Market Microstructure in the Age of Machine Learning

Marcos López de Prado

Lawrence Berkeley National Laboratory
Computational Research Division

A brief History of Market Microstructure Models

- First generation: price sequences
  - The Roll model [1984]
  - High-Low Volatility Estimator: Beckers [1983], Parkinson [1980]
  - Corwin and Schultz [2012]
- Second generation: strategic trade models
  - Kyle's lambda [1985]
  - Amihud's lambda [2002]
  - Hasbrouck's lambda [2009]
- Third generation: sequential trade models
  - Probability of information-based trading (PIN): Easley et al. [1996]
  - Volume-synchronized probability of informed trading (VPIN): Easley et al. [2011]

Advantages of the AI Age

- Financial Big Data (tick/book level)
- New and advanced statistical techniques (Machine Learning)
- Unprecedented computational power (Supercomputers)

Microstructural relationships often are non-linear and hard to parameterize. When used properly, these technologies can uncover relationships unknown to traditional approaches.

Goals of this Presentation

- Investigate all three-generation market microstructure variables jointly on the most recent 10 year market data
- Apply ML based feature importance analysis and test the usefulness of microstructure variables at predicting various market movements
- Study how the feature importance pattern changes with different labels and different time scales

Section I
Data & Variables

Market Data

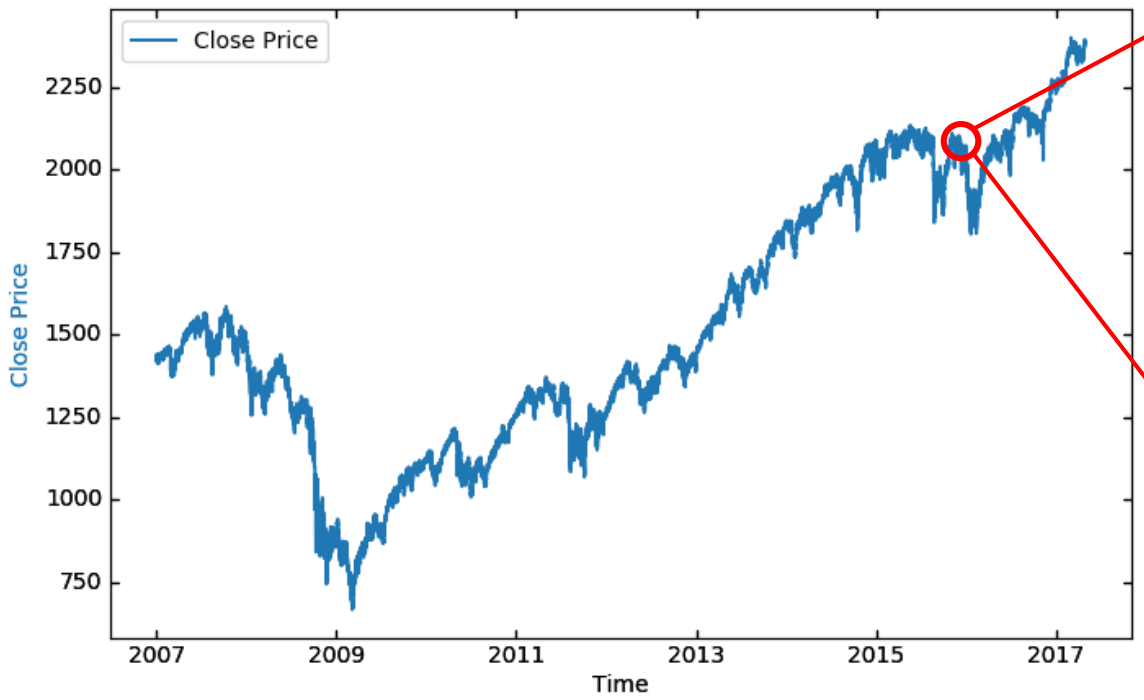
- 87 liquid futures traded globally (index, currency, commodity and fixed-income) with 10 years history
- Tick level trade data, aggregated into bars (price-volume bars). Each bar is formed with a timestamp t when

$$\sum_{\tau=t-1}^{t} p_\tau V_\tau \geq L$$

- The threshold L is chosen to have roughly 50 bars per day in the year 2017. Each bar contains Close, Open, High, Low and Volume

Market Data

A snapshot of ES1 Index data



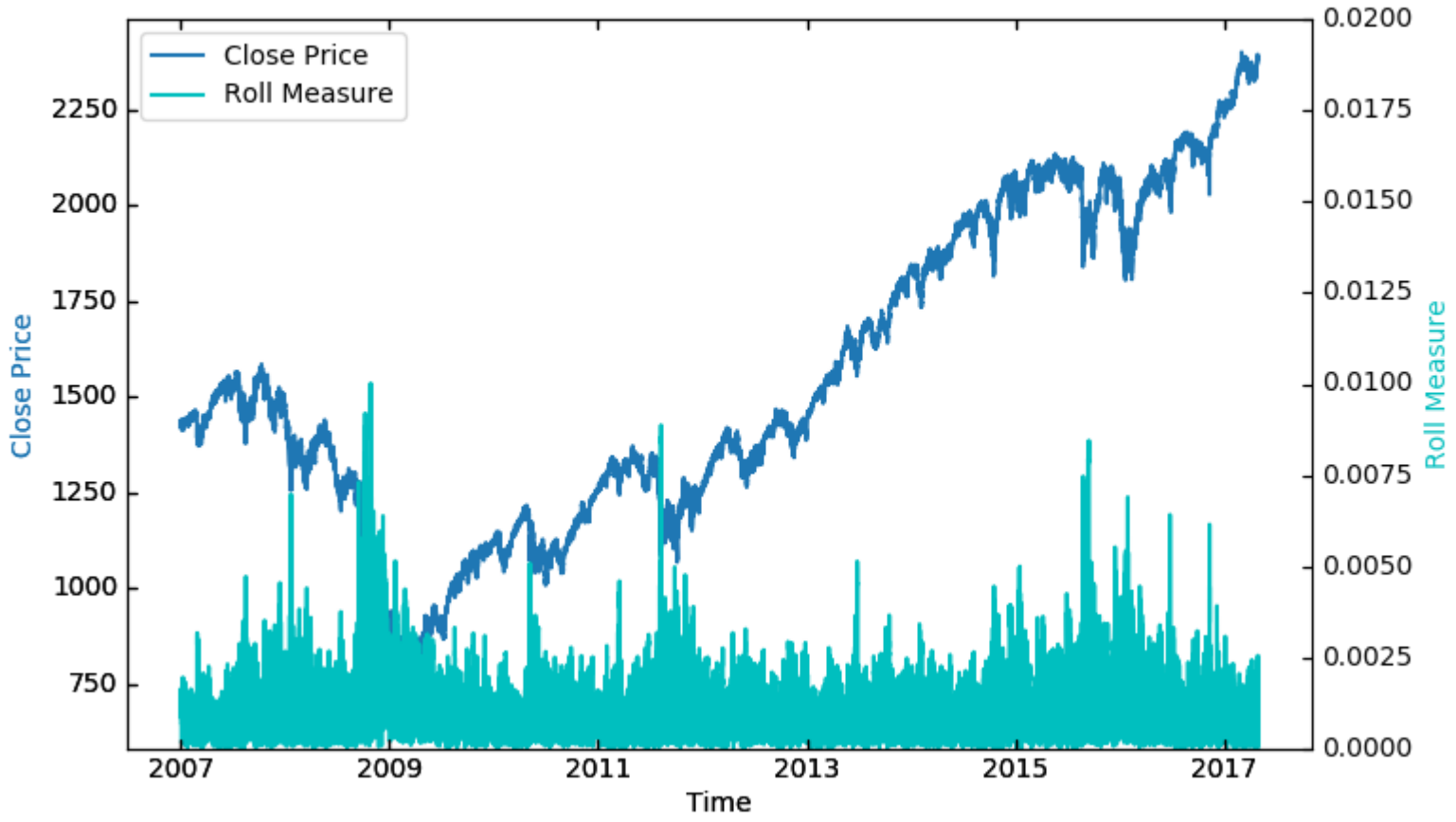
Timestamp	Open	Close	High	Low	Volume
5/23/16 3:14 AM	2052	2045	2053.75	2045	33111
5/23/16 4:10 AM	2045	2052.5	2053.75	2043.5	33033
5/23/16 5:31 AM	2052.5	2051.25	2056	2050.25	32916
5/23/16 7:07 AM	2051.25	2046.25	2052.25	2045	33020
5/23/16 8:43 AM	2046	2049.5	2051.75	2046	33026
5/23/16 9:30 AM	2049.5	2048	2049.75	2047.25	33061
5/23/16 9:35 AM	2048	2049.5	2051.25	2047.5	32996
5/23/16 9:40 AM	2049.5	2050.75	2051	2046.5	33006
5/23/16 9:48 AM	2050.75	2049.5	2051.25	2048	32984
5/23/16 9:54 AM	2049.5	2047	2050	2046.25	33021
5/23/16 10:00 AM	2047	2048.75	2049.75	2046.5	33016
5/23/16 10:06 AM	2048.75	2051.75	2052.25	2048.75	32964
5/23/16 10:14 AM	2051.75	2050	2051.75	2049.25	32972
5/23/16 10:22 AM	2050	2051.5	2052.75	2048.75	33006
5/23/16 10:32 AM	2051.5	2049.5	2052.25	2049.5	32966
5/23/16 10:44 AM	2049.5	2048.75	2051.5	2048.25	33057

