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:

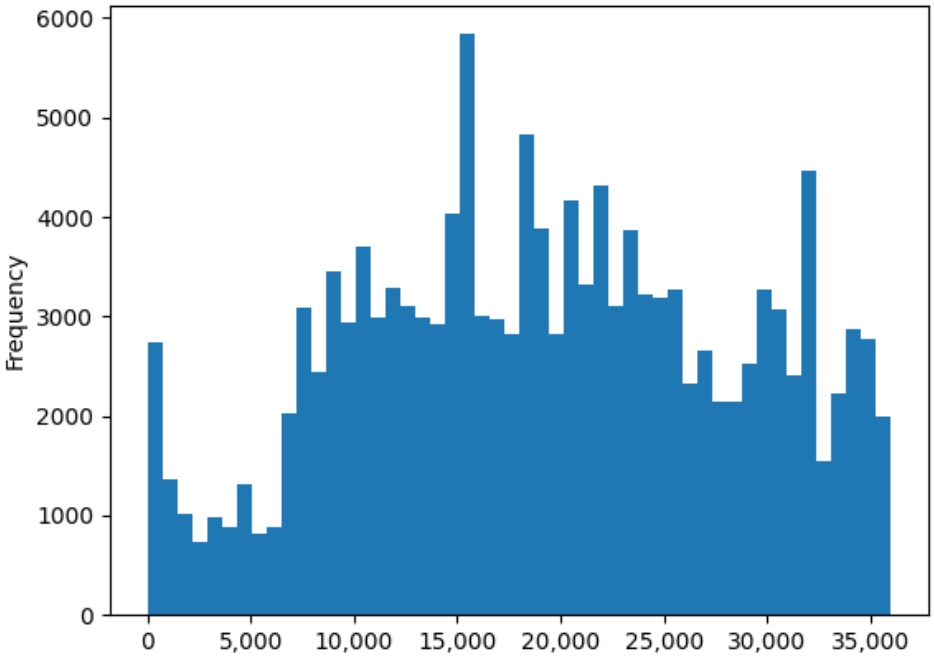

Figure 1: Histogram of order update arrivals for date 20190610

For the analysis the data will be resampled. Specifically, a grid will be used with a time interval of 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 less and a more thorough analysis can be done to optimize, but for practical considerations 100ms will be used as the discretization step.

Thus timestamps will be rounded up to the nearest 10,000 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:

$$P_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0}{bq0 + aq0} \tag{2}$$

  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.

For the target the simple average mid (equation 1) is used rather than the inverse weighted mid. The reasoning being that the inverse weighted mid is a predictor of sorts for the simple average mid. Specifically, the target will be the mid price change time t and $t + \delta$:

$$V(t) = P_{mid,t+\delta} - P_{mid,t} \tag{3}$$

where the δ is set to 100ms. The logic of using the price difference being that for trading predicting the price moves is more useful than the absolute price level.

A plot of the data and some comments:

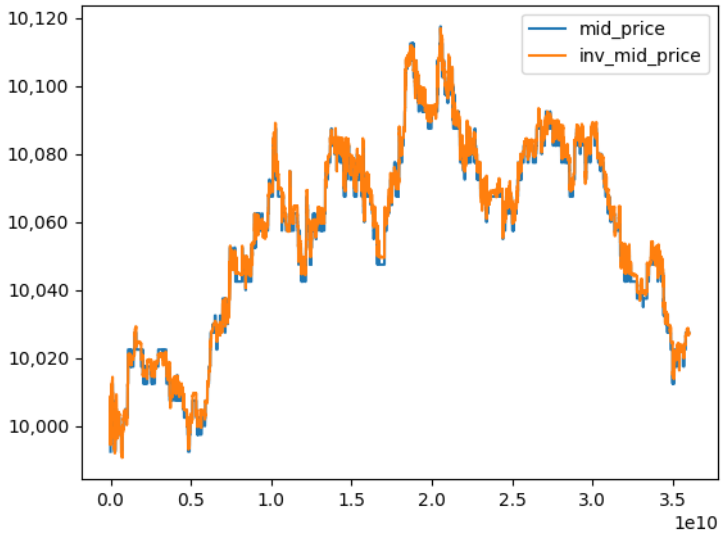

Figure 2

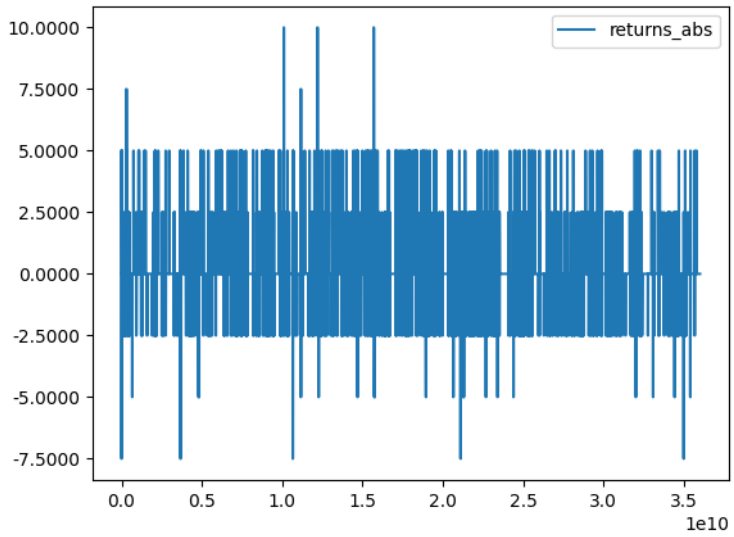

Figure 3

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:

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

  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.

3 Model Selection

The model that will be estimated using a Lasso regression:

$$V(t+\delta) = Spread_t + OI_t + \epsilon \tag{6}$$

4 Results