The Review of Financial Studies

Microstructure in the Machine Age

David Easley
Departments of Economics and Information Science, Cornell University

Marcos López de Prado
True Positive Technologies and the Department of Operations Research and Information Engineering, Cornell University

Maureen O'Hara
Johnson College of Business, Cornell University

Zhibai Zhang
Tandon School of Engineering, New York University

Understanding modern market microstructure phenomena requires large amounts of data and advanced mathematical tools. We demonstrate how machine learning can be applied to microstructural research. We find that microstructure measures continue to provide insights into the price process in current complex markets. Some microstructure features with high explanatory power exhibit low predictive power, while others with less explanatory power have more predictive power. We find that some microstructure-based measures are useful for out-of-sample prediction of various market statistics, leading to questions about market efficiency. We also show how microstructure measures can have important cross-asset effects. Our results are derived using 87 liquid futures contracts across all asset classes. (*JEL* C02, C58, G19, G14, E44)

Received February 27, 2019; editorial decision April 28, 2020 by Editor Wei Jiang. Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

One might have expected that as markets became faster, market data became more copious, and technology superseded human participants, the microstructure of markets would play an ever-decreasing role in explaining

We thank Lee Cohn, Michael Lewis, Michael Lock, and Yaxiong Zeng for useful discussion and help with preparing some of the data used in this study. We also thank seminar participants at Cornell University, including Pamela Moulton and Shawn Mankad; at the Georgia State FinTech Conference, including our discussant at the conference, Haoxiang Zhu; at the NBER, including our discussant at the conference, Joel Hasbrouck; and at Tsinghua University; and the editor (Wei Jiang) and referees of this journal. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to David Easley, Department of Economics, Uris Hall, Cornell University, Ithaca, NY 14850. Email: dae3@cornell.edu.

The Review of Financial Studies 34 (2021) 3316–3363

doi:10.1093/rfs/hhaa078 Advance Access publication July 7, 2020

market behavior. The opposite is true. When time scales shrink to nanoseconds, how the market is structured turns out to be critical in predicting where the market is going. And when data explode to mammoth dimensions, being able to characterize what variables related to market frictions can and should matter for market behavior, a particular focus of microstructure research, takes on even more significance. Yet, despite this continued importance, microstructure research faces some daunting challenges in this new era.[1] The ubiquity of computerized trading, abetted by the rise of big data, has increased the complexity of trading strategies far beyond what is envisioned in simple microstructure models. Similarly, the empirical measures that fill the microstructure "toolbox" were constructed based on simple within-asset relationships that may no longer hold in the high-frequency world of cross-asset trading. The problem, simply put, is that microstructure research needs to evolve.

In this paper, we demonstrate how machine learning techniques can play an important role in that evolution. Much as microstructure research is often used to predict how trading will affect price and liquidity dynamics, machine learning can potentially improve those predictions given complex data and computational constraints. It is not a given that machine learning techniques will lead to improvements in our ability to predict these variables. The nature of financial market data, for example, is often fairly well ordered, meaning that simpler approaches such as logistic methods may do reasonably well in some settings. But the ability of machine learning to process complex data using nonparametric algorithms designed to be adaptive facilitates extracting patterns in data that parametric models may not recognize, setting the stage for higher predictive power. And in financial market settings, even a small advantage executed over a large number of trades can result in significant effects for market participants.

Using a random forest machine learning algorithm, we investigate how well some standard empirical microstructure measures (termed "features" in machine learning parlance) predict variables of interest to market participants. Our focus is on a set of variables typically used in electronic market making, dynamic market hedging strategies, and volatility estimation. Our purpose here is not to provide an exhaustive examination of market data predictability, but rather to illustrate how machine learning can bring new insights to microstructure research by showing what features actually work for out-of-sample predictability. In doing so, we also provide clear evidence of the value of some extant microstructure variables for understanding the new dynamics of market behavior.

Our analysis draws on three generations of market microstructure models to provide specific measures as inputs to our machine learning investigation.

[1] For more discussion, see O'Hara (2015).

These variables include the Roll measure, the Roll impact, a volatility measure, Kyle's λ, the Amihud measure, and the volume-synchronized probability of informed trading (VPIN). We focus on predicting six important outcomes of market price dynamics using a variety of lookback windows (to compute the market microstructure measures) and forecast horizons: sign of change of the bid-ask spread; sign of change in realized volatility; sign of change in Jacques-Bera statistic; sign of change in sequential correlation of realized returns; sign of change in absolute skewness of returns; and sign of change in kurtosis of realized returns. We evaluate the importance of each feature using mean-decreased impurity (an in-sample measure) and mean-decreased accuracy (an out-of-sample measure) methods. We use five years of tick data from the 87 most liquid futures traded globally (including indices, currencies, commodities, short rates, and fixed-income instruments). This extensive sample, one of the largest ever used in a microstructure analysis, epitomizes the big data that can be brought to bear in machine learning analyses. This scale allows us to establish the validity and accuracy of our findings generally, and not merely for a specific contract or asset class.

Market microstructure models typically analyze markets asset by asset, so we begin by restricting our attention to the use of within-asset market microstructure measures for prediction. However, the evolution of markets, particularly the rise of high-frequency and machine-based trading, has made cross-asset trading more the norm. Consequently, we next ask about cross-asset effects: are market microstructure measures in one asset useful for prediction of price and liquidity dynamics in another asset? Here extant market microstructure models provide no guidance, so a machine learning approach is a natural way to ask model-free questions.

Our within-asset research provides a number of results. As expected, we find that the various microstructure measures show different importance for in-sample and out-of-sample estimation, illustrating how variables that may have explanatory power in-sample need not have predictive power out-of-sample. Consistent with previous studies, all of the measures appear to have in-sample explanatory power. Across the six predicted variables, the Amihud measure, the Chicago Board Options Exchange volatility index (VIX), and VPIN have the best performance in-sample, while VPIN has the best out-of-sample performance. For example, predicting the sign of change in the bid-ask spread, in-sample results show that Amihud and VPIN consistently have the largest importance across all window sizes, whereas out-of-sample results show that VPIN predominates. Indeed, out-of-sample prediction results show that VPIN is the most important predictor for five variables, with the Roll measure dominating for the sixth (predicting the sign of change in sequential correlation). The variables we predict should be affected by trade imbalances related to information-based trading, and this is what VPIN was designed to measure, so VPIN's predictive power is not surprising. We interpret these results as showing that simple measures designed to reflect market frictions still work in modern, complex markets dominated by machine-based trading. These results demonstrate not only the importance of particular microstructure-related variables, but also the possibility of successful prediction of future market dynamics. As we discuss, such predictions have wide applicability for areas such as risk management, dynamic trading strategies, and electronic market making.

Our analysis of cross-asset effects provides another set of interesting results. Perhaps most important is that including cross-asset market microstructure measures in the set of features considered by our random forests improves out-of-sample predictability. The particular cross-asset measures that are most useful in prediction differ asset by asset, but there is some regularity in which types of measures are important. The relative importance of own-asset Amihud, Roll, and VPIN measures varies across the variables we are interested in predicting, but for every prediction, all of these own-asset measures remain important. However, with cross-asset measures included, the importance of predictors changes, and own-asset VPIN is no longer the most important predictor. These cross-asset measures are correlated with each other and with own-asset measures, so the change in importances is not surprising. Equally important, we identify several microstructure measures of trade in specific financial futures as being important to prediction across assets. Perhaps surprisingly, we find that measures based on trade in the E-mini are not particularly important. These cross-asset results raise interesting questions about the systemic influences of information in the market.

For many readers and market participants, it may be the predictive power of our machine learning approach, rather than an exploration of what features create it, that is of primary interest. Here we ask a simple question: does the predictive power achieved using a random forest approach exceed that obtained from using a logistic regression? As might be expected when the number of features is small, as is the case when we restrict attention to own-asset measures, logistic regression and random forest have similar predictive power, with the logistic generally being slightly more accurate. When we include numerous cross-asset features, however, predictability using the random forest is generally greater than that obtained from logistic regression. The random forest seems to do a better job of isolating important features and basing prediction on them when many noisy features are being considered. We use a Sharpe ratio analysis to suggest the scale of these accuracy gains, finding that even relatively small accuracy gains matter for investment efficiency.

Our paper joins a growing literature examining the implications of machine learning and big data for economic research. Varian (2014), Abadie and Kasy (forthcoming), and Mullainathan and Spiess (2017) provide excellent discussions of how machine learning can be applied to analyze economic problems involving big data, while recent applications of such techniques can be found in Bajari et al. (2015) and Cavallo and Rigobon (2016). In the finance area, Chinco, Clark-Joseph, and Ye (2018) apply LASSO techniques to make one-minute-ahead equity return forecasts; Rossi (2018) uses boosted regression trees to forecast stock returns and volatility; Krauss, Do, and Huck (2017) use machine learning for statistical arbitrage on the S&P 500; and Lopez de Prado (2018) provides extensive analyses of financial machine learning techniques and applications. Philip (2020) uses reinforcement learning to estimate the permanent price impact of a trade and shows that the nonlinear nature of the price impact results in the reinforcement learning approach preforming better than a traditional VAR approach. In Philip's analysis, as in ours, a simple linear specification does well when the number of variables is small, but the machine learning approach dominates when more variables and a more complex environment are considered. Gu, Kelly, and Xiu (2018) apply multiple machine learning regression algorithms in asset pricing and find that these methods can give rise to better R^2 values than standard econometric models. Our work contributes to this literature by showing how supervised machine learning techniques combined with metrics suggested by microstructure theories can help identify important market variables irrespective of functional form. We believe that machine learning's decoupling of the search for variables from the search for specification will be important for the development of microstructure research.