Microstructure variables

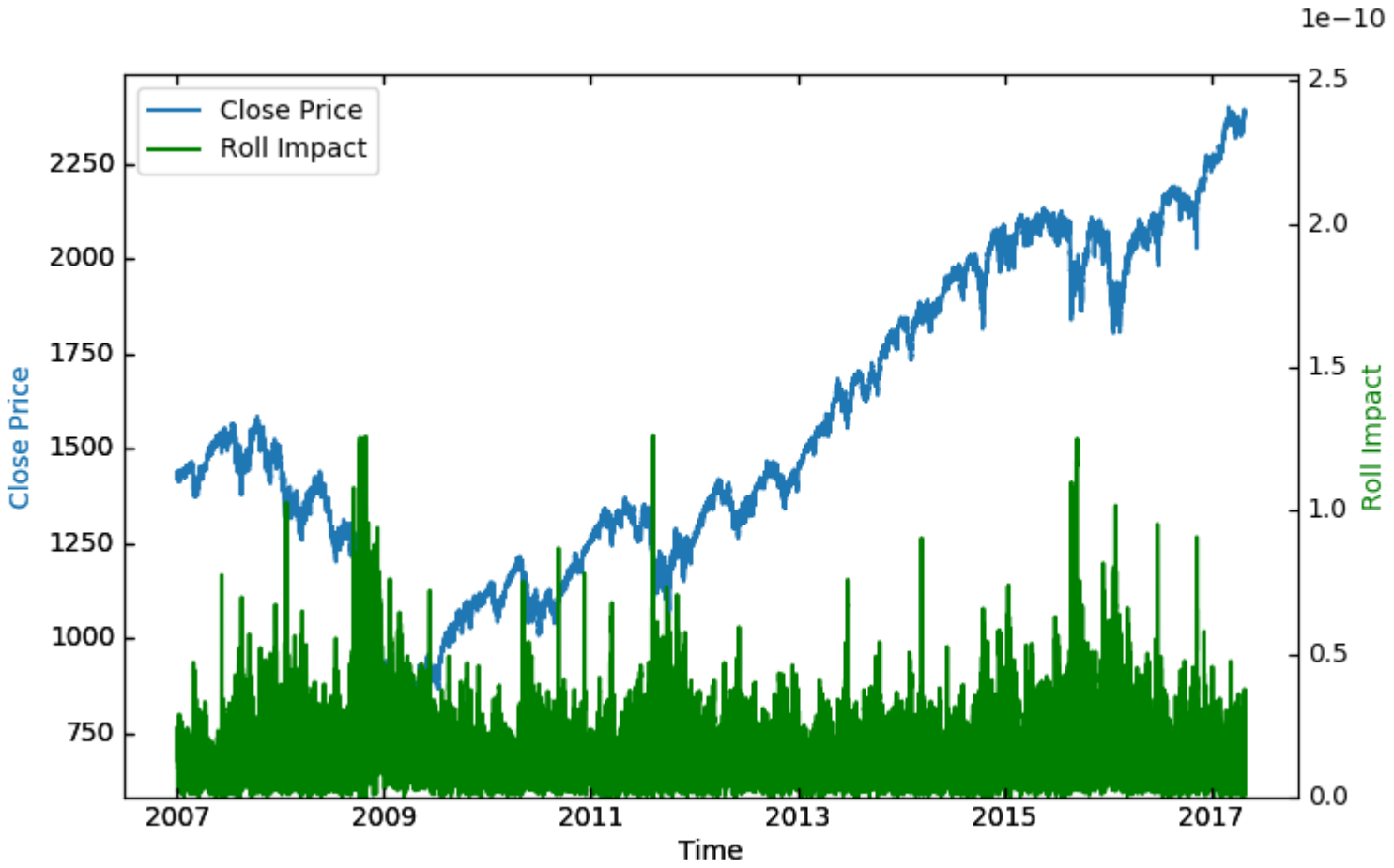
Roll measure

$$2\sqrt{|\mathrm{cov}[\Delta p_t\ , \Delta p_{t-1}]|}$$

- Δp_t is the change in close price between two bars at time t.
- The covariance is evaluated on a rolling basis for different window size (all plots are done with window = 50 bars)



Microstructure variables



Roll impact

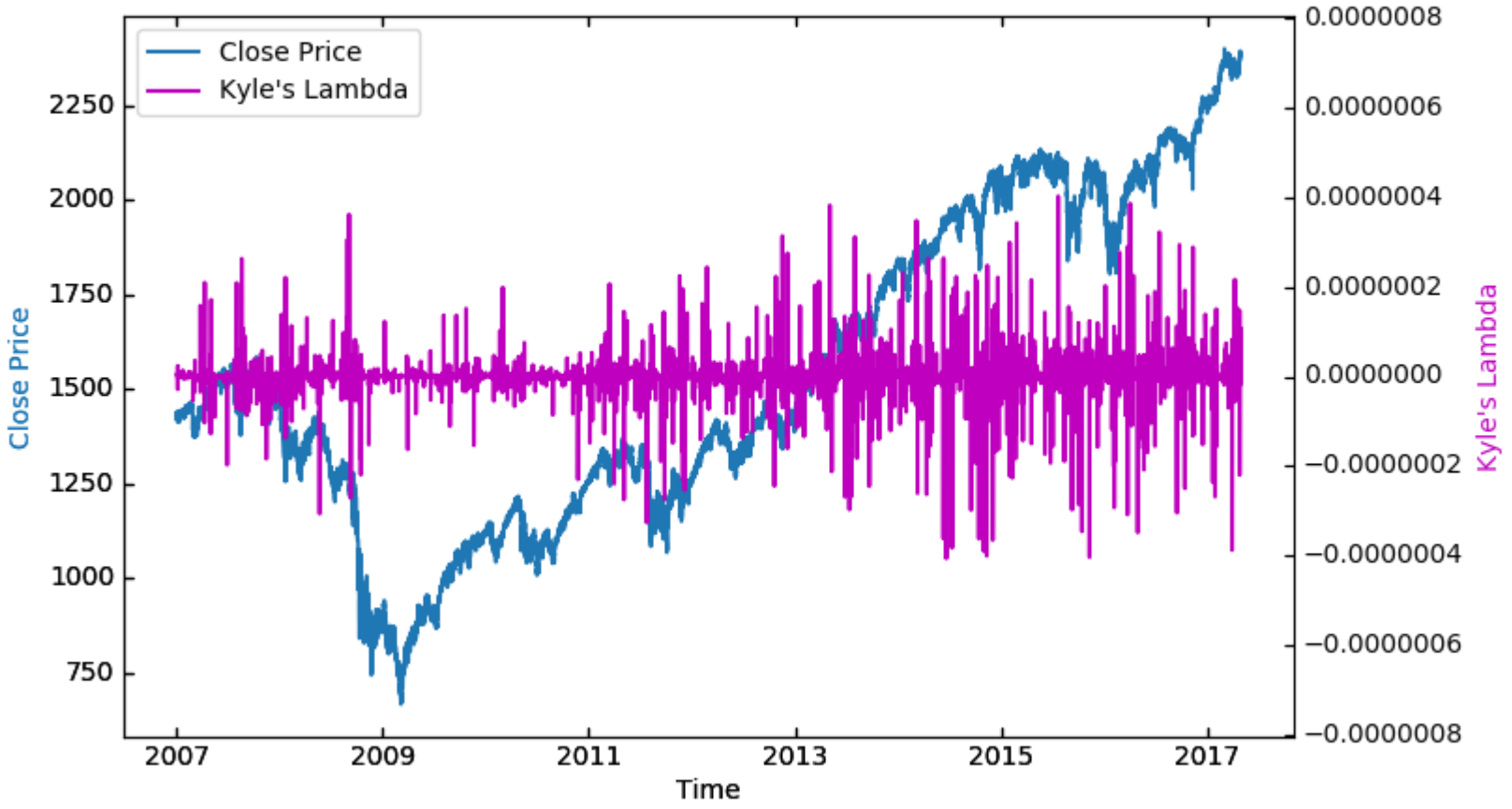
$$\frac{2\sqrt{|\mathrm{cov}[\Delta p_t, \Delta p_{t-1}]|}}{Dollar\ Volume_t}$$

