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:

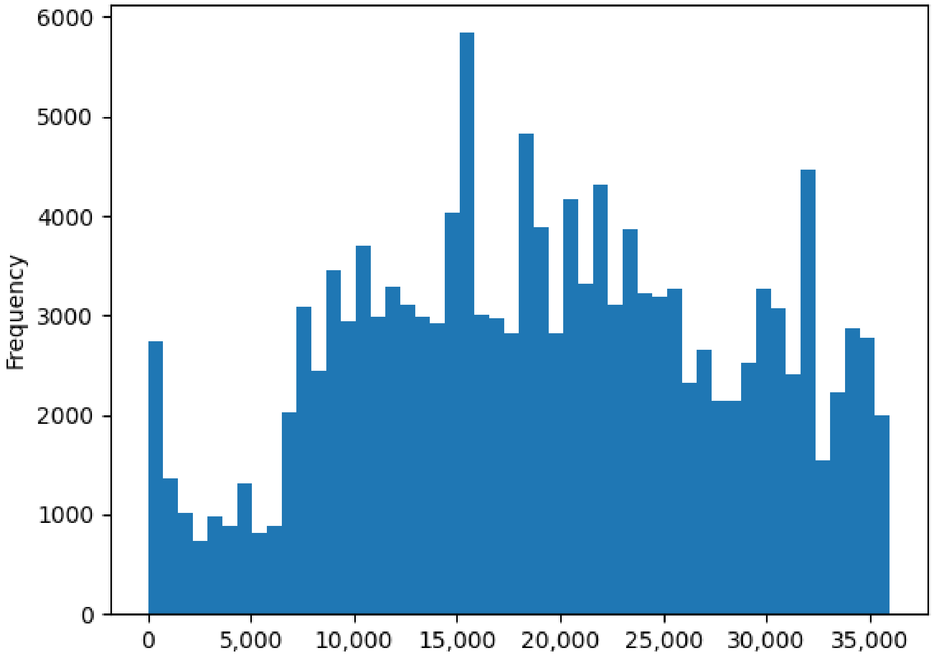

Figure 1: Histogram of order update arrivals for date 20190610

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:

$$P_{mid} = \frac{bp0 + ap0}{2} \tag{1}$$

- The inverse volume weighted mid price:

$$P^1_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0}{bq0 + aq0} \tag{2}$$

  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:

$$P^2_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0 + bp1 \times aq1 + ap1 \times bq1}{bq0 + aq0 + bq1 + aq1} \tag{3}$$

  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 δ 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.