This paper is organized as follows. In the next section, we set out the variables we are interested in predicting and the microstructure variables we use as inputs in our analysis. Section 2 introduces the random forest classification method and feature importance measures. We discuss two such measures: mean decreased impurity (MDI) and mean decreased accuracy (MDA). We also explain how we categorize realized outcomes in terms of binary labels. In Section 3, we discuss the data, how we transform the data into units of analysis called bars, and the microstructure variable definitions we use in the analysis. Section 4 presents our within-asset empirical results and investigates their robustness with respect to various lookback and forecast window sizes, alternative hyperparameter configurations, time periods, and different bar types. We also compare our random forest results with results obtained from logistic regression. In Section 5, we include cross-asset features and again ask about out-of-sample predictive power and the importance of both within-asset and cross-asset features for those predictions. In Section 6, we discuss prediction accuracy and its uses. We also consider its sensitivity to various specifications including using a random forest or logistic regression for prediction, inclusion of lagged returns and volatility, inclusion of cross-asset features, and use of alternative accuracy measures. Section 7 concludes by discussing the implications of our results for trading strategies, considers what we have learned about the explanatory and predictive roles of microstructure variables, and suggests an agenda for future microstructure research in the machine age.

1. Microstructure Variables and Market Movements

Microstructure models provide variables that indirectly measure the implications of market frictions. To the extent that these measures are successful, they should predict the future values or movements in market metrics such as bid-ask spreads, volatility, and other variables related to the shape of the distribution of returns. Some models (which we will term “first generation”) use price data for this task. Examples here are the Roll (1984) measure, which uses price sequences to predict effective bid-ask spreads, and the Corwin and Schultz (2012) bid-ask spread estimator. Second-generation models focus on price and volume data, generating metrics such as the Kyle (1985) lambda, the Amihud (2002) measure, and Hasbrouck’s (2009) lambda. Third-generation models use trade data, inspiring metrics such PIN, the probability of informed trading (Easley et al. 1996), and VPIN, the volume-synchronized probability of informed trading (Easley et al. 2012b). In our analysis, we evaluate the predictive power of measures representative of these three generations of microstructural models.

The specific variables we select are the Amihud measure, the Roll measure, the Roll impact measure, the Kyle lambda, VPIN, and volatility (precise definitions are given in Section 3). The Amihud measure is a general metric of illiquidity that can arise from factors such as market maker inventory pressures, information, or limited risk bearing in markets. The Roll measures capture features of the order flow reflecting sequences of trades, which in turn can influence overall liquidity. The Kyle lambda and VPIN are measures of asymmetric information in markets. Because illiquidity and information effects would be expected to lead to price volatility, we also include a measure of volatility (VIX) in in our analysis.[2]

Being able to forecast future developments in the price process and liquidity has obvious importance for traders, regulators, and researchers, but less apparent is how well these standard microstructure measures work in current markets. The models that produce these measures are relatively simple and were designed at a time when markets were less complex. Those models do not provide much guidance about functional forms describing the relationship between any of these measures for an asset and the price or liquidity dynamics of that asset. They provide no guidance at all for any cross-asset effects. So imposing a particular functional form for these relationships, even a flexible one, and applying standard econometric techniques to estimate it could potentially obscure any relationship.[3]

[2] Some researchers have argued that microstructure metrics work because they capture volatility. For example, Andersen and Bondarenko (2015) claim that VPIN is simply a volatility effect. Including volatility allows us to evaluate these claims. We show that, in our data set, VPIN and VIX are virtually uncorrelated and perform very differently particularly out of sample.

[3] As Mullainathan and Spiess (2017) explain, standard econometric techniques are well suited for variance adjudication; however, they often provide suboptimal forecasts. The reason is that the best forecast estimators

Our interest is in evaluating predictability using various microstructure variables. We begin with data about microstructure variables (such as illiquidity, Kyle's lambda, or VPIN) and data about the market measures (such as bid-ask spreads, volatility, and the like) we are interested in predicting. However, unlike the standard approach in econometrics, we do not attempt to prespecify an underlying data-generating process, and so we do not attempt to estimate parameters of a model relating our microstructure measures to market measures. Our primary interest is in understanding which microstructure variables are useful for prediction and which ones are not useful. We are agnostic about the mechanism relating the variables in our data set to each other, as attempting to specify a mechanism, no matter how complex its structure or underlying probability space, is unnecessarily limiting for our data-exploratory purposes. We believe that this machine learning point of view is more powerful for the questions we want to ask, although we do recognize that for other interesting questions more closely related to developing an understanding of why one measure is a better predictor than another is, specifying a data-generating process and applying standard econometric tools may be more productive.

Thus, we use machine learning to investigate the efficacy of a set of microstructure measures for forecasting a set of variables of wide interest in the market. We discuss in detail in Section 2 how the random forest algorithm we use works, but it is important to stress that we use the algorithm to predict the sign of changes in variables, rather than to provide actual point predictions. While this might seem of limited importance, we explain below why this is not the case and discuss how for our candidate variables such forecasts can be used in practice. The variables we attempt to predict are relevant to all forms of electronic market making and order execution.[4]

1.1 Sign of change of the bid-ask spread

Both market makers and execution traders have an interest in predicting whether the bid-ask spread will widen or narrow over the time frame of their order's implementation. When we expect the bid-ask spread to widen, an execution algorithm could use that expectation to increase volume participation, thereby increasing the portion of the executed order before an increase in transaction costs materializes. Conversely, when we expect the bid-ask spread to narrow, an execution algorithm could use that expectation to decrease volume participation and thereby execute a larger portion of the order after the fall in transaction costs. The magnitude of the change in volume participation would be a function of the

may not be BLUE (best linear unbiased estimator). Unbiasedness is undoubtedly a useful property when the model is properly specified; however, it may be a hindrance when important explanatory variables are missing or when the interaction between variables is not correctly modeled.

[4] We consider only positive and negative changes. One could, and for an investment or trading strategy probably would want to, consider a finer partition of the set of changes at least taking into account a third category in which the change, either positive or negative, is small. One could also focus on predicting only positive or only negative changes.

trader's confidence in the forecast's accuracy. For regulators and researchers, understanding the determinants of bid-ask spread is a long-standing topic of interest.

1.2 Sign of change in realized volatility

When we expect realized volatility to increase, an execution algorithm could use that expectation to increase the volume participation in order to reduce the uncertainty of the average fill price (market risk). It is not necessarily true that we would like to decrease the volume participation if we expect a decrease in realized volatility, because by the time the volatility has decreased, prices may have drifted away from our target. In general, we would like to increase the volume participation rate if we forecast an increase in realized volatility, and reduce the volume participation rate after a decrease in realized volatility has already materialized.

1.3 Sign of change in Jarque-Bera statistic

The Jarque-Bera statistic tests for the null hypothesis that observations are drawn from a normal distribution. This is relevant for risk management purposes, as many risk models assume normality of returns. A higher probability of non-normal returns reduces our confidence in those models. For example, a risk manager may want to reduce the significance level (false-positive rate, type I error probability) of his Gaussian models when returns are expected to be non-normal. In addition, when returns are non-normal, or serially correlated, implementation shortfall estimates may be too small.

1.4 Sign of change in kurtosis/sign of change in absolute skewness of returns

The Jarque-Bera statistic uses skewness and kurtosis to test for normality of observations. This test implies a trade-off between skewness and kurtosis, in the sense that the test may not reject the null hypothesis of normality when an increase in skewness is offset with a decrease in kurtosis. However, offsetting skewness with kurtosis is not without economic meaning. Because skewness is an odd moment, it deforms the normal distribution by shifting its probability toward one side. One possible reason for this deformation is the presence of informed traders, who push prices in an attempt to fill orders before a piece of news is widely known. In contrast, because kurtosis is an even moment, it deforms the normal distribution by shifting its probability symmetrically toward extreme events. One possible explanation for this deformation is a reduction of liquidity, as market makers reduce the size of their quotes in anticipation of a news release, hence increasing the likelihood of extreme outcomes on either side. From an execution and portfolio management perspective, it is important to differentiate between these two causes of non-normality and to forecast them separately.

1.5 Sign of change in sequential correlation of realized returns

Another common assumption of risk models (e.g., in value-at-risk approaches) is that returns are serially uncorrelated. Being able to predict serial correlation offers an insight into how unrealistic this assumption is. When returns are serially correlated, trends occur with a higher frequency than would be otherwise expected. This leads to greater potential drawdowns and time underwater. As in the non-normal case, a higher probability of serially correlated returns reduces our confidence in models that assume an uncorrelated structure. It would therefore be rational to reduce the significance level of this kind of risk model when returns are expected to be serially correlated.

2. The Random Forest Classification Algorithm and Feature Importance Measures

In this section, we introduce the random forest classification algorithm and explain how we use it to evaluate the predictive power of a set of explanatory variables. In machine learning, classification is the practice of using explanatory variables to predict a categorical/discrete target variable. It is analogous to regression in that both are fitted by minimizing an error function built on the explanatory and target variables in the training data set. However, in our machine learning problem, the target variable is discrete (e.g., "yes" or "no"), and so the error functions popular for regression (e.g., mean-squared-error) are not viable. Instead, useful error functions include measures such as cross-entropy and information gain. We refer to the explanatory variables as features and the endogenous variables as labels.

Among machine learning classification methods, random forest is one of the most robust and widely used algorithms. As Varian (2014) notes, "This method produces surprisingly good out-of-sample fits, particularly with highly nonlinear data."[5] It consists of a number of individual classifiers, called decision tree algorithms, and uses the mean of these trees' classifications as its prediction. As the number of low-correlated trees increases, the variance of the forest's forecasting error becomes smaller, hence reducing the chance of the algorithm overfitting the data.