Microstructure variables

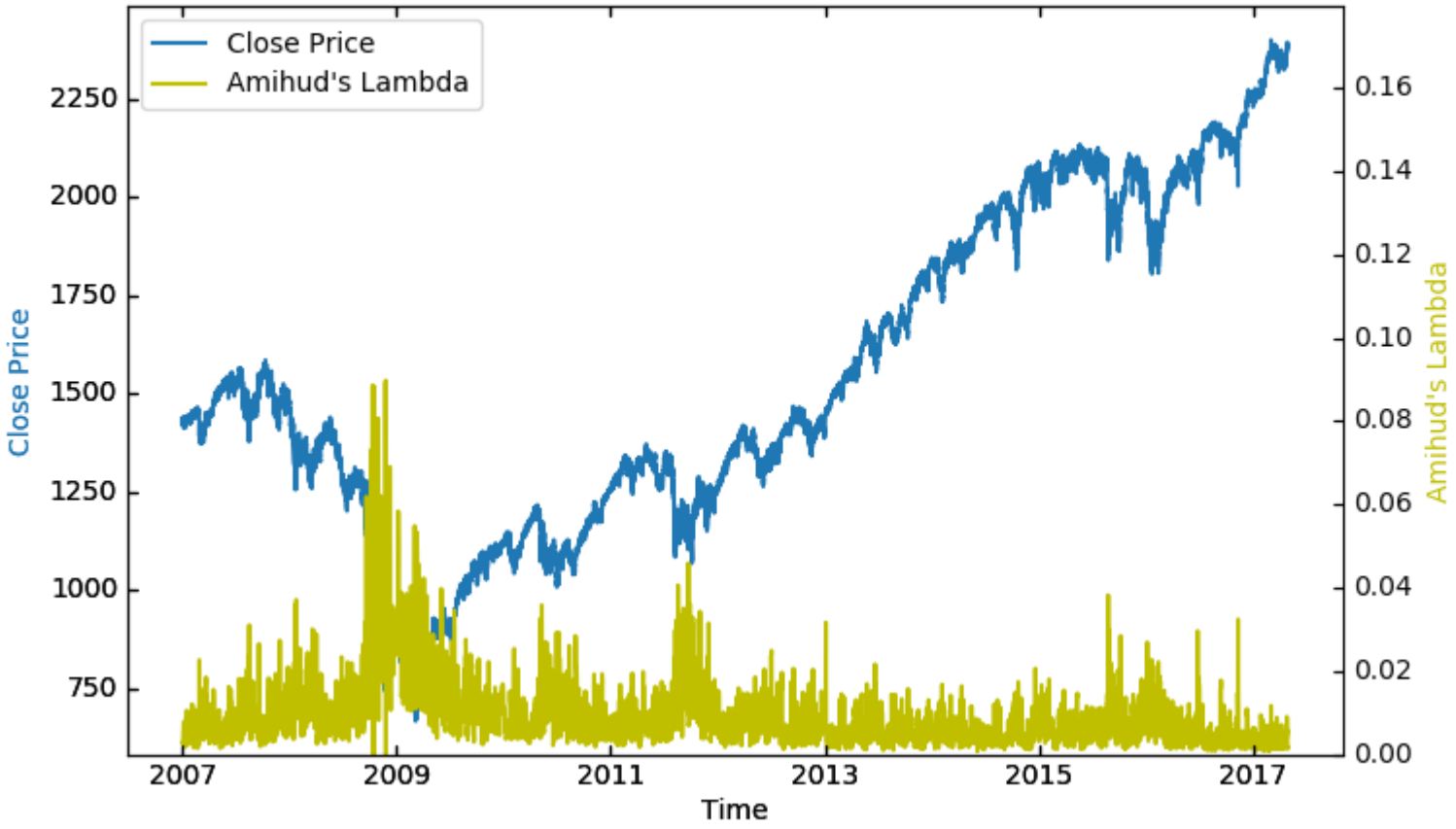
Kyle's λ

$$\Delta p_t = \lambda b_t V_t$$

- λ is derived from the regression with a rolling window
- V_t is the volume and $b_t = \text{sign}[p_t - p_{t-1}]$ which is computed on bar level.



Microstructure variables



Amihud measure

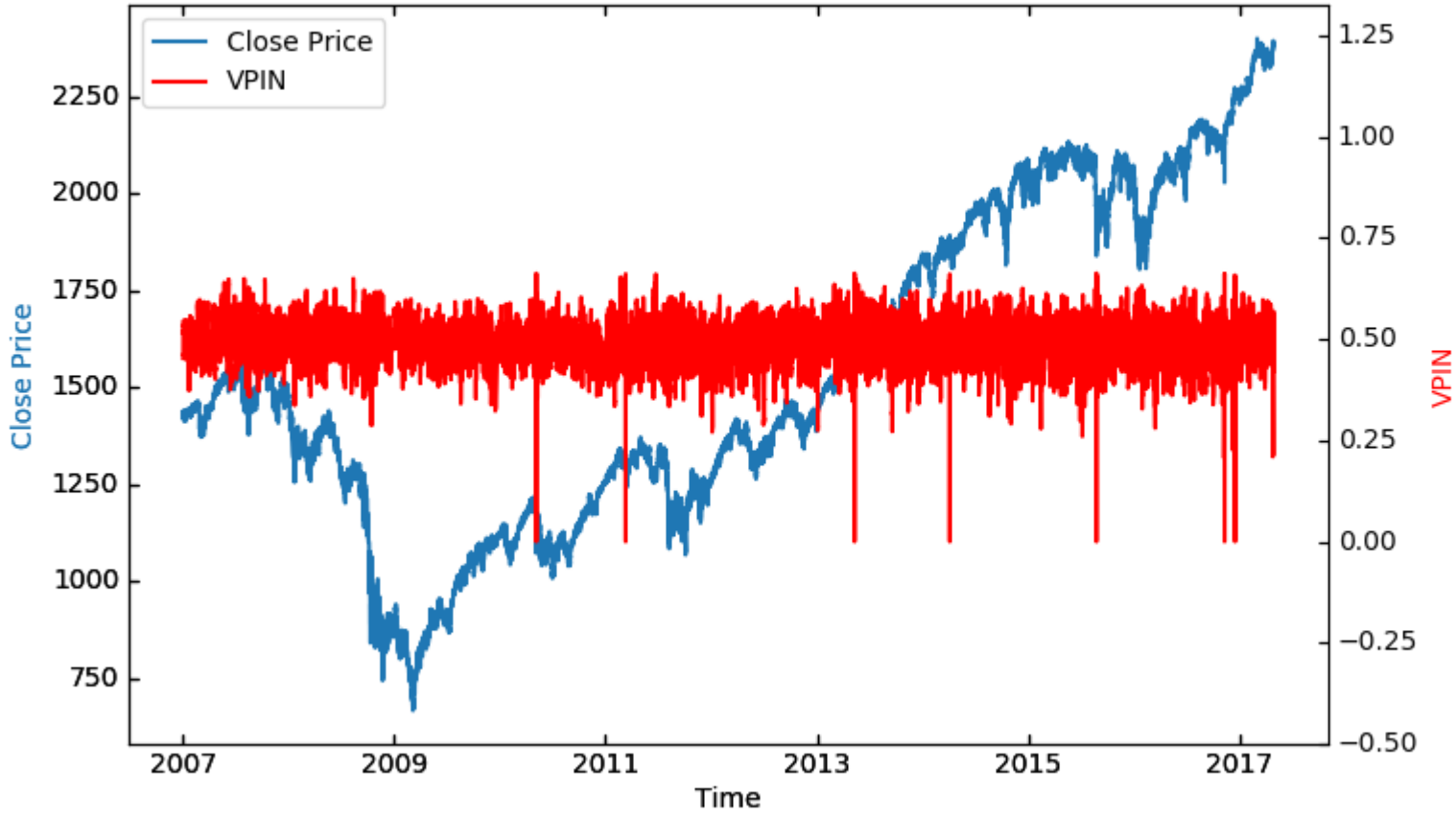
$$Moving\ Average\left[\frac{|Return_t|}{Dollar\ Volume_t}, window\right]$$

Microstructure variables

VPIN

$$\frac{\sum |V_t^S - V_t^B|}{nV}$$

- $V_t^B = V_t Z\left[\frac{\Delta p_t}{\sigma_{\Delta p_t}}\right], \quad V_t^S = V_t - V_t^B$
- n is the number of bars used



Section II
Algorithms & Research Tools

Selection of ML algorithm: Random Forest

- Random Forest is an ensemble method. The bootstrapping process brings in randomness that can reduce potential overfitting, a common issue in finance problems.

- As a tree-based algorithm, Random Forest has a tree-based feature importance method (Mean Decreased Impurity). We can compare it with a more generic ML feature importance method (Mean Decreased Accuracy).

ML-based feature importance analysis

- Mean-Decreased Impurity (MDI). MDI is built for all tree-based ML algorithms including random forest. All tree-based algorithms consist of multiple data splits on selected features, and each split is obtained by minimizing the impurity. MDI measures how much sample size weighted total impurity each feature reduces during training, and rank the feature with largest decreased impurity the highest.

- MDI feature importance is **in-sample** as it is extracted from training data only. It is a statement on explanatory importance rather than predictive importance.

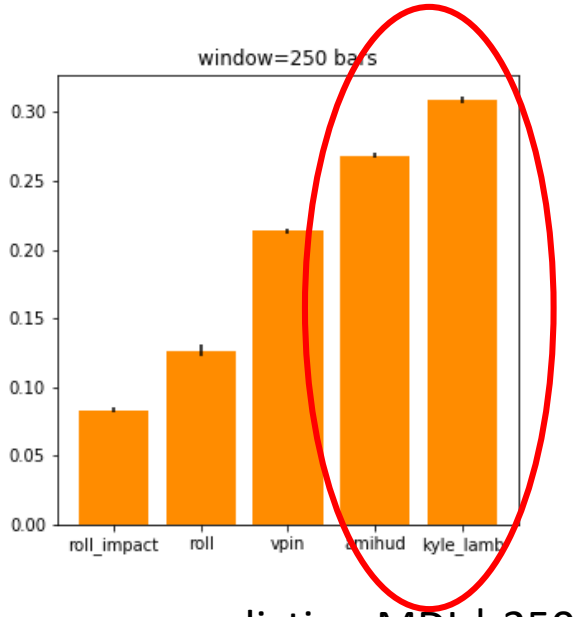
ML-based feature importance analysis

- Mean-Decreased Accuracy (MDA). MDA is a generic feature importance method that applies to all ML algorithms. The idea is to randomly permute the values of each feature and measure how much the permutation decreases the model's out-of-sample accuracy. The more the accuracy decreases indicates the feature is more important.

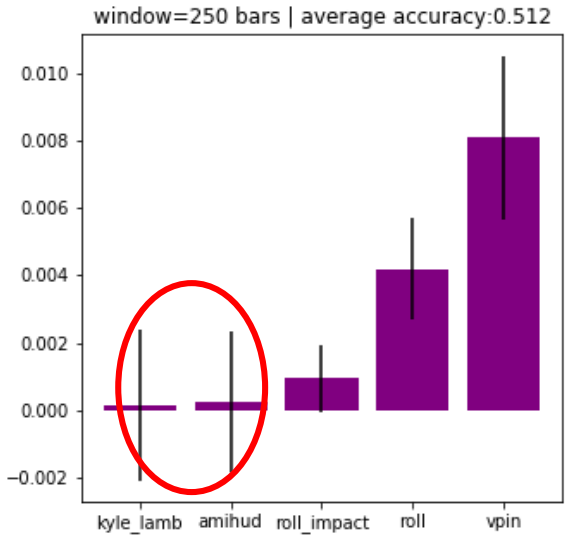
- In contrast to MDI, MDA tests the actual importance for the **out-of-sample** performance.

Analysis

- Feature importance rankings from MDI and MDA are generally different. Important features from MDA might contribute little or even negatively to out-of-sample prediction performance.



