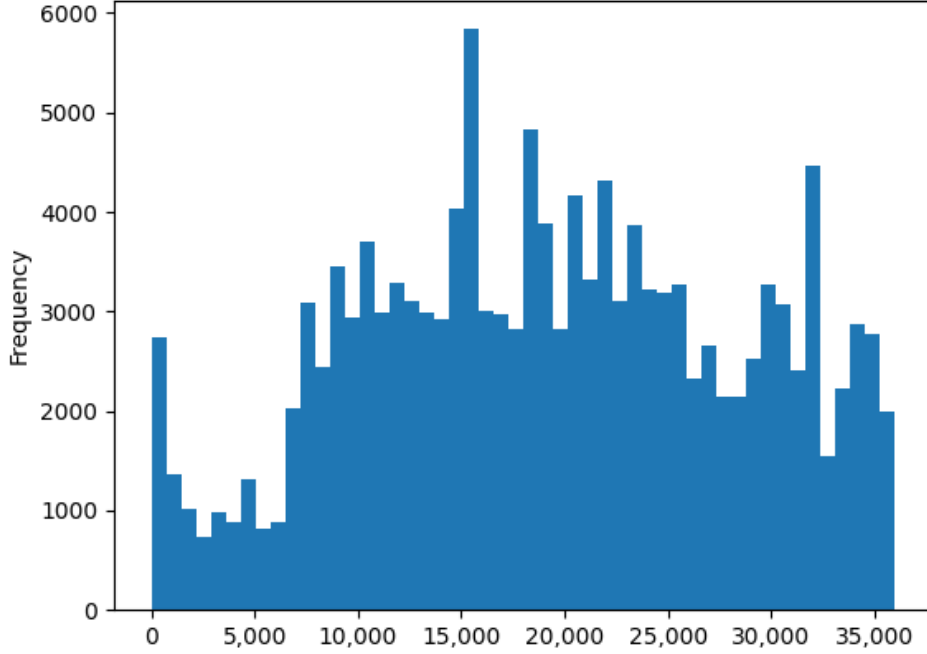
Figure 1: Histogram of order update arrivals for date 20190610

For the analysis the data will be resampled. Specifically, a grid will be used with a time interval of 100,000 micros = 100ms. This is done so as to reduce noise of the raw updates at microsecond level and detect any signals in the data. Obviously, this grid size might cause information loss and a more thorough analysis can be done to optimize, but for practical considerations 100ms will be used as the discretization step, since this will allow for quicker data processing and model estimation times. In practice this means the timestamps will be rounded up to the nearest 100,000 micros. The reason to round up is to avoid look-ahead bias when using the data as a trade signal, since rounding down will match an order update with a timepoint in the past. In addition, when discretizing to a grid, there might an issue when there are no updates available. In that case a forward interpolation is done by using the last known value.

## 2    Feature Selection

### 2.1    Target to predict

For the targets the following was considered:

- The simple average mid price:

$$P_{mid} = \frac{bp0 + ap0}{2} \qquad (1)$$

- The inverse volume weighted mid price:

$$P_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0}{bq0 + aq0} \qquad (2)$$

This mid has the benefit of taking into account the order imbalance at the top level: if the buy order volume is higher, the price will be skewed higher to the ask, and vice-versa if the sell order volume is higher.

- The inverse volume weighted mid price at the first and second level:

$$P_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0 + bp1 \times aq1 + ap1 \times bq1}{bq0 + aq0 + bq1 + aq1} \qquad (3)$$

This mid has the same advantage as the previous one and it also takes into account the second layer of the orderbook

For the target the simple average mid (equation 1) is used rather than one of the two inverse weighted mids. The reasoning being that the inverse weighted mids are predictors of sorts for the simple average mid.

Another interesing target would be the mid price change from time $t$ $t + \delta$ where the $\delta$ is set to a multiple of the discretization step 100ms. The logic being that the price moves is what the market focueses on, rather than the absolute price level. To keep things simple, the target variable to be predicted will be the simple average mid price.

