Market Microstructure in the Big-data Era: Improving HF Price Prediction

5/19/23 2:22 PM

Agostino Capponi (Columbia University)

Abstract:
Traditional empirical market microstructure models use linear models with a small set of features (e.g, trade imbalance, best bid, and ask price) from a single market to study the price discovery process. However, in the big data era, high-frequency full limit-order-book (LOB) data are available from multiple markets, allowing for a much richer set of features. We construct machine learning models such as boosted decision trees and random forests, which endogenously pick the most relevant features from a large universe of LOB variables. We demonstrate that these models outperform classical linear models in high-frequency price prediction and information attribution. We highlight how the gains achieved by these models can be explained by the nonlinear price impact of LOB features, which is missing in traditional models.

Introduction

Microstructure in the machine age

- Big data in the US equities market
  - Extremely fast: algorithmic and high-frequency trading, 20% of trades arrive in < 1ms clusters
  - A highly fragmented market: 16 public exchanges, internalization, dark pools
- The information set of market participants has greatly expanded
- Market data is crucial, for market makers, arbitrageurs & buy-side

Market data policy debate

- Two-tiered market data for US equities:
  - Consolidated (SIP) feeds: slow, top-of-book quotes, SEC-mandated, relatively cheap, used by unsophisticated traders
    (SIP: Securities Information Processer https://polygon.io/blog/understanding-the-sips/)
  - Direct feeds: fast, depth-of-book quotes, sold by exchanges, expensive, used by sophisticated traders such as high-frequency traders
- Fair? Policy debates
  - US: NMS (National Market System) 1.0 (top of book in the SIP) => NMS 2.0 (five-level depth in the SIP)
  - Europe: consolidated feed in the making: what to include? Top-of-book? Depth-of-book?
- Economic questions:
  - Which exchange contributes the most to price discovery?
  - Which part of the data feed contributes the most to price discovery? (Top? When the five best levels? Full depth?)

Contributions

- Empirical microstructure literature
  - Limited set of variables/features
  - Ex-ante specification of price impact function
  - In-sample, ex-post attribution of information shares
- Quantitative finance literature
  - State-of-the-art machine learning models
  - Economics is not clear
- Goal of this paper: bridge the above two strands of literature

Data

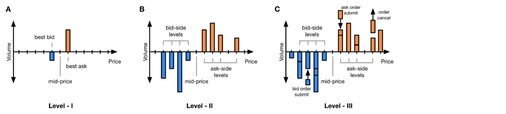
- "Direct feeds" from public exchanges
  - Level 3 order-book messages: all add (new limit orders), cancel/modification of existing orders, and trade messages
  - Time stamped to microsecond precision
- 30 constituent stocks of DJI, 54 trading days spanning from the year of 2017 to 2021
- For each exchange, we build the entire order book based on the direct feed messages

Limit order book (LOB) market

- Most liquid markets use limit order books for trading
- A limit order book is essentially a collection of unexecuted quotes
  - Each quote specifies the price and quantity the trader is willing to trade
  - New quotes can be continuously added and existing quotes can be canceled, modified, or executed against incoming marketable orders

LOB data types

- Level-I: the best bid/ask prices and volumes,
- Level-II: price and aggregated volume across a certain number of price levels
- Level-III: non-aggregated orders placed by market participants



Direct Feeds

(a) Add message

datetime	mtype	micros	seq	delta	source	symbol	oid	size	price	side	flags
2014-07-15 14:01:00.465239	add	50460465239	2542087 11	60	INET	AAPL	193553906	59	95.46	S	-

(b) Cancel/modification message

datetime	mtype	micros	seq	delta	source	symbol	oid	size
2014-07-17 13:01:00.976771	mod	46860976771	210279229	43	INET	MSFT	162545830	0

(c) Trade message

datetime	mtype	micros	seq	delta	source	symbol	oid	size	price	side	flags
2014-08-12 14:02:35.414432	trd	50055414432	219933620	102	INET	YHOO	167243090	100	35.290	B	-