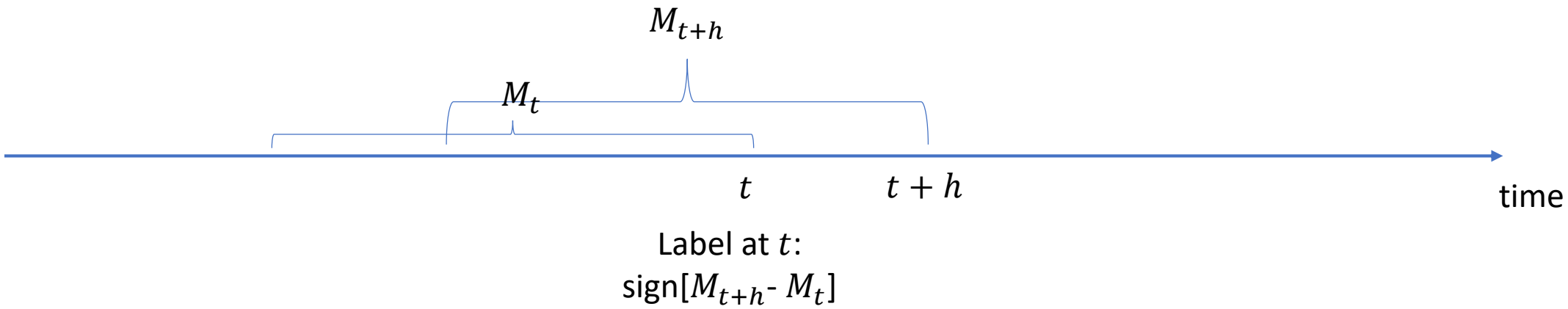
return skewness prediction MDI | 250 bars



return skewness prediction MDA | 250 bars

Prediction labels

Binary Label generation: at time t, we generate a binary label by computing a measure M_t that is constructed with a backward window and comparing it to its value at $t + h$, with $h = 250$ bars (around 5 trading days) for all labels hereafter in this study.



Prediction labels

Label 1. Sign of change in bid-ask spread estimator (Corwin-Schultz)

$$\text{sign}[S_{t+h} - S_t]$$

$$S_t = \frac{2(e^{\alpha_t}-1)}{1+e^{\alpha_t}}, \quad \alpha_t = \frac{\sqrt{2\beta_t}-\sqrt{\beta_t}}{3-2\sqrt{2}} - \sqrt{\frac{\gamma_t}{3-2\sqrt{2}}}, \quad \beta_t = E\left[\sum_{j=0}^{1}\left[log\left(\frac{H_{t-j}}{L_{t-j}}\right)\right]^2\right], \quad \gamma_t = \left[log\left(\frac{H_{t-1,t}}{L_{t-1,t}}\right)\right]^2$$

Label 2. Sign of change in realized volatility

$$\text{sign}[\sigma_{t+h} - \sigma_t]$$

σ_t is the standard deviation of realized return

Prediction labels

Label 3. Sign of change in Jarque-Bera statistics of realized returns

$$\text{sign}(JB[r_{t+h}] - JB[r_t]) \,,$$

$JB[r] = \frac{n}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right)$, S is the skewness and C is the kurtosis of realized return r in the past window

Label 4. Sign of change in serial correlation of realized returns

$$\text{sign}[sc_{t+h} - sc_t]$$

$sc_t = \text{corr}[r_t, r_{t-1}]$, the correlation between returns of two consecutive bars