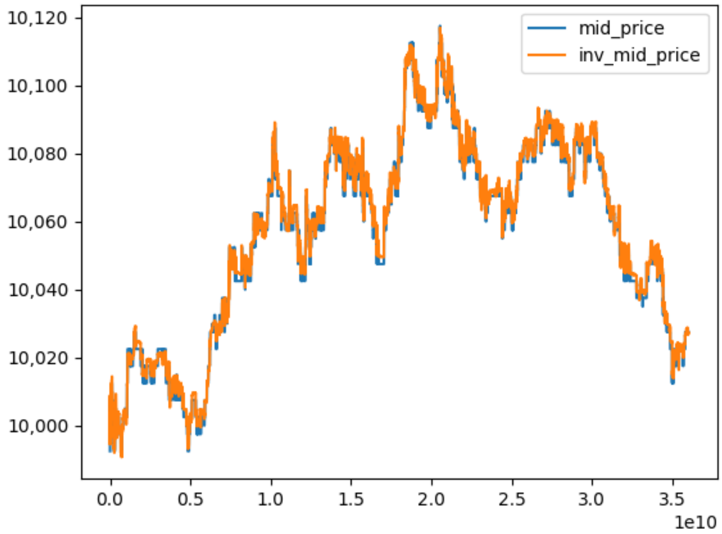

Figure 2: Day 2019-06-10

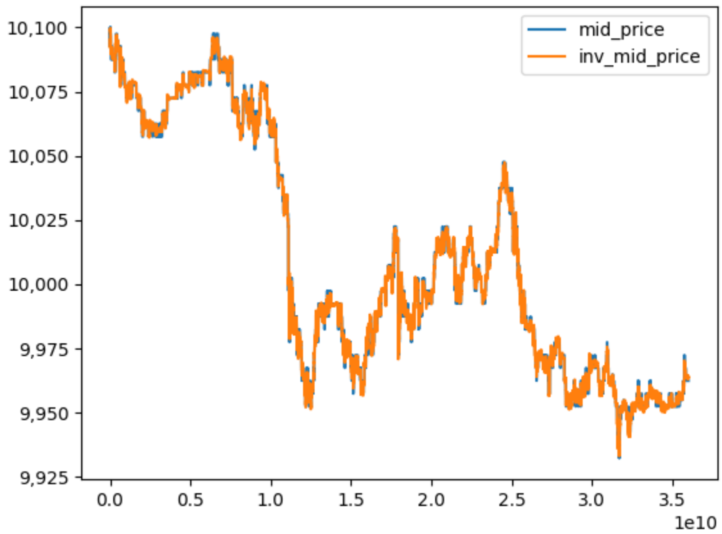

Figure 3: Day 2019-06-12

2.2 Features

For the features, the ones below are considerd:

- The bid-offer spread calculated as:

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

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:

$$OI_0 = \frac{bq0}{bq0 + aq0} \tag{5}$$

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:

$$OI_1 = \frac{bq1}{bq1 + aq1} \tag{6}$$

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

$$\Delta bq0 = bq0_t - bq0_{t-\delta} \tag{7}$$

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:

$$\Delta aq0 = aq0_t - aq0_{t-\delta} \tag{8}$$

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:

$$P_{mid,t+\delta} = BOspread_t + OI_{0,t} + OI_{1,t} + \Delta bq_{0,t} + \Delta aq_{0,t} + P^2_{mid,t} + \beta_0 + \epsilon \tag{9}$$

where ϵ is the error term and β_0 is the intercept. For ease of notation, the coefficients are omitted in the formula above. Note that the inverse weighted mid price of two layers $P^2_{mid,t}$ will be heavily correlated with P_mid. Hence, the regression will be run twice: with and without the inverse weighted mid price.

4 Results

The model above is estimated using a Lasso regression. We try out different alphas and report the R^2 score of both the test and train set. The results are given below:

date	test score	train score
20190610	0.999716	0.999710
20190611	0.999766	0.999767
20190612	0.999935	0.999936
20190613	0.999639	0.999640
20190614	0.999900	0.999900

Table 1: Model with $\alpha = 0.04$

date	test score	train score
20190610	0.996999	0.997010
20190611	0.995976	0.995977
20190612	0.998965	0.998966
20190613	0.993737	0.993749
20190614	0.998459	0.998463

Table 2: Model with $\alpha = 1$

date	test score	train score
20190610	0.873826	0.873907
20190611	0.815287	0.815320
20190612	0.952993	0.952984
20190613	0.771241	0.771078
20190614	0.954335	0.954352

Table 3: Model with $\alpha = 10$

Note that as the alpha is increased, the score overall go down which is the effect of the increasing alpha.

The high scores are mostly due to the inclusion of $P^2_{mid,t}$. In the following tables, that the term is omitted:

date	test score	train score
20190610	0.005131	0.005719
20190611	0.015000	0.014752
20190612	0.003620	0.003753
20190613	0.074478	0.078335
20190614	0.186116	0.187572

Table 4: Model with $\alpha = 0.04$

date	test score	train score
20190610	0.003133	0.003457
20190611	0.011403	0.011313
20190612	0.002394	0.002521
20190613	0.068101	0.071400
20190614	0.184384	0.185805

Table 5: Model with $\alpha = 1$

date	test score	train score
20190610	0.000000	-0.000000
20190611	0.000000	-0.000001
20190612	0.000000	-0.000002
20190613	0.000000	-0.000012
20190614	0.109777	0.110458

Table 6: Model with $\alpha = 10$

The coefficients are:

feature	coefficient
$BOspread$	0.237939
OI_0	2.080445
OI_1	-0.701317
Δbq_0	-0.092875
Δaq_0	0.117881

Table 7: $\alpha = 0.04$ and 2019-06-10

feature	coefficient
$BOspread$	-1.000401
OI_0	-0.785708
OI_1	-2.308759
Δbq_0	0.000000
Δaq_0	0.002927

Table 8: $\alpha = 0.04$ and 2019-06-12

feature	coefficient
$BOspread$	4.542120
OI_0	13.765068
OI_1	10.569343
Δbq_0	-0.987877
Δaq_0	0.441522

Table 9: $\alpha = 0.04$ and 2019-06-14