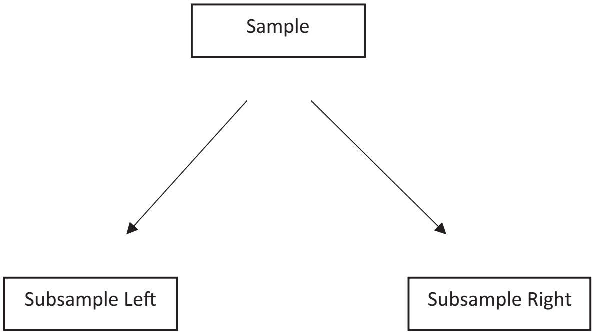
We initially apply our machine learning algorithm one futures contract at a time.[6] So it operates on a data set, or sample, $\{(x_t, y_t)\}_{t=1}^{T}$, consisting of observations of features (x) and a label (y) for the selected contract, with t indexing T observations.[7] The first step in creating a random forest is to build a decision tree by splitting the sample into two subsamples, and then splitting each of these subsamples into two subsamples, and so on. Graphically, the

[5] See Varian (2014, p. 14). This article provides a description of the random forest technique, as does Low et al. (2018, ch.6).

[6] In Section 4, we extend our analysis to consider the effect of cross-asset features on the variables we predict.

[7] We discuss creation of the sample in subsequent sections and the creation of a forest of trees later in this section.

decision tree consists of numerous sequential splits, each of which takes the following form:



To create the split, we first compute for each feature the information gain that would be created by splitting the sample using that feature. For any split of a sample S at node n in the tree into two subsamples, L and R, this information gain is

$$IG(S,n) = I(S) - \frac{N_L}{N_S} I(L) - \frac{N_R}{N_S} I(R), \tag{1}$$

where we use the Gini Index I(S) $= \sum_i p_i(1-p_i)$ as our purity measure for any data set S; p_i is the fraction of labels of the ith class in data set S; and N_S, N_L, and N_R are the number of data points in the sample, the left subsample, and the right subsample. The information gain from using a particular feature to split the sample is then defined to be the maximal gain that can be obtained by choosing a value of the feature and splitting the sample such that all data points with smaller values of that feature are in the left subsample and those with larger values of that feature are in the right subsample. Intuitively, the information gain is maximal when a feature is able to split the data into two pure subsets (subsets with a single label). If the data could be split using the selected feature so that each subsample was pure (contained only increases or only decreases in the label), then the information gain would take on its maximum possible value I(S), while a less pure split would produce a smaller gain.

The actual split of sample S at node n is the one that maximizes the information gain over the choice of features used to create the split. This procedure is repeated for each subsample and for each of the new subsamples created by any split until either a predetermined stopping criterion is reached or until no additional splits yield any information gain. We allow our trees to grow without bound; we consider the effect of bounds in Section 4.4. Although the information gain from alternative splits has to be computed many times, this approach is computationally tractable because each split is done using a greedy algorithm—there is no attempt to choose the split by looking ahead to implications of the current split for possible future information gains.

For each contract, and any sample for that contract, we create a random forest by modifying the simple procedure above in two ways. First, we create multiple decision trees and assign each tree a bootstrapped sample from our underlying sample. The averaging produced by bootstrapping reduces variance that could otherwise result from fitting a single tree to noise in the data set. Second, at each node in each tree, we consider only two randomly selected features as candidates to determine the optimal split.[8] This second modification is done to take into account the possibility that one feature dominates splits even if a second highly correlated feature would produce similar but slightly smaller information gains. Without restricting attention to randomly selected sets of features, we would attribute too much importance to the first feature relative to the second one. For any contract, and data set for that contract, we create 100 trees in our random forest. Finally, given any feature vector, the prediction made by our random forest for the sign of the label is determined by majority vote across the trees in the forest.[9]

For decades, researchers have recognized the prevalence of hierarchical relationships in economic and financial systems. As Nobel laureate Simon (1962) put it, “The central theme that runs through my remarks is that complexity frequently takes the form of hierarchy, and that hierarchic systems have some common properties that are independent of their specific content.” How exactly to measure the contribution of features to the hierarchal structure of the random forest is a critical issue. In our analysis, we use two standard measures of feature importance—mean decreased impurity and mean decreased accuracy.[10]

2.1 Mean decreased impurity–based feature importance

MDI feature importance evaluates the information gain of each feature in all trees, weights them with the number of samples of each split, and sums and then normalizes the score to be one in total. The importance of a feature is its contribution to the building of trees as quantified by the information gain on the splits. Given some data set, the MDI for feature i in that data set is

$$MDI(i)=(1/100)\sum_{N}\sum_{n\in N:v(s_n)=i} p(t)IG(s_n,n), \tag{2}$$

where $v(s_n)$ is the feature used in the split of s_n, with s_0 being the initial data set.

[8] The number of features to consider is a parameter that can be adjusted. We use the standard rule of selecting $int(\sqrt{6})=2$ features, where six is the total number of features we consider.

[9] For more detail on the creation of a random forest for financial data, see Lopez de Prado (2018).

[10] The interested reader can find a detailed explanation of these techniques in Lopez de Prado (2018, ch. 8).

2.2 Mean decreased accuracy–based feature importance

It is worth pointing out that MDI is an in-sample method, as it is derived from the same information used to fit the trees. This makes it similar to a p-value in regression analysis. In contrast, MDA evaluates feature importance out of sample, and, unlike MDI, it can be used with any classifier. The MDA procedure computes feature importance as follows: (a) the data set is split into nonoverlapping training and testing sets; (b) a classifier is trained on the training set using all features; (c) predictions are made on the test set, and a performance measure (e.g., accuracy) is recorded as p_0; (d) values of one of the features, i, in the test set are randomly shuffled, and predictions are remade on the test set; (e) the performance associated with the shuffling of feature i is recorded as p_i. The MDA feature importance of i in the given data set is then

$$MDA(i)=\frac{p_0-p_i}{p_0}. \tag{3}$$

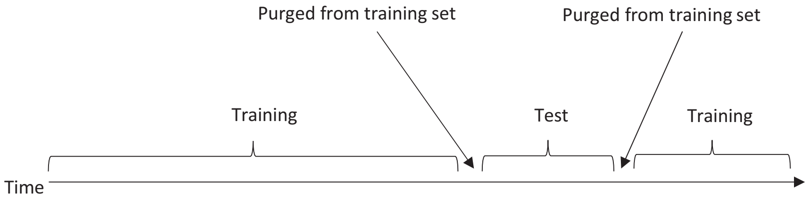
Thus, MDA's feature importance is determined by how the out-of-sample prediction worsens because of shuffling the values of a particular feature. The more deterioration there is in performance, the more important this feature is.

Finally, we turn to the issue of prediction accuracy. We define accuracy to be the number of correct predictions divided by the total number of predictions generated from a given data set split into a training set and a test set.[11] If we applied this idea once to our data set for a futures contract, we would not use the information in the data set efficiently as we would lose the opportunity to train the random forest on all of the data. In particular, the data held out as a test set is not used in training. To use all of the data while still computing out-of-sample predictions, we apply a 10-fold purged cross-validation method to train the forest and compute accuracy Specific details of this approach can be found in Lopez de Prado (2018), but summarily, we (i) partition the entire data set into 10 intervals, (ii) take one as a test set, (iii) purge approximately one week of data from the training set to remove observations that could contain leaked information from the test set (see the figure below), (iv) train the algorithm on the remaining data, and (v) make a prediction on the test set. This procedure is repeated 10 times (once for each test set) so the entire period is tested. Accuracy is computed using all the predictions from the 10 test sets.

Note that the test set may precede some of the trading set in calendar time (as in the figure below). Our data are a time series, so it is important to not include data in the training set that may be influenced by data in the test set. We do this by removing approximately one week of data (equivalent to 250 dollar-volume bars) before and after the test set before we form the training set. This method

[11] One concern with this accuracy measure is that if the data are heavily skewed toward either positive or negative values, then simply predicting the overall more likely outcome can be misleading. For that reason, an alternative accuracy measure (receiver operating characteristics area under the curve [ROC-AUC]) that is not biased by class skewness is often used. We report ROC-AUC results in Section 6.

of purging potentially contaminated data is standard in the machine learning literature; see chapters 11–14 of Lopez de Prado (2018).



3. Data

In this section, we turn to the data and the definitions of the labels and features we use in our analysis. We also address a variety of implementation issues. Our analysis uses both time bars and dollar-volume bars, so we set out how we use tick data to form bars across the various contracts in our sample. In this section, we focus on dollar-volume bars; time bars are discussed in Section 4. Because we use futures data, our data have to "roll" across contract expirations to create a continuous price sequence. We describe how we effectuate that transition using a process akin to creating an exchange-traded fund (ETF) on the contract. Finally, we discuss measurement issues connected with viewing microstructure variables in volume bars as opposed to time-based units.

Our analysis is done on the 87 most liquid futures contracts traded globally, with details of each contract given in Table A1 in Internet Appendix A. We use these futures contracts rather than equities for two reasons. First, we are able to examine the universe of active futures, so there is no issue of selecting a sample out of some larger collection of financial assets. Second, we have complete trade data about the trade of these assets. Futures contracts trade almost continuously (23.75 hours per day), and we have tick data for the entire period. Our sample period begins on July 2, 2012, and ends on October 2, 2017. Tick-level data are available for most of these contracts over a longer period, but we are interested in VIX as a feature and the futures contract on VIX (ticker UX1) only began trading in July 2012. We note that two commodity contracts in our sample (IK1 and BTS1) have shorter sample periods beginning in October 2015.

3.1 Creating dollar-volume bars

We obtain tick-level trade data for each futures contract and aggregate the data into intervals, or bars, based on dollar volume. Aggregating data into bars variously defined over time or volume increments is standard practice in industry and in academic research (see, e.g., Engle and Lange 2001; Easley et al. 2012a; Chakrabarty et al. 2012; Easley et al. 2016; Low et al. 2018). Barardehi, Bernhardt, and Davies (2019) also propose a trade time approach in their measurement of liquidity and show that it works better than a clock time approach. Easley and O'Hara (1992) demonstrated that the time between trades should be correlated with the existence of new information, providing our basis for looking at trade time (volume) instead of clock time. Information arrival results in patterns in volume, essentially akin to intraday seasonalities.[12] By drawing a sample whenever the market exchanges a constant volume, we attempt to mimic the arrival to the market of news of comparable relevance. We use dollar volume to allow comparability across the 87 contracts in our sample. Also, Lopez de Prado (2018) presents evidence that the sampling frequency of dollar-volume bars is more stable than the sampling frequency of time bars or volume bars. One reason for this stability is that dollar-volume bars take into account price fluctuations, hence normalizing the dollar value transacted across different time periods.