Prediction labels

Label 5. Sign of change in absolute skewness of realized returns

$$\text{sign}[skew_{t+h} - skew_t]$$

Label 6. Sign of change in kurtosis of realized returns

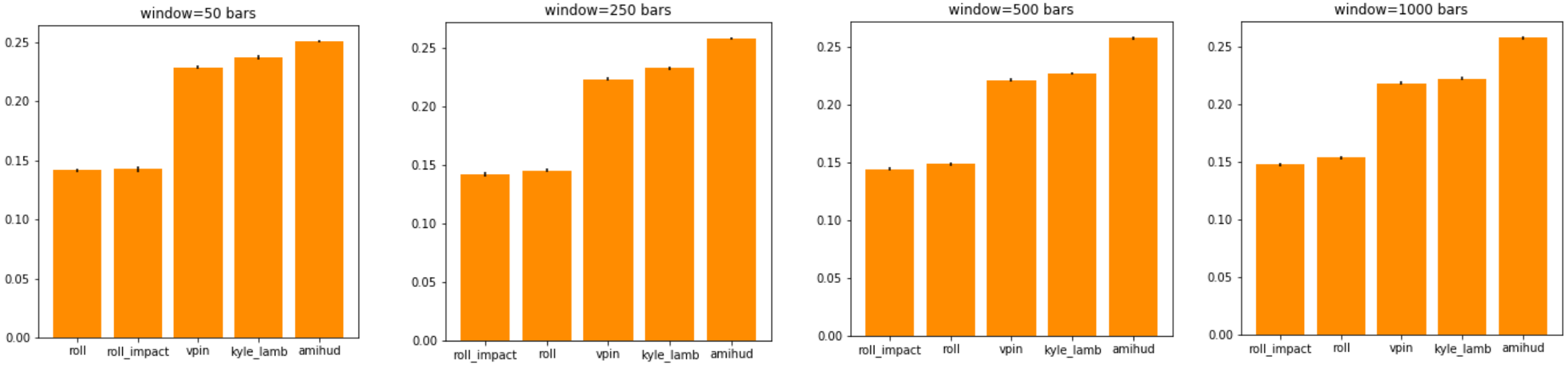
$$\text{sign}[Kurt_{t+h} - Kurt_t]$$

Section III
Results

Microstructure variable feature importance

1. Sign of change in Corwin-Schultz estimator

MDI result

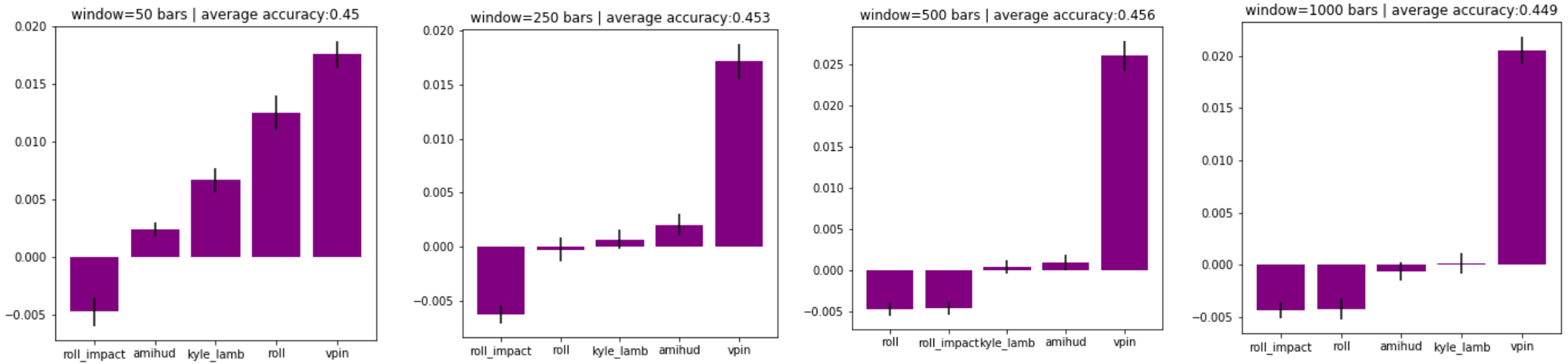


Increasing backward window size for feature generation

Microstructure variable feature importance

1. Sign of change in Corwin-Schultz estimator

MDA result

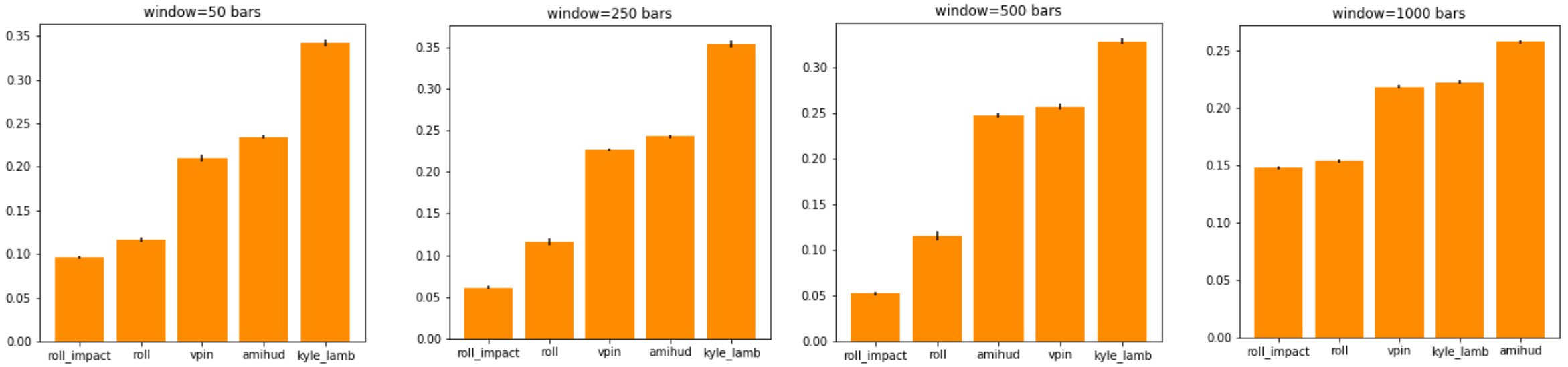


Increasing backward window size for feature generation

Microstructure variable feature importance

2. Sign of change in realized volatility

MDI result

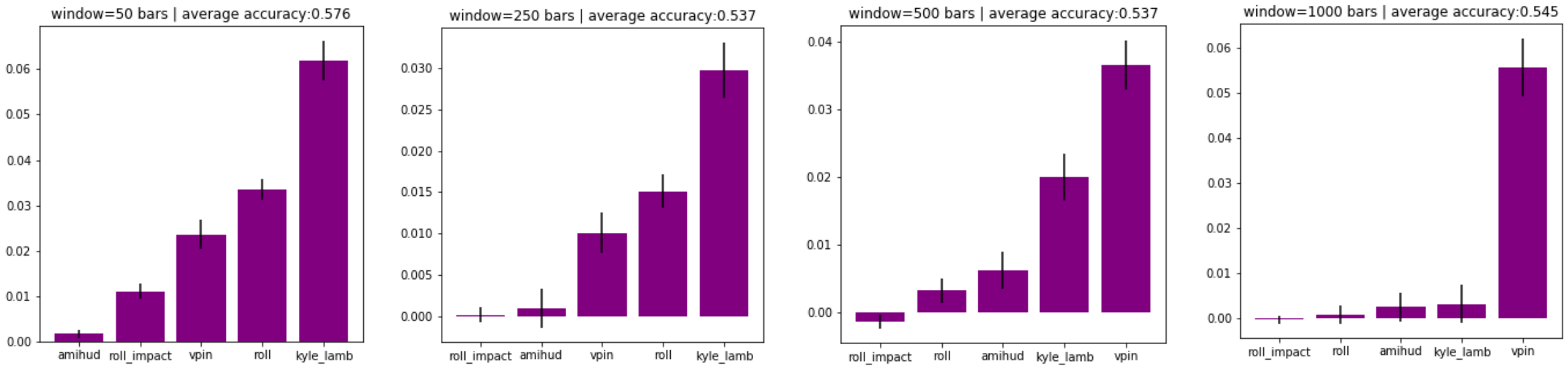


Increasing backward window size for feature generation

Microstructure variable feature importance

2. Sign of change in realized volatility

MDA result

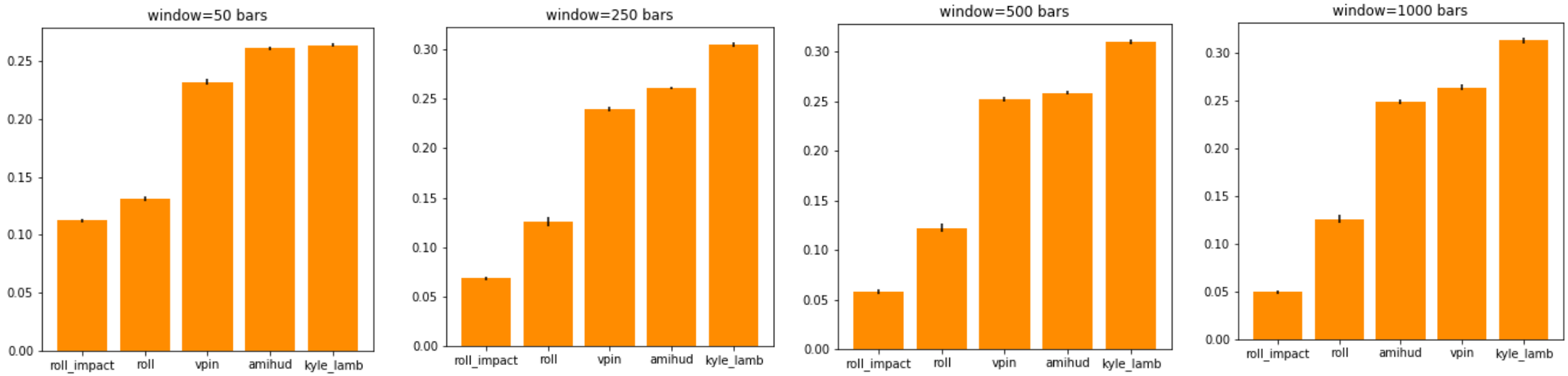


Increasing backward window size for feature generation

Microstructure variable feature importance

3. Sign of change in Jarque-Bera statistics of realized returns

MDI result

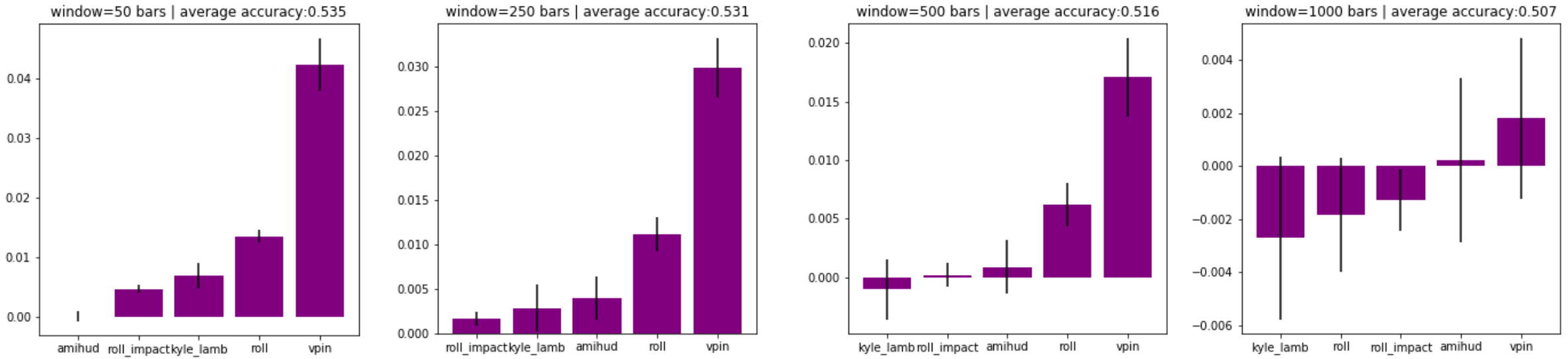


Increasing backward window size for feature generation

Microstructure variable feature importance

3. Sign of change in Jarque-Bera statistics of realized returns

MDA result

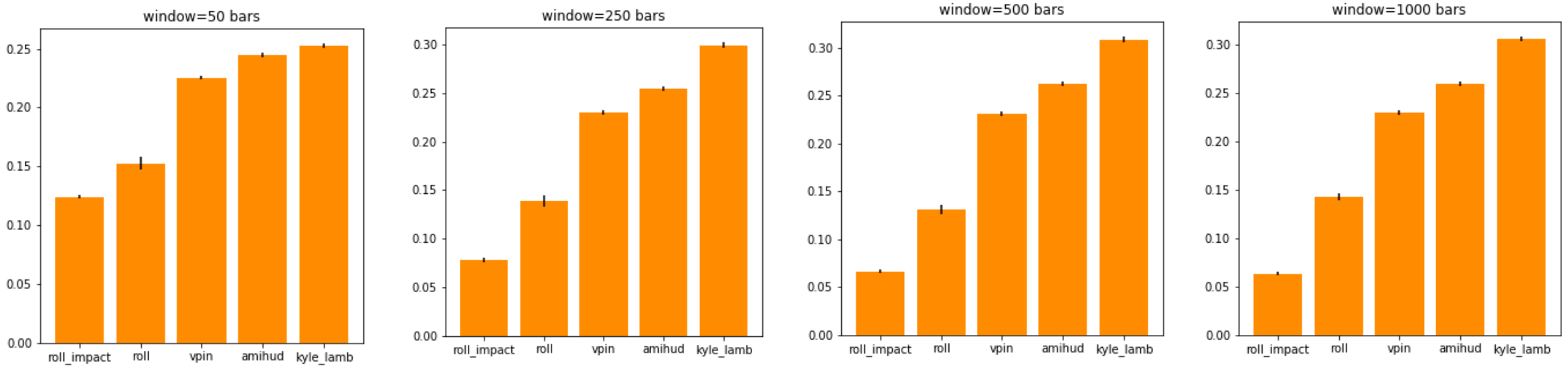


Increasing backward window size for feature generation

Microstructure variable feature importance

4. Sign of change in serial correlation of realized returns

MDI result

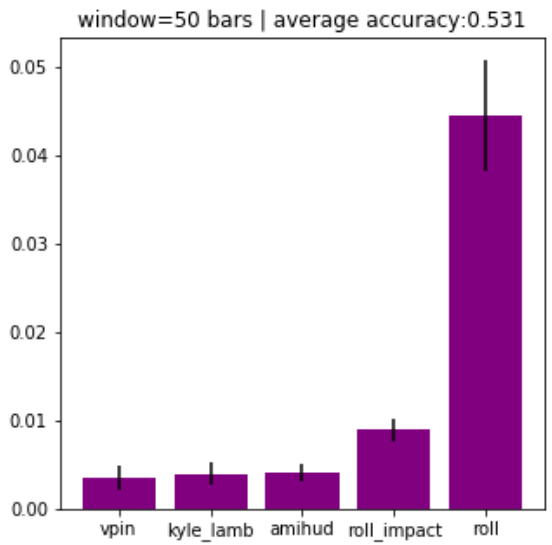


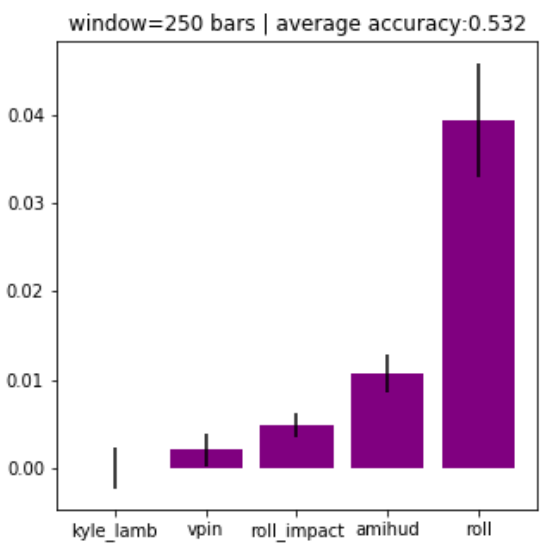
Increasing backward window size for feature generation

