Cleaned Signals Data Fields Details and Cleaning

Fields

Bids	Bids will be as string when imported from csv. If need the price and volume at each level for input to models, e.g. NNs, use regex to extract.
Asks	Same as above…
b	The price of the best bid.
a	The price of the best ask.
s	The spread, aka the difference between the best ask and the best bid (b subtracted from a). This is a good reflection of the liquidity in the market for the asset at any given timestamp. If it is small, there is good liquidity. If it is large, the market is less liquid.
m	The mid-price, aka the mid-point between the best bid and best ask (b + ((a-b)/2)).
vol_b	The volume at the best bid (b).
vol_a	The volume at the best ask (a).
dW	IGNORE: Used to work out OFI.
dV	IGNORE: Used to work out OFI.
e	IGNORE: Used to work out OFI.
all_vol_b	The total volume across all levels on the bid side of the LOB.
all_vol_a	The total volume across all levels on the ask side of the LOB.
all_vol_diff	all_vol_b – all_vol_a. This is a measure that looks at the difference in total volumes between the bid and ask sides. If positive, it can be implied that there might be more buy pressure (price likely to increase). If negative, it can be implied that there might be more sell pressure (price likely to decrease).
OBI	Order Book Imbalance. (vol_b – vol_a) / (vol_a + vol_b). This feature looks at the relative imbalance between the volume at the best bid and best ask as a proportion of the total volume at both. This gives indications on whether the best bid or ask are likely to move up or down in price.
OFI_{interval}	Order Flow Imbalance. A measure of the net flow of orders (and cancellations) into the order book at the best bid and ask (b and a). If positive, there has been a greater flow of orders that may increase the price over the given interval. If negative, the opposite is true.
signal_{interval}	TARGET VARIABLE: 0: Sell signal – If a trader placed a sell market order, and therefore sold at the current best bid price, at the current timestamp and then bought back the asset at the nearest timestamp to the given {interval}, then the trader would make a profit of >=2 (this acts as a margin of safety). 1: Neutral – The trader would not make a profit of >=2 if they bought or sold at the current timestamp and did the reverse after the given {interval}. Therefore, do nothing. 2: Buy signal – If a trader placed a buy market order, and therefore bought at the current best ask price, at the current timestamp and then sold the asset at the nearest timestamp to the given {interval}, then the trader would make a profit of >=2 (this acts as a margin of safety).

Recommended Cleaning:

See ‘decision_tree_notebook.ipynb’ stored in mw_analysis on the github for an example of importing the data and cleaning each dataframe…

1. Make the ‘Timestamps’ column a timedelta and make it the index.
2. Get rid of the first 500 and last 500 timestamps (approx. 1m of the data at the start and end of the day) as this is unreflective of normal operations and some of the features may be dysfunctional for these timestamps.

Model Considerations

Things to think about, test and note down when running models:

1. Explainability – can you explain why the model made the decisions it did
2. Overfitting – how overfit is the model and can this be reduced.
3. How does/would the model deal with anomalies (e.g. financial crash etc.)
4. Accuracy, recall, precision, F1, confusion matrix, etc.
5. Training time? How long does it take to train?
6. Profit performance when used with an algorithm and what does the algorithm need to take into account?
7. Variability in profit performance
8. Amount of training data required