The τ-th bar is formed at tick t when

$$\sum_{j=t_0^\tau}^{t} p_j V_j \geq L, \tag{4}$$

where t_0^τ is the index of the first tick in the τth bar, p_j is the trade price at tick j, V_j is the trade volume at tick j, and L is a predetermined threshold that gives roughly 50 bars per day for the year 2016.[13] Note that because average daily trading volume differs across contracts, the dollar volume in each bar will differ across specific futures contracts, but the average daily number of bars will not (in 2016) For each individual contract, on an active day bars will fill faster and there can be more than 50 bars in a day; on an inactive day, bars will fill more slowly and there can be fewer than 50 bars in a day.

We compute each microstructure variable in our analysis at each bar τ by applying a rolling "lookback window" of size W. For example, at bar τ we use bars within the set {τ-W+1, τ-W+2,..., τ-1, τ} to compute the microstructure variables and labels. In our analysis, we consider lookback windows ranging from 25 bars to 2,000 bars.

3.2 Rolling contracts

As futures contracts expire, we need to "roll" the contracts (i.e., sell the expiring one and enter the new one) to form a price series as if it were a continuous instrument. To do so, we transform the price of a futures contract to the value of an ETF that perfectly tracks the futures with $1 initial capital.[14] To understand this process, consider the following example.

Assume we would like to take a long position in the front contract of the E-mini S&P 500 futures (Bloomberg code: ES1 <Index>) from January 2, 2015

[12] As futures often trade over a 23.75-hour day, volume patterns are very pronounced.

[13] We chose 2016 because it is the last full year before the end of our sample.

[14] For a detailed discussion of this technique, see Lopez de Prado (2018).

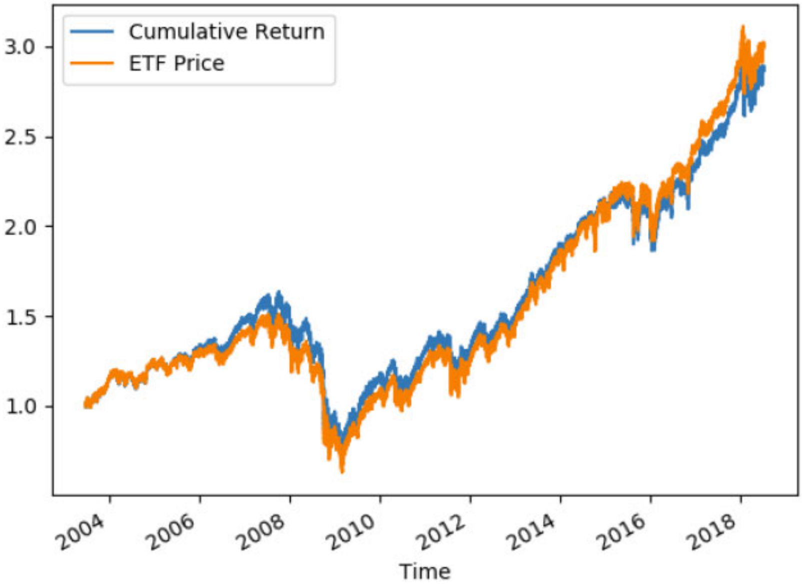


Figure 1
ES1 Index's cumulative return and ETF price
Figure 1 provides a time series graph of the cumulative return on the ES1 Index and the price of the ETF we create to track it. Abbreviation: ETF, exchange-traded fund.

onwards. On March 20, 2015, the front month futures contract is soon to expire and we have to sell it and buy the then second month futures contract, hence "rolling to the next contract." In this rolling process, there is no change in the value of our investment except for the tiny transaction cost. However, there is usually a difference in raw price between the front and second month contracts. If the front month contract was trading at $2,000 while the second month contract was at $2020, then if we simply switched the price time series from front month to second month, there would be a 1% difference. The machine learning algorithm would incorrectly think that there was a sudden jump in price and consider it some sort of a signal.

To avoid this problem, we produce a new time series we call the ETF price of the futures series, which reflects the value of $1 invested in the futures contract assuming one can hold fractional shares. This series starts with 1, and its current value equals the investment's cumulative return (see Internet Appendix, Table A2 for an example). When the futures contract rolls, one sells the old contract and invests all the money in the new contract. During this event, there is no change to the investment assuming zero transaction cost, so the ETF price is unaffected by the artificial change in raw price. Figure 1 provides a plot for ES1 Index's cumulative return and ETF price series.

In Internet Appendix B, we provide the calculation details for this process. In the following analyses, for each futures contract, we use the ETF-based price and the corresponding volume instead of the raw price and volume unless noted otherwise.

3.3 Features and labels

As discussed in the previous section, we focus on a few well-known market microstructure variables. These features are all constructed from the bar data described above. One issue that arises in our construction of these microstructure measures is that they initially were not defined using the same concepts of time periods or bars or using lookback windows. Therefore, for each measure we have to adapt the original definition to our setting. We call these measures by their original names, but we note that they are actually our translation of the measure to our setting.

More specifically, we have:

- Roll measure, given by

$$\begin{aligned} R_t &= 2\sqrt{|cov(\boldsymbol{\Delta P_{\tau}}, \boldsymbol{\Delta P_{\tau-1}})|}, \\ \boldsymbol{\Delta P_{\tau}} &= [\Delta p_{\tau-W}, \Delta p_{\tau-W-1}, \ldots, \Delta p_{\tau}], \\ \boldsymbol{\Delta P_{\tau-1}} &= [\Delta p_{\tau-W-1}, \Delta p_{\tau-W}, \ldots, \Delta p_{\tau-1}], \end{aligned} \tag{5}$$

  where Δp_τ is the change in close price between bars $\tau-1$ and τ and W is the lookback window size.
- Roll impact, which is the Roll measure divided by the dollar value traded over a certain period, is

$$\tilde{R}_t = \frac{2\sqrt{|cov(\boldsymbol{\Delta P_{\tau}}, \boldsymbol{\Delta P_{\tau-1}})|}}{p_\tau V_\tau}. \tag{6}$$

  We evaluate the denominator at each bar and note that for dollar-volume bars, the denominator varies very little between consecutive bars.
- Kyle's lambda is given by

$$\lambda_\tau = \frac{p_\tau - p_{\tau-W}}{\sum_{i=\tau}^{\tau} b_i V_i}, \tag{7}$$

  where $b_i = \text{sign}[p_i - p_{i-1}]$, which is computed at the bar level, and W is the lookback window size.
- Amihud's measure is given by

$$\lambda_\tau^A = \frac{1}{W} \sum_{i=\tau-W+1}^{\tau} \frac{|r_i|}{p_i V_i}, \tag{8}$$

  where r_i, p_i, V_i are the return, price and volume at bar i and W is the lookback window size. Our version of Amihud's measure using dollarvolume bars is closely related to the Barardehi, Bernhardt, and Davies (2019) trade time analogue of the Amihud measure.

Table 1
Correlation matrix of microstructure variables

	Roll	Roll_impact	Kyle lambda	Amihud	VPIN	UX1 (VIX)
Roll	1.0000	0.9275	0.0001	0.3441	0.0190	0.1574
Roll_impact	0.9275	1.0000	0.0001	0.3255	0.0141	0.1506
Kyle lambda	0.0001	0.0001	1.0000	0.0002	−0.0001	0.0008
Amihud	0.3441	0.3255	0.0002	1.0000	0.0776	0.2971
VPIN	0.0190	0.0141	−0.0001	0.0776	1.0000	0.0320
UX (VIX)	0.1574	0.1506	0.0008	0.2971	0.0320	1.0000

Table 1 reports the correlations of the six measures used as features in our machine learning analysis. These five microstructure measures and VIX are created from dollar volume bars for the futures contracts specified in the Internet Appendix, Table A1. The sample period is July 2, 2012, to October 2, 2017. Abbreviations: UX1, VIX front month futures contract; VIX, Chicago Board Options Exchange volatility index; VPIN, volume-synchronized probability of informed trading.

- Volume-synchronized probability of informed trading is estimated as

$$VPIN_\tau = \frac{1}{W} \sum_{i=\tau-W+1}^{\tau} \frac{|V_i^S - V_i^B|}{V_i}, \tag{9}$$

  where volume is signed using the BVC method, $V_i^B = V_i Z\left[\frac{\Delta p_i}{\sigma_{\Delta p_i}}\right]$, $V_i^S = V_i - V_i^B$, and W is the lookback window size. See Easley et al. (2016) for additional details.
- We use VIX's front month futures (Bloomberg code: UX1 <Index>) tick-level trade data to represent VIX. For each bar, we take the price of the closest tick from the UX1 Index before that bar's timestamp as VIX's value.

Table 1 provides a correlation matrix of these six variables over our sample period, taking the vector of the values of the six variables for each asset, for each bar as an observation.[15] As is apparent, while the Roll and Roll impact measures are highly correlated, the other microstructures are not highly correlated with each other or with the Roll measures, suggesting that they may have very different properties for forecasting purposes. Although the Roll measures and the Amihud measures are positively correlated with VIX, the two microstructure measures most directly designed to measure the presence of informed trading, Kyle lambda and VPIN, are not correlated with VIX. Note that these correlations are all calculated based on dollar-volume bars. For VPIN, calculation in dollar-volume bars is a natural milieu, but the other variables were traditionally derived based on fixed time intervals, such as daily bars. A natural concern is that this specification may bias our results against finding significance for these types of variables. As part of our robustness testing, we also calculated all variables using hourly time bars and reran our analysis using this alternative data specification.

[15] These variables are also correlated across assets. In particular, the Roll and Amihud measures tend to be positively correlated, VPIN is correlated with both of them and all three of these measures tend to be positively correlated with VIX. These cross-asset correlations are not important in our initial analysis as it is done on a asset-by-asset basis.