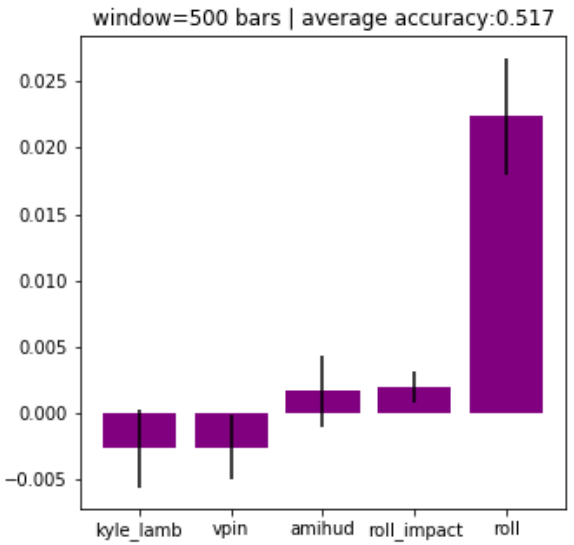
Microstructure variable feature importance

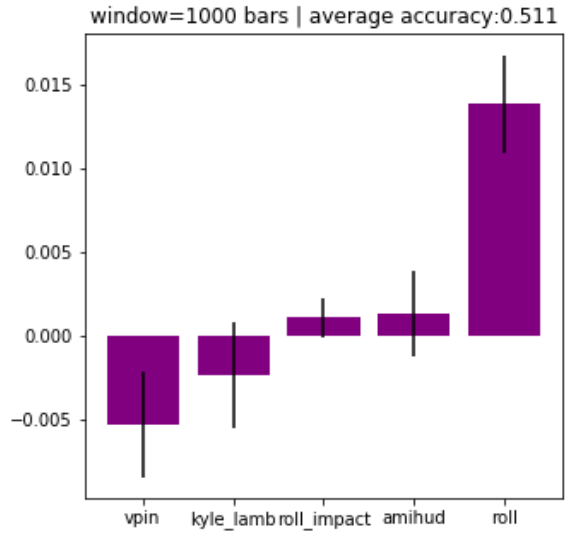
4. Sign of change in serial correlation of realized returns

