Order Book Analysis

Artagan Malsagov

January 16, 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:

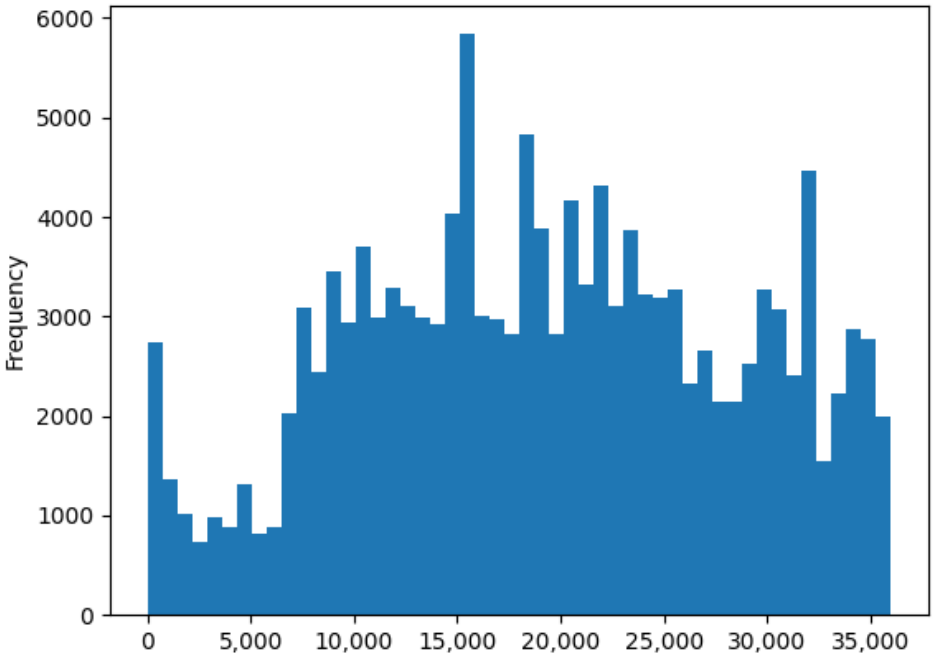

Figure 1: Histogram of order update arrivals for date 20190610

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 considering table **??**, the 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.

2 Feature and Target Selection

For the features and targets the following things are considered:

- The simple average mid price:

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

- The inverse volume weighted mid price:

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

First a simple plot of the data. The simple mid price is considered:

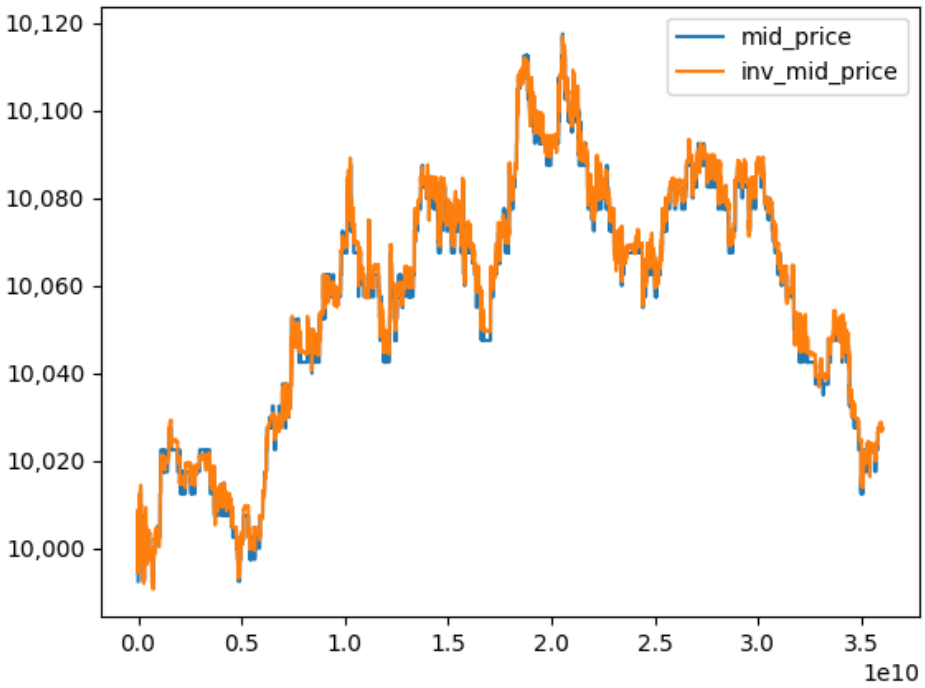

Figure 2

3 Model Selection

4 Results