We discuss these results in Section 4.5 but note upfront that we find slightly greater prediction accuracy using dollar-volume bars instead of time bars and that measures of feature importance are largely unchanged.

For the classification, we are interested in predicting the sign of the change in several important variables. Note that these labels are binary, as their value is either positive +1 or negative –1, reflecting that we are forecasting the sign of changes in the relevant variable. In particular, we label observations according to:

- the sign of change in bid-ask spread. The spread is computed via the Corwin-Schultz estimator:

$$S_\tau = \frac{2(e^{\alpha_{\tau-1}})}{1+e^{\alpha_\tau}},$$

$$\alpha_\tau = \frac{\sqrt{2\beta_\tau}-\sqrt{\beta_\tau}}{3-2\sqrt{2}} - \sqrt{\frac{\gamma_\tau}{3-2\sqrt{2}}},$$

$$\beta_\tau = E\left[\sum_{j=0}^{1}\left[\log\left[\frac{H_{\tau-j}}{L_{\tau-j}}\right]\right]^2\right], \tag{10}$$

$$\gamma_\tau = \left[\log\left[\frac{H_{\tau-1,\tau}}{L_{\tau-1,\tau}}\right]\right]^2,$$

  where $H_{\tau-j}$ and $L_{\tau-j}$ are the high and low prices at $\tau-j$, and $H_{\tau-1,\tau}$ and $L_{\tau-1,\tau}$ are the high and low prices over the 2 bars $(\tau-1,\tau)$. For a given forecasting horizon h, the label is then

$$\text{sign}[S_{\tau+h}-S_\tau], \tag{11}$$

  and effectively we are predicting whether the estimated spread will widen or narrow. Note that there is a window size variable in computing β_τ
- the sign of change in realized volatility, or simply

$$\text{sign}[\sigma_{\tau+h}-\sigma_\tau], \tag{12}$$

  where σ_τ is the realized volatility of onebar returns over a lookback window of size W. In this case we are predicting whether the realized volatility will go up or down.
- the sign of change in Jarque-Bera statistics of realized returns:

$$\text{sign}[JB[r_{\tau+h}]-JB[r_\tau]] \tag{13}$$

$$JB[r_\tau] = \frac{W}{6}\left(skew_\tau^2+\frac{1}{4}(Kurt_\tau-3)^2\right),$$

  where $Skew_\tau$ is the skewness and $Kurt_\tau$ is the kurtosis of realized returns over the lookback window of size W. This label can be viewed as a higher moment generalization of the realized return volatility above.

- the sign of change in the firstorder autocorrelation of realized returns:

$$\text{sign}[AR_{\tau+h} - AR_\tau] \tag{14}$$

$$AR_\tau = \text{corr}[r_\tau, r_{\tau-1}],$$

  where the correlation is evaluated over the returns of the past W bars.
- the sign of change in absolute skewness of realized returns:

$$\text{sign}[Skew_{\tau+h} - Skew_\tau]. \tag{15}$$

- the sign of change in kurtosis of realized returns:

$$\text{sign}[Kurt_{\tau+h} - Kurt_\tau]. \tag{16}$$

In the current analysis, we fix the forecast horizon h to be 250 bars ahead, which roughly corresponds to a week of trading. It is natural to question whether this week-long forecast window obscures market microstructure so we also considered a shorter forecast window. In section 6.1 we analyze forecast horizons of 25 and 50 bars and find similar results with slightly lower overall accuracy.

4. WithinAsset Results and Analysis

In this section, we restrict attention for each asset to its own features. In Section 5, we consider for each asset both its own features and features of a collection of other potentially important assets. We first set out the parameters of our random forest classification methodology. We then present the main results of this analysis, namely the feature importance of the microstructure variables and the prediction accuracy we obtain with them. This is followed by a sensitivity analysis in which we tune the parameters of the random forest and by various robustness checks, including a comparison between dollar-volume bar and time bar results and a comparison of our machine learning results with the results of logistic regression.

Our analysis uses a standard open-source machine learning software package, Scikit-learn (Pedregosa et al. 2011). We begin by specifying the configuration (hyperparameters, in machine learning parlance) of the random forest machine learning algorithm. For our analysis, we choose the default values for the random forest's hyperparameters[16]:

- number of trees (*n_estimators*) = 100;
- maximal features per split (*max_features*) = int($\sqrt{6}$) = 2;
- sample weight (*class_weight*) = inverse of total number of samples in the sample's class ('*balanced*')

[16] Scikit-learn's corresponding notations are in parentheses.

The number of trees is a parameter that controls how many decision trees the random forest contains. The maximal number of features here is the square root of the total number of features, a common choice for random forest. Sample weight is the weight one assigns to each sample in the training class, and we use a balanced approach to reduce the bias that can come from label imbalance. We report results from an unregularized random forest (i.e., one in which the decision trees are allowed to grow without limit). In Section 4.4, we report the results from using a regularized random forest to check that our results are stable and robust and to allay fears that the original random forest is overfit.

4.1 MDI results

We first examine feature importance using MDI. As a reminder, MDI is an in-sample method that is based on the explanatory power of each feature and gives rise to normalized values for feature importance (all positive and sum to one). Table 2, panels A through F, reports the MDI feature importance for each of the six predicted variables we consider in our analysis. Each row corresponds to a specific lookback window size, as indicated by the first column. Each entry is formulated as "mean MDI feature importance score" ± "MDI feature importance score standard deviation," where the mean and standard deviation are evaluated across all 87 instruments.[17] The highest importance is bolded for each window size.

To provide some intuition for how features contribute to in-sample explanation of labels, we provide in Figure 2 a scatter plot of predicted changes in spread for the ES1 index as a function of the VPIN and Roll measures. The random forest assigns a predicted increase or decrease in spread given any list of all the features. Plotting this assignment against feature vectors produces a plot in R^6 which we project to R^2in Figure 2. The right angle shape in that figure indicates a cut for spread predictions at just below 0.002 for the Roll measure and just above 0.05 for VPIN. The random forest predicts decreases in spread in the northwest quadrant and increases elsewhere.[18]

For bid-ask spread estimation, panel A of Table 2 shows that the Amihud measure has the highest feature importance, followed by the VPIN metric. Feature importance increases with the window size for Amihud, Kyle, and VPIN, but not for the Roll measures or for VIX. The Amihud measure is also the most important for absolute skewness prediction (panel E).

Panel B provides feature importance results for volatility prediction. Here we find mixed results depending on the window size. For both the shortest (25) and longest ($\geq$1,000) bars, VPIN dominates. Amihud is the most important if measured over a 250–500-bar window, while VIX prevails for the 50-bar

[17] These standard deviations are the empirical standard deviation of the MDI feature importance scores across the 87 instruments.

[18] This appears noisy in the figure because we are conditioning on only two of the six features. Otherwise, there would be sharper regions.

Table 2
MDI feature importance

Panel A: MDI feature importance for bid-ask spread prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN
25 bars	**0.181**±0.001	0.17±0.001	0.155±0.002	0.15±0.002	0.165±0.002	0.179±0.001
50 bars	**0.184**±0.001	0.17±0.001	0.153±0.002	0.15±0.002	0.167±0.001	0.177±0.0
250 bars	**0.194**±0.001	0.171±0.001	0.15±0.002	0.148±0.002	0.157±0.001	0.18±0.001
500 bars	**0.197**±0.001	0.171±0.001	0.149±0.002	0.148±0.002	0.151±0.001	0.184±0.001
1,000 bars	**0.198**±0.001	0.172±0.001	0.148±0.003	0.148±0.002	0.146±0.001	0.187±0.001
1,500 bars	**0.197**±0.001	0.173±0.001	0.148±0.003	0.148±0.002	0.145±0.001	0.19±0.001
2,000 bars	**0.197**±0.001	0.172±0.001	0.147±0.003	0.147±0.002	0.145±0.001	0.192±0.002

Panel B: MDI feature importance for realized volatility prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN
25 bars	0.169±0.003	0.157±0.002	0.149±0.003	0.157±0.003	0.178±0.003	**0.178**±0.004
50 bars	0.185±0.001	0.155±0.001	0.135±0.002	0.144±0.002	**0.205**±0.001	0.175±0.003
250 bars	**0.223**±0.002	0.148±0.001	0.098±0.002	0.119±0.002	0.22±0.002	0.192±0.002
500 bars	**0.234**±0.001	0.146±0.001	0.089±0.002	0.114±0.002	0.206±0.001	0.211±0.002
1,000 bars	0.234±0.002	0.143±0.002	0.085±0.002	0.117±0.003	0.18±0.002	**0.241**±0.004
1,500 bars	0.23±0.002	0.143±0.002	0.083±0.002	0.119±0.003	0.168±0.002	**0.257**±0.004
2,000 bars	0.226±0.002	0.142±0.002	0.082±0.002	0.119±0.003	0.165±0.002	**0.267**±0.004

Panel C: MDI feature importance for Jarque-Bera test prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN
25 bars	**0.184**±0.002	0.167±0.002	0.14±0.001	0.143±0.002	0.183±0.002	0.183±0.004
50 bars	0.19±0.001	0.161±0.001	0.13±0.001	0.135±0.001	**0.203**±0.001	0.181±0.003
250 bars	0.22±0.001	0.149±0.001	0.098±0.001	0.118±0.002	**0.22**±0.002	0.195±0.002
500 bars	**0.233**±0.002	0.147±0.001	0.091±0.002	0.117±0.002	0.206±0.001	0.206±0.002
1,000 bars	**0.24**±0.001	0.147±0.001	0.089±0.002	0.121±0.003	0.181±0.001	0.221±0.002
1,500 bars	**0.241**±0.002	0.147±0.001	0.088±0.002	0.124±0.003	0.168±0.001	0.232±0.002
2,000 bars	0.24±0.002	0.144±0.001	0.088±0.002	0.125±0.003	0.161±0.001	**0.241**±0.002

(*Continued*)

window (although VIX and VPIN are very similar for the 25-bar window as well). Feature importance for predicting the Jacques-Bera test in panel C also shows mixed results. Overall, Amihud is most important, but for some windows VIX and VPIN predominate. The Amihud measure also does well when using longer window sizes for sequential correlation prediction (panel D), while VIX dominates for shorter windows. Interestingly, the Roll measures, which might have been expected to do well with correlation change predictions, do not fare well. The results for kurtosis prediction again favor Amihud for long windows, but VIX and VPIN for shorter widows.