MDA result







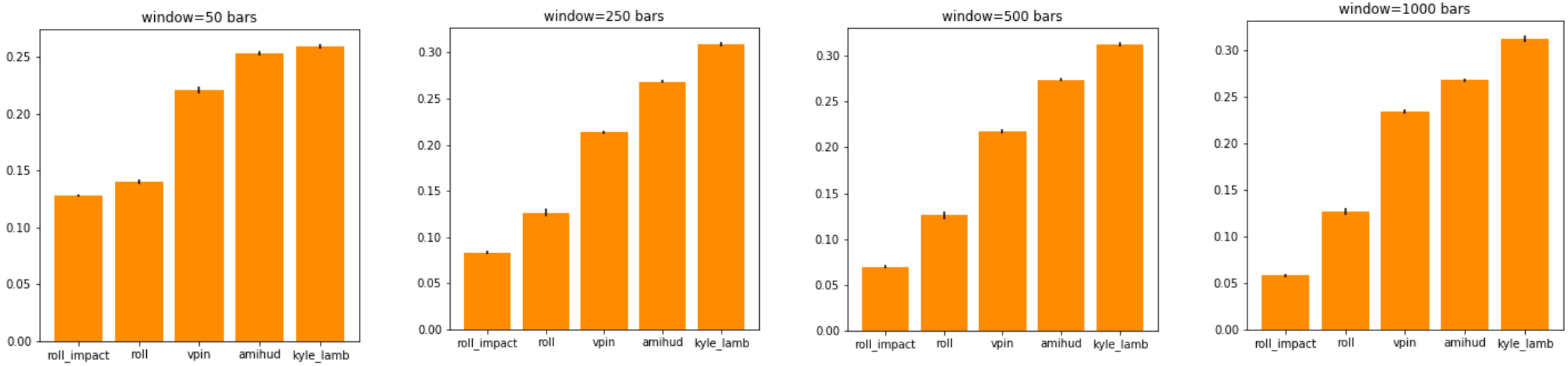


Increasing backward window size for feature generation

Microstructure variable feature importance

5. Sign of change in absolute skewness of realized returns

MDI result

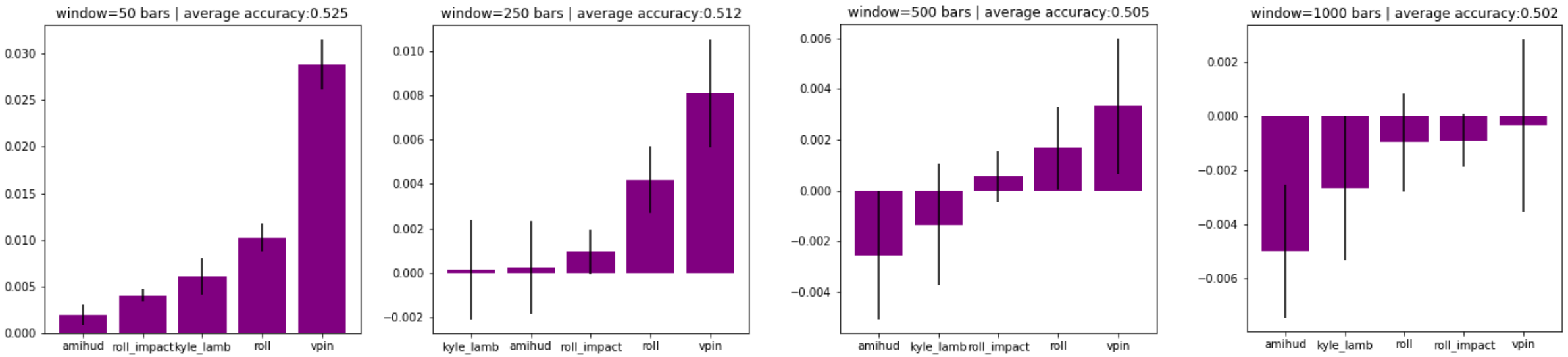


Increasing backward window size for feature generation

Microstructure variable feature importance

5. Sign of change in absolute skewness of realized returns

MDA result

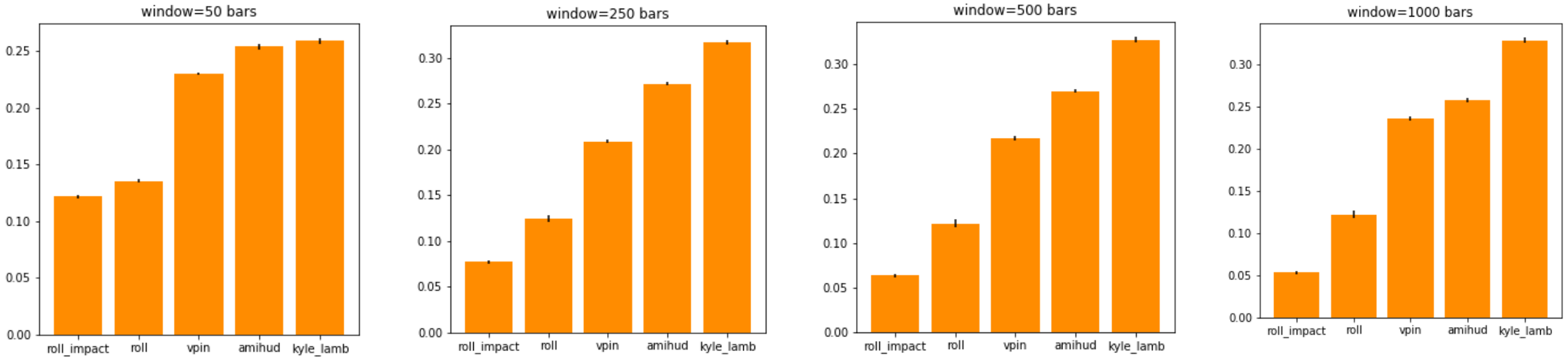


Increasing backward window size for feature generation

Microstructure variable feature importance

6. Sign of change in kurtosis of realized returns

MDI result

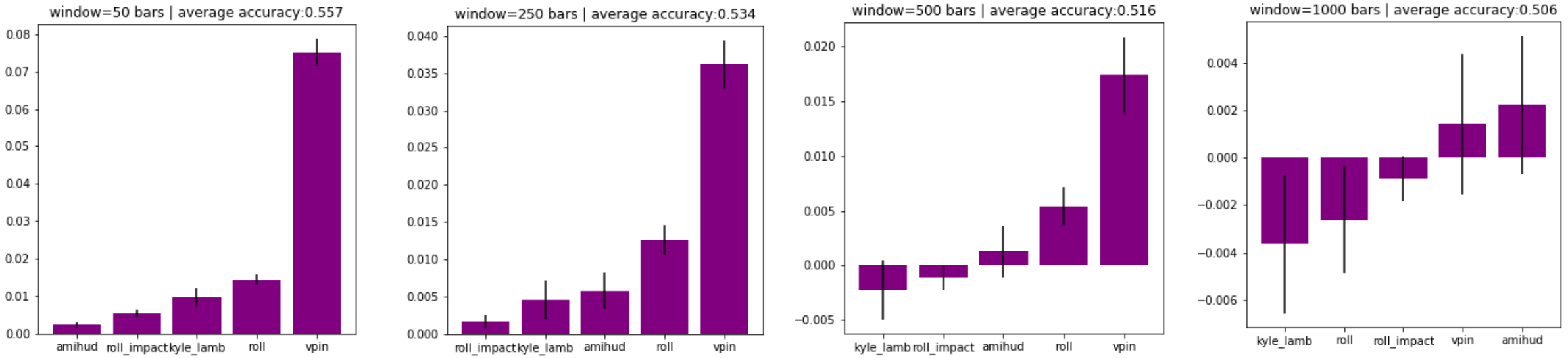


Increasing backward window size for feature generation

Microstructure variable feature importance

6. Sign of change in kurtosis of realized returns

MDA result

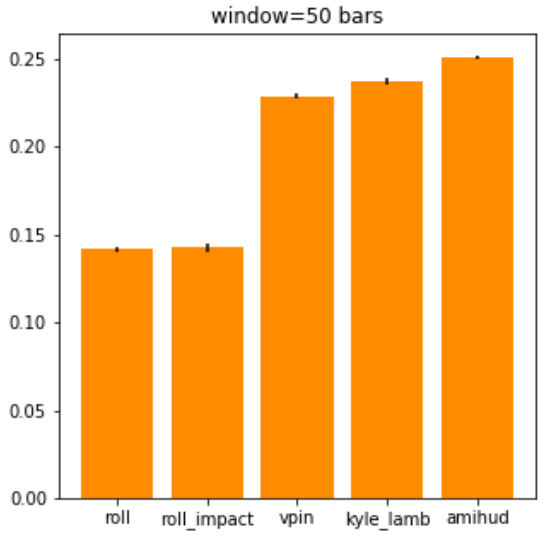


Increasing backward window size for feature generation

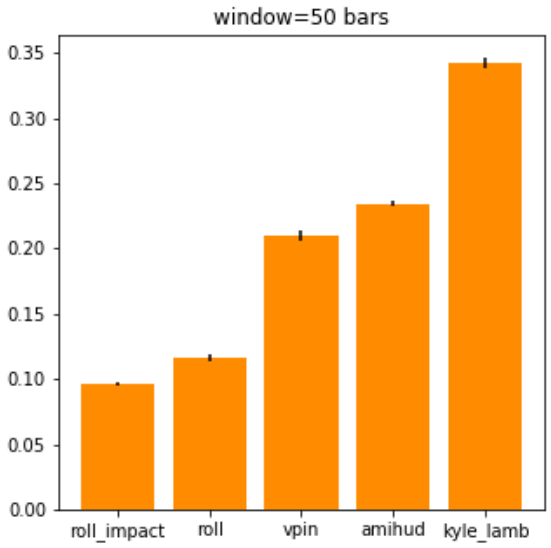
Section IV
Analysis

Kyle & Amihud are best In-Sample (1/2)

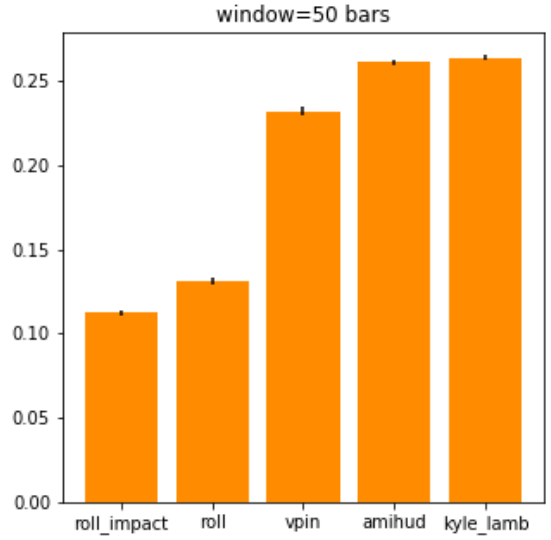
- MDI results have strong similarity across all labels. It is observed that MDI is biased towards features with higher variance (Altmann et al. [2010]). See below for MDI results with 50 bars window.



Corwin-Schultz



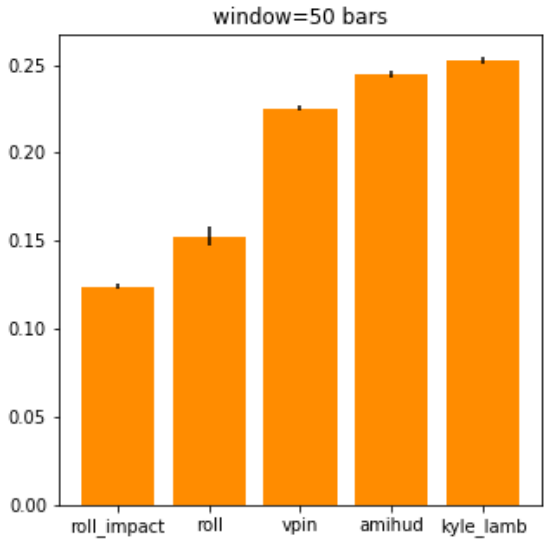
Realized volatility



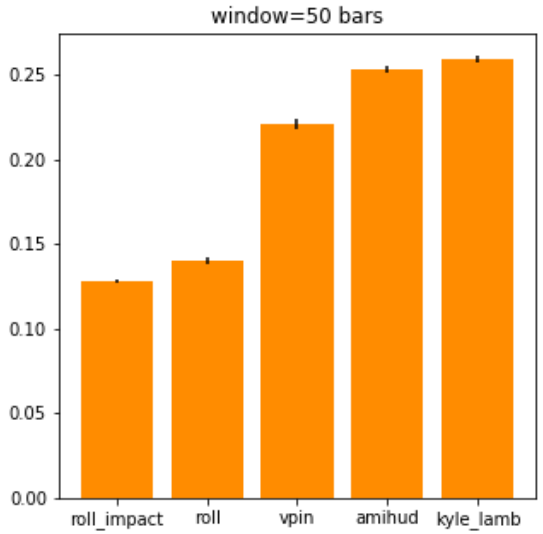
JB statistics

Kyle & Amihud are best In-Sample (2/2)

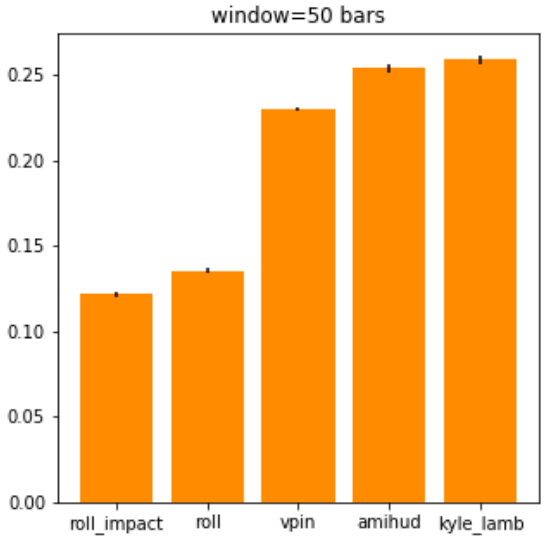
- MDI results have strong similarity across all labels. It is observed that MDI is biased towards features with higher variance (Altmann et al. [2010]). See below for MDI results with 50 bars window.



Sequential correlation



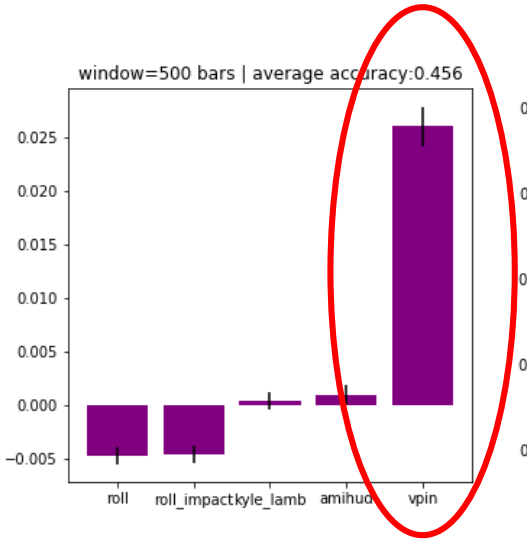
Return skewness



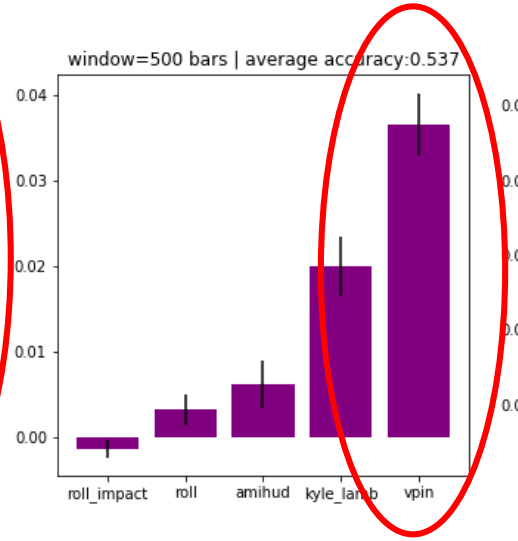
Return kurtosis

VPIN is best Out-Of-Sample (1/2)

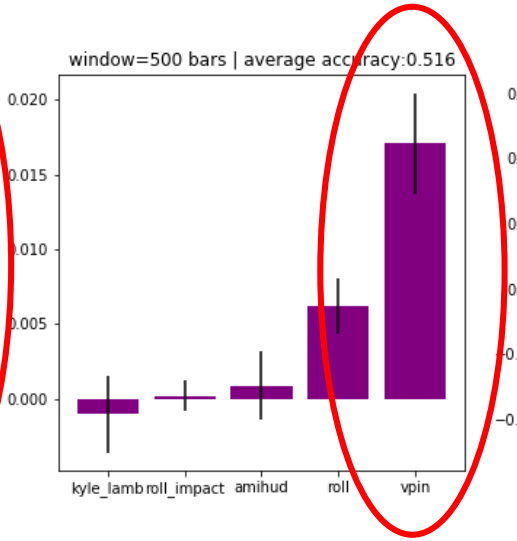
- When the backward window is large, only VPIN can contribute positively to out-of-sample prediction across all labels except sequential correlation. Below are MDA result with 500 bar window



