Where will the next trade take place?

Given recent trades and order books from a set of trading venues, predict on which trading venue the next trade will be executed

Quentin Jacob

Challenge Data 2020

April 5, 2022

Overview

1. **Feature engineering**
2. **Model selection**
3. **Data augmentation**

Feature engineering

Initial remarks

- Hybrid data: trades like time series, LOB snapshots with specific structure (normalization across venues, timestamp of last update) $\rightarrow$ need for feature engineering
- Benchmark: next venue = trading venue of the most recent trade
- Intuition: the LOB of one venue taken in isolation does not give much information, comparison between venues is key
- Handcraft features representing:
  - attractiveness of one venue VS the other (w.r.t. price and quantity), current level of liquidity fragmentation (full size only available on one venue? evenly spread?)
  - intensity, aggressiveness and sign imbalance of the order flow
  - historical properties of each stock, properties of the current day, of the current time window (not causal)
- Gradient boosting trees on tabular data is a good way to start

Is this venue showing a competitive price?

Feature	Dimension
Rank of venue first limit prices (vs 5 other)	$\|\text{venues}\| \times \|\text{sides}\|$
Rank of venue second limit prices (vs 5 other)	$\|\text{venues}\| \times \|\text{sides}\|$
Number of venues showing best price	$\|\text{sides}\|$
Redundancy of first limit prices (entropy-like)	$\|\text{sides}\|$
Standard deviation of first limit prices	$\|\text{sides}\|$
Binary variable representing a crossed LOB	$\|\text{venues}\|$
Ranking of spreads	$\|\text{venues}\|$
Redundancy of spreads	1
Spreads / tightest	1
Target encod. of stocks (hist. proportion of trades on each venue)	6 per stock

Is this venue showing competitive sizes?

Feature	**Dimension**
LOB imbalance of each venue	$\lvert\text{venues}\rvert$
Rank of venue first limit sizes (vs 5 other)	$\lvert\text{venues}\rvert \times \lvert\text{sides}\rvert$
Rank of venue second limit sizes (vs 5 other)	$\lvert\text{venues}\rvert \times \lvert\text{sides}\rvert$
Measure of fragmentation of aggregated LOB per side	$\lvert\text{sides}\rvert$
Redundancy of first limit sizes (entropy-like)	$\lvert\text{sides}\rvert$
Total size available at best price	$\lvert\text{sides}\rvert$
Size of first limit $\times \mathbb{1}_{\text{venue shows best price}}$	$\lvert\text{venues}\rvert \times \lvert\text{sides}\rvert$
Ranking of last feature	$\lvert\text{venues}\rvert \times \lvert\text{sides}\rvert$
Redundancy of last feature	$\lvert\text{sides}\rvert$

Temporal features from last trades

Feature	Dimension
Time elapsed between first and last trade	1
Time elapsed between 5th trade and last trade	1
Mean time elapsed, for this (stock, date) between first and last trade	1 per stock $\times$ date
Mean inter-trade time for this stock	1 per stock
Mean inter-trade time for this day, all stocks included	1 per date
Mean trade time in the window, size weighted	1
Mean trade time in the window	1

Price and size features from last trades

Feature	Dimension
mean, std, min, max prices of last trades	$1 \times Nb_{metrics}$
mean trade price / bast ask price	1
mean trade price / best bid price	1
mean absolute value of differences between consecutive trade prices	1
total traded size	1
largest trade	1
std of trade sizes	1
total traded size per second	1
total signed traded size	1
total signed traded size per second	1
total traded size per side	$\lvert\text{sides}\rvert$

Locations of last trades and last updates

Feature	Dimension
$\mathbb{1}_{\text{last two trades on same venue}}$	1
Number of consecutive trades on last traded venue	1
Number of trades with same timestamp as last trade	1
Number of consecutive trades on same side as last trade	1
Ranking of timestamp updates	$\lvert\text{venues}\rvert$
For each venue, number of trades and size among last ten	$2 \times \lvert\text{venues}\rvert$
For each venue, number of trades and size among last ten as fract. of max	$2 \times \lvert\text{venues}\rvert$
Mean of last features for (stock, date)	
Mean of last features for (stock, date, 5-min window)	

Features from the very last trade

Feature	Dimension
Time elapsed between last trade and last LOB update	1
Price of last trade	1
Size of last trade	1
Price of last trade divided by best price	$\lvert\text{sides}\rvert$
Price of last trade divided by mean first limit price	$\lvert\text{sides}\rvert$
Size of last trade divided by total size at best	$\lvert\text{sides}\rvert$

Model selection

Cross-validation scheme

- 4-fold cross-validation on dates
- Stability between folds
- Consistency between improvements on cross-validation and leader-board

Model

- Not much time spent on hyper-parameter tuning
- LightGBM DART model with multi-class one-vs-all objective function

Data augmentation

Exploit symmetries

- Intuition: in an imaginary world where the order-flow is reversed (buy orders become sell orders and vice versa) as well as order books (ask becomes bid, bid becomes ask), one could expect the next trade venue to be the same as in the true world
- Create a mirrored version of the true dataset: if symmetry is not in the model or directly in the features, enforce it in the data
- Small but significant improvement in accuracy (10^{-3}) on the leader-board after training on this extended dataset
- Score summary: Benchmark (0.3621, 0.3564), final model (0.5062, 0.4925)

Thank you!