Overall, the data suggest that measured by in-sample performance the Amihud measure does best, with VPIN and VIX also having strong feature importance. The Kyle lambda and Roll measures are never the most important measure for predicting any of the six variables. However, all of the measures have similar MDI results for most of the variables. Perhaps most importantly, these measures all provide significant in-sample explanatory power even though they are simple measures designed for a simpler world.[19]

[19] It should be noted, however, that at each split in the trees we consider only two features. A feature that was never useful would have an MDI of zero, but one that sometimes is better than the single alternative it is compared with will have a nonzero MDI.

Table 2
(Continued)

Panel D: MDI feature importance for sequential correlation prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN
25 bars	0.183±0.002	0.163±0.002	0.142±0.002	0.15±0.002	**0.191**±0.003	0.171±0.002
50 bars	0.188±0.002	0.159±0.002	0.132±0.002	0.141±0.002	**0.206**±0.002	0.174±0.002
250 bars	0.215±0.002	0.146±0.001	0.109±0.002	0.131±0.003	**0.216**±0.002	0.184±0.001
500 bars	**0.228**±0.001	0.146±0.001	0.1±0.002	0.125±0.003	0.2±0.001	0.201±0.001
1,000 bars	**0.233**±0.002	0.146±0.001	0.099±0.002	0.133±0.003	0.176±0.001	0.213±0.002
1,500 bars	**0.234**±0.002	0.143±0.001	0.1±0.002	0.139±0.003	0.163±0.001	0.221±0.002
2,000 bars	**0.232**±0.002	0.143±0.001	0.1±0.003	0.143±0.003	0.156±0.001	0.224±0.002
Panel E: MDI feature importance for absolute skewness prediction						
25 bars	**0.181**±0.002	0.172±0.002	0.141±0.001	0.144±0.001	0.18±0.002	0.181±0.003
50 bars	0.189±0.001	0.169±0.001	0.132±0.001	0.136±0.001	**0.2**±0.001	0.174±0.002
250 bars	**0.219**±0.001	0.155±0.001	0.101±0.001	0.12±0.002	0.218±0.001	0.187±0.001
500 bars	**0.23**±0.001	0.152±0.001	0.096±0.002	0.12±0.002	0.201±0.001	0.201±0.001
1,000 bars	**0.237**±0.001	0.152±0.001	0.092±0.002	0.124±0.003	0.18±0.001	0.216±0.001
1,500 bars	**0.236**±0.001	0.153±0.001	0.091±0.002	0.127±0.003	0.168±0.001	0.225±0.002
2,000 bars	**0.238**±0.001	0.151±0.001	0.09±0.002	0.127±0.002	0.161±0.001	0.232±0.002
Panel F: MDI feature importance for kurtosis prediction						
25 bars	0.181±0.002	0.159±0.002	0.135±0.002	0.139±0.002	0.182±0.001	**0.205**±0.003
50 bars	0.188±0.001	0.157±0.001	0.127±0.001	0.133±0.001	**0.202**±0.001	0.193±0.002
250 bars	0.219±0.001	0.149±0.001	0.097±0.001	0.117±0.002	**0.221**±0.001	0.197±0.001
500 bars	**0.233**±0.001	0.147±0.001	0.091±0.002	0.117±0.002	0.206±0.001	0.207±0.002
1,000 bars	**0.24**±0.001	0.146±0.001	0.089±0.002	0.121±0.003	0.182±0.001	0.221±0.002
1,500 bars	**0.241**±0.002	0.146±0.001	0.088±0.002	0.124±0.003	0.168±0.001	0.232±0.002
2,000 bars	0.24±0.002	0.145±0.001	0.088±0.002	0.125±0.003	0.161±0.001	**0.241**±0.002

Panels A through F of Table 2 provide average MDI measures for each of the six labels we predict. For each label and each lookback window (given by rows of the panels), we provide the mean and empirical standard deviation of the MDI measure for each feature (given by the columns of the panels). For each window size, the largest importance is in bold. Abbreviations: MDI, mean decreased impurity; VIX, Chicago Board Options Exchange volatility index; VPIN, volume-synchronized probability of informed trading.

4.2 MDA results

We next turn to evaluating MDA feature importance. Table 3 summarizes the results of MDA feature importance for each predicted variable. In contrast with MDI, MDA is an out-of-sample method that captures the predictive power of each feature. Accordingly, MDA's outputs are not guaranteed to be positive (some features may actually be detrimental for forecasting purposes), nor are they normalized. As can be seen in the table, there are several entries with negative yet close to zero scores, and the interpretation is that they contribute little to the out-of-sample prediction despite the explanatory power they might have in sample. Every row corresponds to a specific lookback window, as indicated by the first column. Each cell is formulated as "mean MDA feature importance score" ± "MDA feature importance score standard deviation," where the mean and standard deviation are evaluated across all 87 instruments.[20] The highest importance is bolded for each window size. The last

[20] These standard deviations are the empirical standard deviation of the MDA feature importance scores across the 87 instruments.

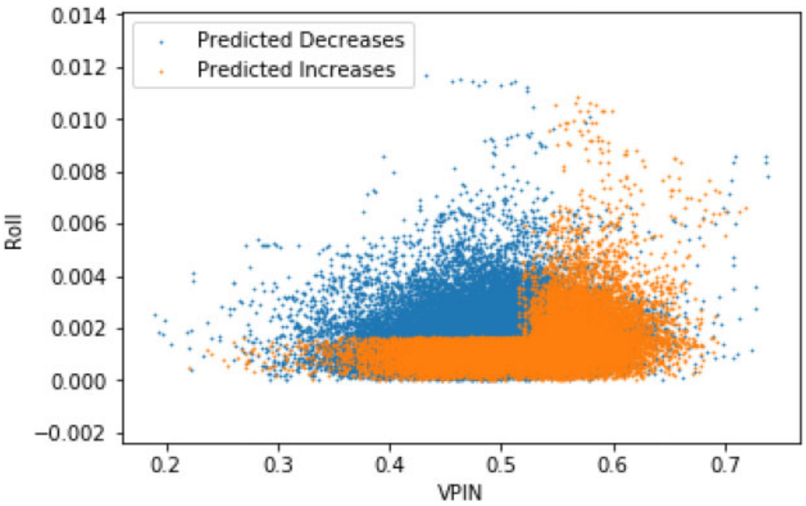


Figure 2
Scatter plot of predicted changes in spread as a function of the VPIN and Roll measures; the plot is for the ES1 Index with a lookback window of 25 bars
Figure 2 provides a scatter plot of predictions made by our random forest for the ES1 Index as a function of the values of the VPIN and Roll measures. The analysis was done using a lookback window of 25 bars and forecast horizon of 250 bars. Predicted increases are indicated in orange, and predicted decreases are indicated in blue. Abbreviation: VPIN, volume-synchronized probability of informed trading.

column summarizes the out-of-sample prediction accuracy averaged across all contracts.

To illustrate how the random forest works, we provide in Figure 3 a final decision tree that would be reached for Kurtosis using a 25-bar lookback window and a limit of three levels on the depth of the tree.[21] In a random forest, multiple decision trees are trained and aggregated. In this decision tree, each node is labeled with the variable used to split the node's sample, the Gini coefficient for the node's sample, the weighted-by-population proportion size of the sample, the number of weighted positive and negative values of the label in the node's sample, and finally an indicator of up or down that is determined by whether there are more weighted positive labels or more weighted negative labels in the node's sample. The directed edge coming out of each node is labeled true for the subsample sent to the left node and false for the subsample sent to the right node. So, for example, the entire sample is included at the first node, and it is split according to whether VPIN is below or above 0.518. The subsample with VPIN $\leq$0.518, which is of size 32,578, is assigned to the left node at level two of the tree, and in this subsample there are more negative than positive values for the sign of the change in kurtosis, so this node is labeled down. The rest of the sample (in which VPIN >0.518) is assigned to the right

[21] We limit our attention to three levels just for illustration as the size of the tree grows exponentially with depth