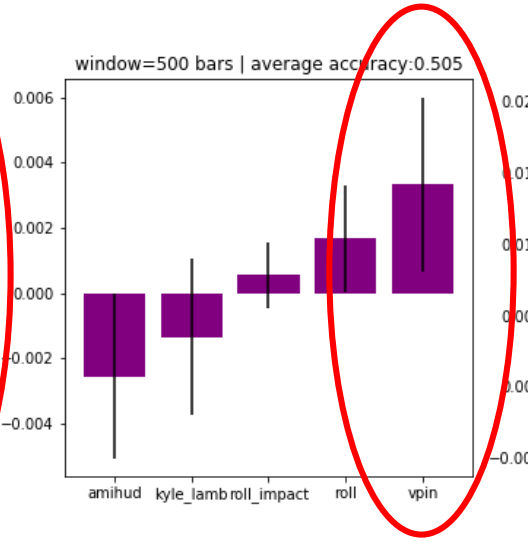
Corwin-Schultz



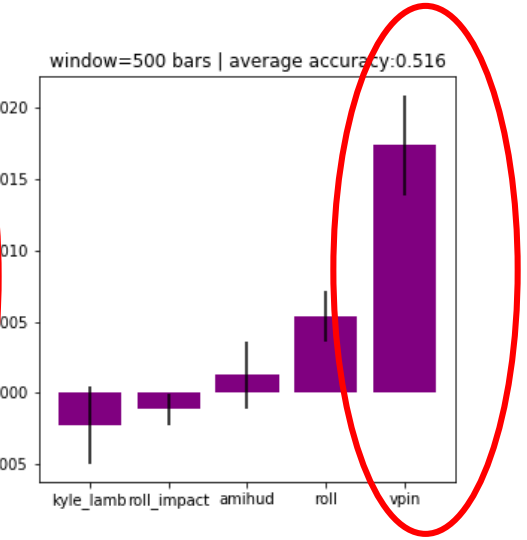
Realized volatility



JB statistics



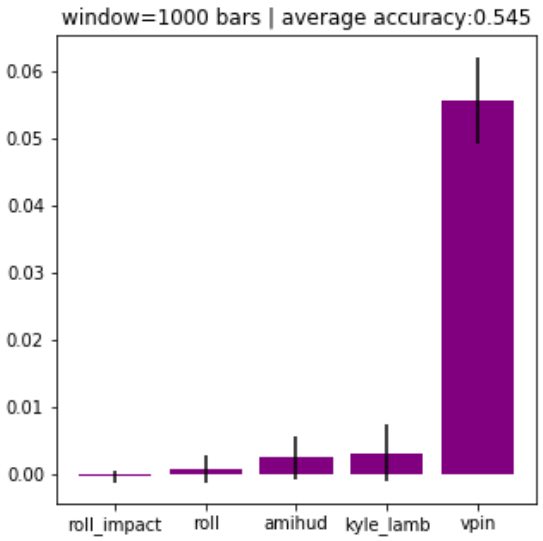
Return skewness



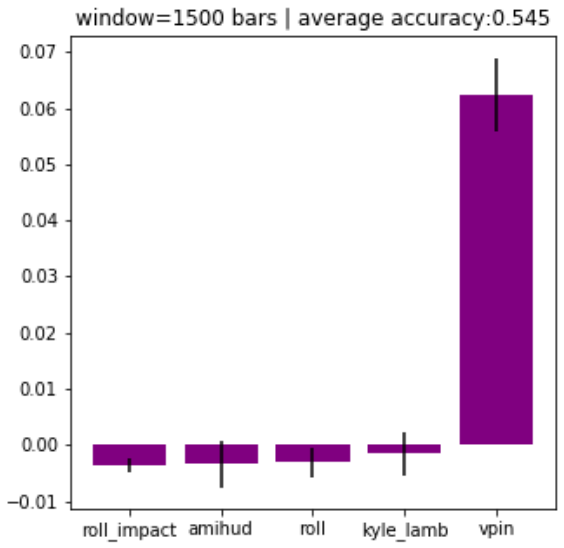
Return kurtosis

VPIN is best Out-Of-Sample (2/2)

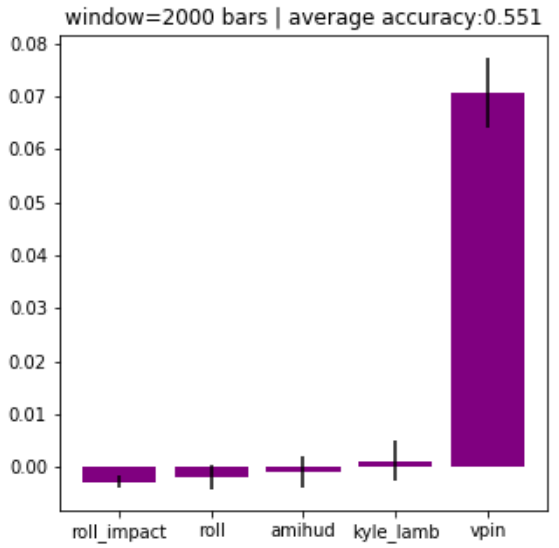
- VPIN's MDA importance at predicting realized volatility remains significant with large window size, even when other variables become irrelevant



1000 bars



1500 bars



2000 bars

Section V
Conclusions

Conclusions (1/3)

- For the prediction of the estimated bid-ask spread, the feature importance remains almost the same across all window sizes, for both MDI and MDA, indicating universality.

- For the prediction of the realized volatility, while Amihud and Roll measures remain stable, VPIN's importance increases while Kyle' lambda decreases as the window size expands, indicating the growth of predictability of VPIN with larger look back window size.

- For the prediction of the JB test the result is similar to realized volatility. VPIN's importance increases while Amihud measure decreases as the window size expands, while the others remain almost the same, indicating the growth of predictability of VPIN with larger look back window size.

Conclusions (2/3)

- For the prediction of the sequential correlation, MDA results demonstrate the Roll measure is much more predictive than all other variables, corresponding to the fact that it is built on past sequential correlation of returns.

- For the prediction of various moments of realized return (volatility, skewness and kurtosis), MDA feature importance shows that VPIN gives the largest contribution consistently, indicating universality.

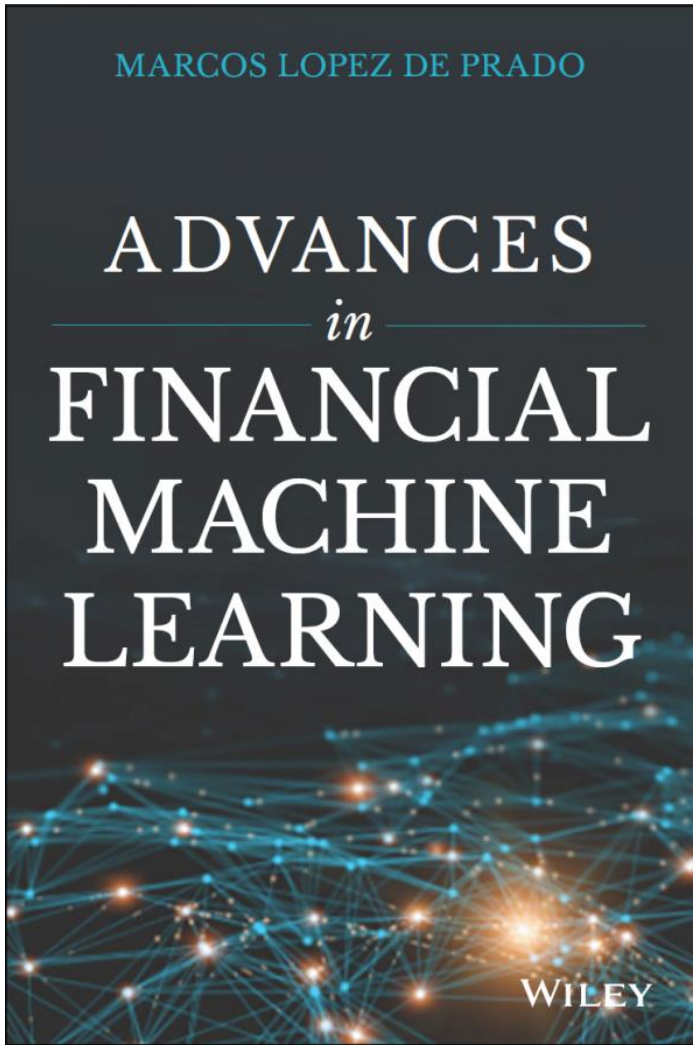
Conclusions (3/3)

- Technologies at the Age of AI provide new methods and perspectives for market microstructure study.
- ML offers new prediction framework with classification on categorical output, compared to traditional regression-based models.
- The 5 prototypical market microstructure variables (Roll Measure, Roll Impact, Kyle's Lambda, Amihud's Lambda and VPIN) show different importances in-sample and out-of-sample. This demonstrates explanatory power does not fully relate to predictability.
- For all prediction labels tested, VPIN is shown to have consistently high importance.

Future directions

- Incorporate more market indicators as features (e.g. VIX)
- Experiment with different bar formations (Volume/Dollar Imbalance Bars)
- Vary forecast horizon and test other prediction labels

For Additional Details



How does one make sense of todays' financial markets in which complex algorithms route orders, financial data is voluminous, and trading speeds are measured in nanoseconds? For academics and practitioners alike, this book fills an important gap in our understanding of investment management in the machine age.
— Prof. **Maureen O'Hara**, Cornell University. Former President of the American Finance Association.

The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

THANKS FOR YOUR ATTENTION!

Bio

Dr. Marcos López de Prado is the chief executive officer of *True Positive Technologies*. He founded *Guggenheim Partners*' Quantitative Investment Strategies (QIS) business, where he developed high-capacity machine learning (ML) strategies that consistently delivered superior risk-adjusted returns. After managing up to $13 billion in assets, Marcos acquired QIS and spun-out that business from Guggenheim in 2018.

Since 2010, Marcos has been a research fellow at *Lawrence Berkeley National Laboratory* (U.S. Department of Energy, Office of Science). One of the top-10 most read authors in finance (SSRN's rankings), he has published dozens of scientific articles on ML and supercomputing in the leading academic journals, and he holds multiple international patent applications on algorithmic trading.

Marcos earned a PhD in Financial Economics (2003), a second PhD in Mathematical Finance (2011) from *Universidad Complutense de Madrid*, and is a recipient of Spain's National Award for Academic Excellence (1999). He completed his post-doctoral research at *Harvard University* and *Cornell University*, where he teaches a Financial ML course at the School of Engineering. Marcos has an Erdős #2 and an Einstein #4 according to the *American Mathematical Society*.

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-