(Additional info: https://www.investopedia.com/terms/l/level3.asp)

Methodology

LOB actions

- LOB is constantly changing due to add, modifications, and executions

Feature Engineering

- LOB actions and their lagged values, from each exchange
- (BBO: Best-Bid-offer)
  - **Trade-BBO-Changing**: Executions moving BBO
  - **Trade-NonBBO-Changing**: Execution not moving BBO
  - **Add-BBO-Improving**: Add orders improving BBO
  - **Cancel-BBO-Worsening**: Cancel orders worsening BBO
  - **Add-at-BBO**: Add orders adding depth at the current BBO
  - **Cancel-at-BBO**: Cancel orders removing depth at the current BBO
  - **Add-<=5lvlBBO**: Add orders adding depth <= 5 levels from BBO
  - **Cancel-<=5lvlBBO**: Cancel orders removing depth <= 5 levels from BBO
  - **Add->5lvl-BBO**: Add orders adding depth > 5 levels from BBO
  - **Cancel->5lvl-BBO**: Cancel orders removing depth > 5 levels from BBO
- Midquote changes, from each exchange

Target and performance evaluation

- Target: Short-term NBBO (National Best Bid Offer) change (e.g. next 5 events)
- Evaluation:
  - Mean Squared Error
  - R square

Models

- OLS
- Linear with penalties
- Tree-based models
  - RF
  - Boosted regression trees

Training, validation and testing sample split

- We split each trading day into 13 half-an-hour intervals
- Training, validation and testing based on inter-day rolling windows, for example -
  - 9:30-10:00 as training
  - 9:30-10:00 as validation
  - 9:30-10:00 the day after as testing
- We follow hyperparameters in Gu, Kelly and Xiu (2020)

Results

Prediction (MSE)

- Consider the five stocks: American Express (AXP), Boeing (BA), Caterpillar (CAT), Disney (DIS), and Goldman Sachs (GS).
- Boosted Regression Tree consistently outperforms other models

ticker model	AXP	BA	CAT	DIS	GS
OLS	0.0149	0.1267	0.0407	0.0213	0.3925
Elastic-net	0.0149	0.1206	0.0368	0.0213	0.3808
RF	0.0151	0.1214	0.0363	0.0218	0.3727
BRT	**0.0145**	**0.1158**	**0.0343**	**0.0204**	**0.3593**

Prediction (R Squared)

- Boosted Regression Tree consistently outperforms other models

Permutation importance

- To access the importance of a feature or several features, permutate (randomly shuffle the ordering) them in the testing set
- Then compare the change in MSE or R2 from the testing set
- Different from in-sample feature importance
- Agnostic to model choice

Permutation importance (exchange)

- Which exchange contributes the most to price discovery?
- Conclusion: larger exchanges are more important, but the drop in R2 is mild

Permutation importance (data feeds)

- Which part of data feed (beyond the best five levels, or within the best five levels) contributes the most to price discovery?
- Data feeds beyond the five best levels have limited information, within five levels much more important

Conclusion

- ML have consistently better prediction performance for LOB misquote changes than linear models
- From an economic perspective:
  - Larger exactness are more important, but R2 drop is mild
  - Data feeds beyond five levels have limited info
- Future extensions:
  - LOB events have time-series dynamics, e.g. autoregressive structure
  - Suitable for time-aware machine learning models:
    - LSTM
    - Transformers

SIP: Securities Information Processor

- The SIPs essentially link the U.S. markets by processing and consolidating all bid/ask quotes and trades from every exchange (i.e., core data) into one data feed
- Every broker-dealer in the US is required to report their best bids and offers to the securities exchanges. The securities exchanges consolidate this information from the broker-dealer and provide it to the SIPs to create the consolidated data. Every broker-dealer is then required to purchase the consolidated data from the SIPs to comply with their best execution obligations and compete in the market.