Table 3
MDA feature importance

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN	Accuracy
Panel A: MDA feature importance for bid-ask spread prediction							
25 bars	0.0033±0.00084	0.0042±0.00068	−0.0031±0.00108	0.0174±0.00154	0.0011±0.00059	**0.0248**±0.00142	0.4535
50 bars	0.0042±0.00094	0.0045±0.00078	−0.004±0.00088	0.0126±0.00142	0.0002±0.00056	**0.0167**±0.00117	0.4525
250 bars	0.0048±0.00125	0.0018±0.00088	−0.0047±0.00087	0.0031±0.00124	0.0021±0.0009	**0.0161**±0.00179	0.4572
500 bars	−0.0001±0.00094	−0.0003±0.00082	−0.003±0.00103	−0.003±0.00097	0.0031±0.00115	**0.0268**±0.00214	0.4587
1,000 bars	−0.002±0.00106	0.0007±0.00084	−0.002±0.00103	−0.0033±0.00108	0.0039±0.00138	**0.0198**±0.00166	0.4546
1,500 bars	−0.0023±0.00094	−0.0007±0.00083	−0.002±0.001	−0.0017±0.00109	0.0053±0.00133	**0.015**±0.00114	0.4513
2,000 bars	−0.0005±0.00093	0.0012±0.00093	0.0002±0.00099	−0.0025±0.00115	0.0058±0.00148	**0.0102**±0.0011	0.4498
Panel B: MDA feature importance for realized volatility prediction							
25 bars	0.0013±0.00289	0.0185±0.00138	0.0237±0.00211	**0.0558**±0.00288	−0.0006±0.00095	0.0531±0.00482	0.61
50 bars	0.0057±0.0011	0.0133±0.00125	0.019±0.00155	**0.0435**±0.00229	0.0004±0.00093	0.0402±0.00432	0.5813
250 bars	0.0163±0.00304	0.006±0.00163	0.0063±0.00148	**0.025**±0.00245	0.0002±0.0022	0.0172±0.0037	0.5493
500 bars	0.0133±0.00373	0.0005±0.00229	−0.0029±0.00149	0.0028±0.00231	−0.002±0.00251	**0.0307**±0.0044	0.5399
1,000 bars	0.0063±0.00315	0.0024±0.00265	−0.0029±0.00124	−0.0002±0.00251	0.01±0.00288	**0.0477**±0.0056	0.5578
1,500 bars	0.002±0.00377	0.004±0.00293	−0.0072±0.00187	−0.0034±0.0032	0.0101±0.00311	**0.0513**±0.00564	0.559
2,000 bars	0.0036±0.00386	0.002±0.00293	−0.0076±0.00191	−0.0111±0.00287	0.0187±0.00418	**0.056**±0.00544	0.5668
Panel C: MDA feature importance for Jarque-Bera test prediction							
25 bars	0.0027±0.00083	0.0219±0.00235	0.0041±0.0008	0.0095±0.00122	0.0001±0.00082	**0.0334**±0.00996	0.5406
50 bars	0.0008±0.00091	0.0132±0.00189	0.0032±0.00086	0.0083±0.00101	0.0001±0.00101	**0.0428**±0.00589	0.5416
250 bars	0.002±0.00264	−0.0006±0.0018	0.0004±0.00122	0.0086±0.00182	−0.0024±0.00192	**0.0411**±0.00424	0.5415
500 bars	−0.0043±0.00349	−0.002±0.00181	−0.0013±0.00145	0.002±0.002	−0.0011±0.00251	**0.0244**±0.0039	0.5232
1,000 bars	−0.0059±0.00319	−0.003±0.00257	−0.001±0.00167	**–0.0001**±0.00245	−0.0049±0.0027	−0.0019±0.00421	0.5066
1,500 bars	−0.0051±0.00382	−0.0049±0.00263	−0.0007±0.00182	−0.006±0.0028	−0.0042±0.00273	**0.0026**±0.00421	0.5051
2,000 bars	−0.0087±0.00331	−0.0042±0.00282	−0.0031±0.00224	−0.005±0.00342	−0.0025±0.00342	**0.0003**±0.00428	0.5074

(*Continued*)

Table 3
Continued

Panel D: MDA feature importance for sequential correlation prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN	Accuracy
25 bars	0.0048±0.00112	0.0042±0.00149	0.0158±0.00209	**0.0548**±0.00733	0.0012±0.00134	0.0053±0.00141	0.5401
50 bars	0.005±0.00135	0.0012±0.00072	0.0096±0.00147	**0.0433**±0.00716	0.0007±0.0011	0.0057±0.00177	0.5357
250 bars	0.0128±0.00268	0.0002±0.00136	0.0112±0.00204	**0.0391**±0.00671	−0.0013±0.00214	0.0021±0.0024	0.5394
500 bars	0.0081±0.00278	0.0017±0.00186	0.0069±0.00186	**0.023**±0.0046	0.0021±0.00254	0.0045±0.00289	0.5265
1,000 bars	0.0031±0.00287	−0.0015±0.00195	0.0013±0.00215	**0.0129**±0.00343	−0.0041±0.00208	−0.0008±0.00336	0.5173
1,500 bars	−0.0109±0.00825	−0.0032±0.00204	0.0019±0.0022	**0.0072**±0.00517	−0.0112±0.00429	−0.0051±0.00478	0.5113
2,000 bars	0.0006±0.00362	−0.0033±0.00212	0.0028±0.00204	**0.011**±0.00321	−0.006±0.00245	0.0023±0.00359	0.5113

Panel E: MDA feature importance for absolute skewness prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN	Accuracy
25 bars	0.0013±0.00074	**0.0374**±0.00494	0.0042±0.00123	0.0106±0.0015	−0.0009±0.0009	0.0328±0.00302	0.5447
50 bars	0.0012±0.00087	**0.0288**±0.0045	0.002±0.00072	0.0047±0.00094	−0.0002±0.00094	0.0276±0.00246	0.537
250 bars	0.0028±0.00264	0.0098±0.00248	0.0001±0.00117	0.0044±0.0018	0.0002±0.00213	**0.0179**±0.00315	0.5264
500 bars	−0.0005±0.00293	0.0073±0.00221	0.0006±0.0013	0.0011±0.00203	−0.0018±0.00217	**0.0092**±0.00344	0.5166
1,000 bars	−0.0033±0.00303	**–0.0012**±0.00269	−0.0037±0.00171	−0.0082±0.00272	−0.0085±0.00216	−0.008±0.00293	0.5024
1,500 bars	−0.0096±0.00398	−0.0041±0.00306	−0.0028±0.00165	**–0.0021**±0.00251	−0.0039±0.0024	−0.0084±0.00421	0.4995
2,000 bars	−0.0047±0.00349	−0.0026±0.00277	**–0.0019**±0.00166	−0.0052±0.00286	−0.0027±0.00255	−0.0025±0.00406	0.504

Panel F: MDA feature importance for kurtosis prediction

Window size	Amihud	Kyle lambda	Roll impact	Roll measure	VIX	VPIN	Accuracy
25 bars	0.0014±0.00071	0.0062±0.00121	0.005±0.00075	0.0114±0.00134	0.0001±0.00061	**0.0968**±0.00597	0.5694
50 bars	0.0012±0.00096	0.0047±0.00137	0.0032±0.00088	0.0094±0.00104	−0.0007±0.00094	**0.0844**±0.00482	0.5641
250 bars	0.0022±0.0024	−0.0016±0.00168	−0.0007±0.0012	0.0077±0.00198	−0.0023±0.00184	**0.0461**±0.00421	0.5444
500 bars	−0.0039±0.00315	−0.0035±0.00185	−0.0018±0.00156	0.0026±0.00208	−0.0025±0.00246	**0.0258**±0.00416	0.5227
1,000 bars	−0.0065±0.00308	−0.0025±0.00265	**–0.0008**±0.00179	−0.0006±0.00254	−0.0028±0.00242	−0.002±0.00441	0.5057
1,500 bars	−0.0059±0.00374	−0.0039±0.0029	−0.0024±0.00197	−0.0053±0.00311	−0.0031±0.0029	**0.0023**±0.00369	0.5046
2,000 bars	−0.0089±0.00317	−0.0043±0.00292	−0.0026±0.0022	−0.0068±0.00334	−0.0022±0.00349	**0.0009**±0.00411	0.5057

Panels A through F of Table 3 provide average MDA measures for each of the six labels we predict. For each label and each lookback window (given by rows of the panels), we provide the mean and empirical standard deviation of the MDA measure for each feature (given by the columns of the panels). Abbreviations: MDA, mean decreased accuracy; VIX, Chicago Board Options Exchange volatility index; VPIN, volume-synchronized probability of informed trading.

node at level two of the tree, and this node is labeled up as there are more positive values than negative values of the sign of the change in kurtosis.

Table 3, panel F shows that for kurtosis prediction VPIN is the most important feature (measured by MDA), the Roll measure is the second most important feature, and all of the other features are at least an order of magnitude less important. So it is not surprising that the first cut in the data is based on VPIN. Note that the decision tree associates high VPIN observations with increasing kurtosis. This is consistent with the intuition that high VPIN is a reflection of unbalanced trade indicating the possible presence of information and thus leading to fat tails in the distribution of returns. The next cuts, those made at level two to generate level three, are again made on VPIN. Now as the left and right nodes at level two are so unbalanced in their numbers of positive and negative values of the label, the left cuts yield subsamples both labeled down and the right cuts yield subsamples both labeled up. Finally, the cuts made at level three of the tree are made based on the Roll measure; no other feature is important enough to enter the decision tree by level three. By this point in the tree, it is difficult to interpret the splits as the samples have been split based on VPIN values, leading to unbalanced samples correlated with VPIN values, and so correlation between VPIN and the Roll measure also matters in the splitting.

For bid-ask spread prediction, VPIN has the highest feature importance for every widow size, and it has the highest importance for 5 or 6 window sizes for kurtosis prediction and Jarque-Bera test prediction. The Roll measure dominates for sequential correlation prediction. For realized volatility prediction, the Roll measure is better for shorter windows, with VPIN being a close second. Over longer lookback windows, however, VPIN again provides greater feature importance for realized volatility prediction, while the Roll measure generally contributed little to out-of-sample prediction. Interestingly, VIX has little out-of-sample prediction power regardless of the window size. The differential (and lower) performance of VIX relative to VPIN refutes the notion that VPIN is simply picking up volatility effects. Finally, for absolute skewness prediction, the feature importance results are mixed, with Kyle lambda, VPIN, Roll impact, and Roll measure each having greater importance for specific window sizes.

We interpret these results as providing support for the predictive power of microstructure measures that reflect frictions in the market. VPIN is generally the most important among these features at predicting variables that should be influenced by the presence of information-based trade: spread and measures of fat tails in the distribution of returns. The Roll measure is created from correlation in price changes, so it is not surprising that this measure has some explanatory power for serial correlation in returns. Finally, although we include VIX in our set of features, it is not intended to reflect microstructure frictions, so it is not surprising that it has little explanatory power for the variables we attempt to predict.