Below are the plots of the mid price and iverse weighted mid price for different days. Looking at the plots there are no weird outliers. Also clearly the inverse weighted mid price closely track the simple average mid price.
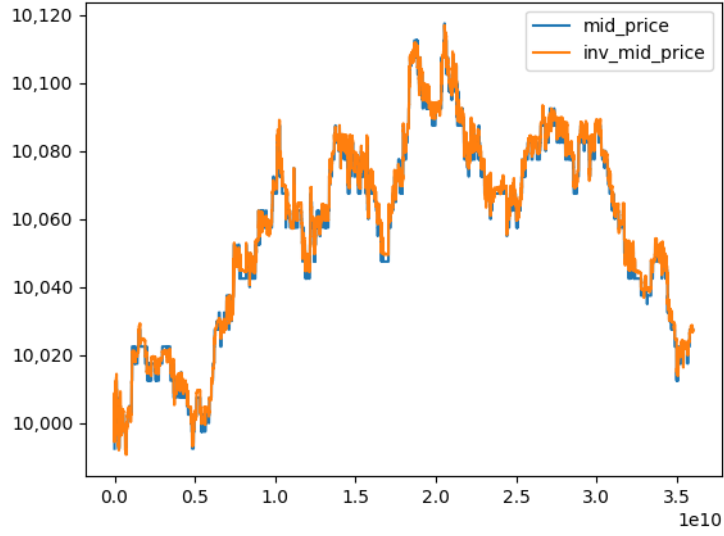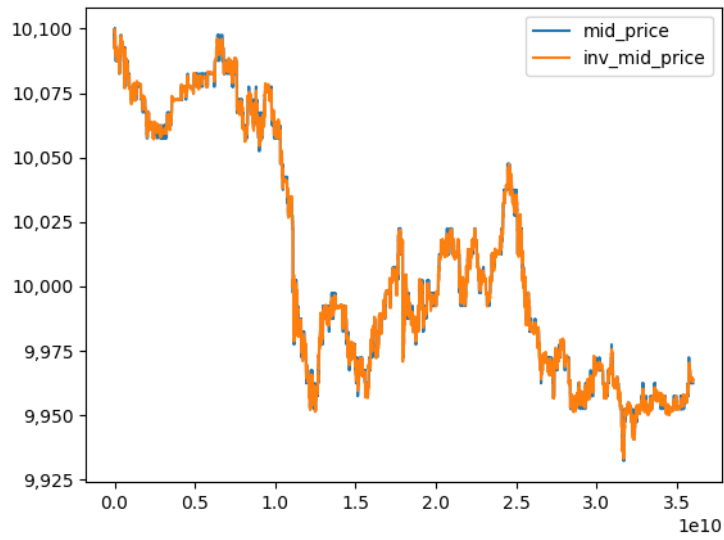
Figure 2: Day 2019-06-10



Figure 3: Day 2019-06-12

## 2.2 Features

For the features, the ones below are considerd:

- The bid-offer spread calculated as:

$$Spread = \frac{ap0 - bp0}{P_{mid}} \tag{4}$$

The intuition of using the spread as a predictor for the change of the mid-price is that if say the spread is relatively wide, then the probability of a non-markeatable limit order arriving whose price is inside the spread is also higher. Whereas if the bid-offer spread was very narrow, then the mid-price can only change when side of the order-book is depleted.

- Order imbalance at the top level:

$$OI = \frac{bq0}{bq0 + aq0} \tag{5}$$

This intuition behind this metric is that if the order queue on the bid side larger than on the ask side, the ask side will be depleted sooner and therefore its predictive of an upward price move. This ratio is chosen over a simple subration, is because its normalized, which reduces bias in the model fitting. Note that if the top level bid size queue is larger than the one on the ask side, the ratio will be close to 1. And if the ask side has a much larger queue the ratio will be cloe to 0.

- Order imbalance at the second level: same as the above metric, but at the second level of the order book

- Change in $bq0$ relative to the previous period:

$$bq0_t - bq0_{t-\delta} \tag{6}$$

The intuition behind this is that the arrival of a large order on the bid side indicates more buying pressure and vice versa when a large buy order is removed by a markeatable action. Both cases make it likely for the mid price to move.

- Change in $aq0$ relative to the previous period:

$$aq0_t - aq0_{t-\delta} \tag{7}$$

The intuition behind this is the same as the one for the bid quantity increasing.

- Finally the inverse mid price will also be considered as defined in equation 3

# 3  Model Selection

The model that will be estimated using a Lasso regression:

$$V(t + \delta) = Spread_t + OI_t + \epsilon \tag{8}$$

# 4  Results