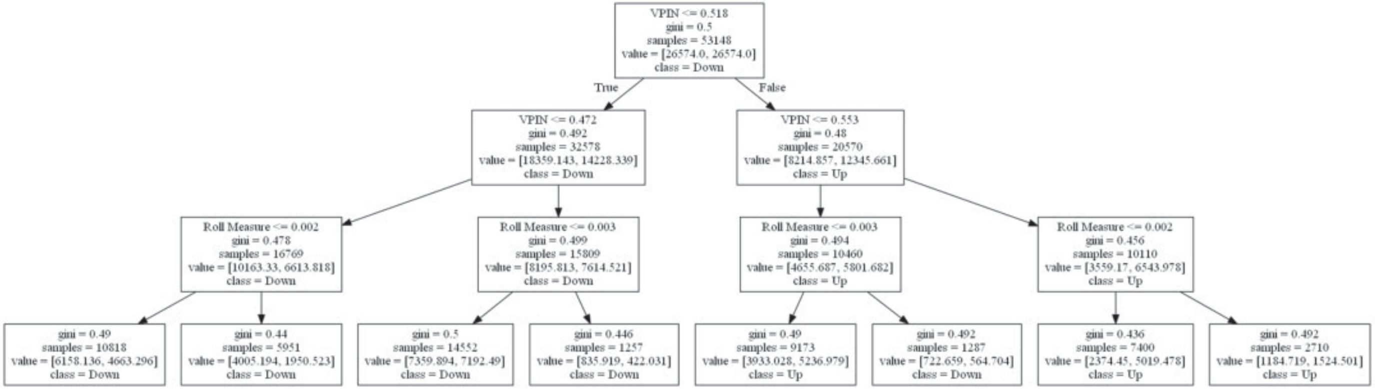


Figure 3
A three-level decision tree for predicting the sign of change in kurtosis, created using a 25-bar lookback window
Abbreviation: VPIN, volume-synchronized probability of informed trading.

4.3 Why are the MDI and MDA results so different?

Our finding that microstructural features with good explanatory power can have poor predictive power, and vice versa, may be surprising at first. The reason is, in the MDI feature importance analysis, each tree is fit on the entire sample, and the inference is conducted on the output of that fit. In the MDI approach, the trees are not exposed to out-of-sample, never-seen-before data points. As a result, MDI explains the past, even if each label was determined after the associated feature was observed. This is not dissimilar to the way inference is conducted in standard econometric approaches: A particular functional form is fit on an entire sample, and the estimated coefficients are subjected to a number of hypothesis tests. In a sense, MDI is an econometric-like feature important analysis, analogous to $p-$values of estimated betas. In an MDA analysis, the trees are not fit on the entirety of the data. Instead, each tree is fit on a fraction of the data, and after the fit has taken place, the tree is exposed to a never-before-seen sample. This type of K-fold cross-validation analysis, although commonplace in the machine learning literature, is less common in the market microstructure literature.

That MDI and MDA have such different results on microstructural features should give researchers pause. Most of the empirical research on market microstructure has been built on in-sample, MDI-like methods, absent of systematic cross-validation. When in-sample analyses are overfit to the entire sample, some features appear to be more important than they truly are for out-of-sample prediction. It is essential to recognize that an econometric forecasting specification, when fitted on the entire sample, leads to in-sample (MDI-like) results that may be overfit.

4.4 Sensitivity to hyperparameters, time periods, forecast windows, and additional features

As the random forest algorithm is highly nonparametric and can be tuned easily, one should ask about the stability of the results above with respect to tuning of the model parameters. After all, if the feature importance changes drastically when a random forest is constructed differently, then the results are not consistent. For this reason, we conduct multiple sensitivity tests for the feature's importance. All of our tests confirm that the feature importance score is consistent across different parameters, models, and time.

First, we tune two different model parameters intrinsic to all tree-based machine learning algorithms: maximal depth and minimal weight fraction per leaf. Changing these parameters transforms our unregularized random forest into a regularized random forest. The first parameter sets a depth threshold (the maximal number of sequential splits) for all decision trees that compose a random forest. For instance, if we set the maximal depth to be 5, then each tree cannot go beyond 5 sequential splits.[22] After adding this parameter

[22] In the scikit-learn library, this parameter is controlled by the argument "*max_depth,*" and we set *max_depth* = 5.

Table 4
MDA feature importance correlation between original random forest and adding max_depth = 5

Variable	25 bars	250 bars	50 bars	500 bars	1,000 bars	1,500 bars	2,000 bars
Kurtosis	0.998353	0.998687	0.997921	0.999611	0.999924	0.999904	0.999857
Bid-ask spread	0.994484	0.99634	0.997186	0.993873	0.996883	0.997981	0.997974
Return variance	0.999724	0.99978	0.999759	0.999471	0.99923	0.999054	0.999315
Sequential correlation	0.999838	0.999787	0.999747	0.999869	0.999924	0.999784	0.999847
Skewness	0.999728	0.999669	0.999655	0.999896	0.99991	0.999963	0.999893
Jarque-Bera test	0.999695	0.999104	0.999393	0.999652	0.999931	0.999869	0.999815

Table 4 provides correlation coefficients between the average MDA measures generated from a random forest with unbounded trees and those generated from a random forest with trees of maximal depth 5. These coefficients are provided for each variable we predict (the rows) and for each lookback window we consider (the columns). Abbreviation: MDA, mean decreased accuracy.

Table 5
MDA feature importance correlation between original random forest and adding min_weight_fraction_leaf = 0.01

Variable	25 bars	250 bars	50 bars	500 bars	1,000 bars	1,500 bars	2,000 bars
Kurtosis	0.99845	0.999174	0.998438	0.99976	0.999984	0.999891	0.999861
Bid-ask spread	0.996369	0.998045	0.998327	0.99646	0.998141	0.99838	0.998313
Return variance	0.999759	0.999895	0.999845	0.999754	0.999666	0.999218	0.999587
Sequential correlation	0.999918	0.999892	0.999823	0.999928	0.999867	0.999945	0.999906
Skewness	0.999677	0.999788	0.999632	0.999895	0.99996	0.999955	0.999968
Jarque-Bera test	0.999644	0.999318	0.999532	0.999794	0.999982	0.999899	0.999812

Table 5 provides correlation coefficients between the average MDA measures generated from a random forest with no restriction on minimal leaf size and those generated form a random forest with leafs restricted to have at least a sample fraction of 0.01. These coefficients are provided for each variable we predict (the rows) and for each lookback window we consider (the columns). Abbreviation: MDA, mean decreased accuracy.

to the random forest, we compute the MDA feature importance correlation between the original random forest and the regularized random forest across all 87 instruments. As shown in Table 4, correlation coefficients for every predicted variable and window size are virtually one. This indicates that the feature importance results are consistent and robust to changes in the tree depth hyperparameter.

The second parameter, which controls the least sample fraction on a leaf required to stop splitting, has a similar functionality.[23] A leaf is the name given to the node at the end of each branch or split. Restricting the minimal weight fraction per leaf essentially limits how big the tree can grow and thus limits the chances of overfitting. Again, we compute the MDA feature importance correlation between the unregularized and regularized random forest. These results are given in Table 5. Just like the case in Table 4, correlation coefficients for every predicted variable and window size are close to one, which further confirms the robustness of our feature importance analysis.

As our data are time series, a natural question is whether feature importance is stationary across time. For instance, is it possible that the Roll measure is good at

[23] In the scikit-learn library, this parameter is controlled by the argument "*min_weight_fraction_leaf.*"

predicting realized kurtosis when the market is less volatile, and not otherwise? To answer this, we run the MDA test presented in Section 4.2 on a yearly basis. More specifically, we split the data set into 5 parts in chronological order. As the total length of data in time is 5 years, each part covers a year-long period. For simplicity, we only show the results for the 250-bar lookback window in Table 6. It is evident that the feature importance, and especially the ranking, does not vary much across time, indicating that the feature importance is stationary. A related question is whether the feature importance is stationary across different instruments. This is partially proven by the small empirical standard deviations in the feature importance shown in Tables 2 and 3. In addition, we include a list of feature importance for kurtosis prediction per instrument with a lookback window fixed at 250 bars in Table A3 in the Internet Appendix.

Next, we ask about stability of our results with respect to the forecast window. All of our results are for a 250-bar forecast window, and it is worth asking how feature importance changes as the window size varies. Table 7 provides the correlation in MDA feature importance results between a 250-bar forecast window and a 50-bar forecast window. As the table shows, most entries are high (particularly for short lookback windows), which indicates that the results are reasonably stable across different forecast scales.

4.5 Dollar-volume bar versus time bar accuracy

The analysis above is based on a dollar-volume bar formulation. These bars have the desirable feature of aligning the sampling of data with the arrival of information, which seems an appropriate property for the high-frequency world characterizing futures trading. There are other bar types that could be used, and in this subsection we compare the accuracy using dollar-volume bars with the accuracy that results from using another popular bar method, the time bar method. A time bar is formed when the difference between the close tick and open tick's timestamps exceeds a predefined value. We formulate hourly time bars for all the futures instruments and apply the same cross-validation with the same random forest configuration. The results of out-of-sample prediction accuracy averaged over all lookback windows are given in Table 8.

Table 8 shows that accuracy results are very close for the two metrics. For four of the metrics, dollar-volume bars have higher accuracy, while for the Jarque-Bera test and bid-ask spread, time bars are slightly more accurate. This similarity is important for allaying fears that variables originally calculated over fixed time intervals may be distorted when cast in a volume-based metric. Additionally, we find that even though overall the time bar formulation has slightly lower accuracy, in many cases it gives rise to similar feature importance ranking as dollar-volume bars. For example, in Figure 4 we present the MDA feature importance for both bar methods for kurtosis prediction using a window size of 50 bars. The similarity in feature importance ranking is evident. Finally, Table 8 shows that regardless of how we measure bars, the accuracy of out-of-sample bid-ask spread prediction is low. Thus, prediction difficulties here

Table 6
Yearly feature importance with lookback window fixed at 250 bars

Panel A: MDA feature importance for bid-ask spread prediction

		Kyle	Roll	Roll			
Period	Amihud	lambda	impact	measure	VIX	VPIN	Accuracy
2012–2013	0.0013	0.0002	−0.0051	0.0014	0.0033	**0.0142**	0.4555
2013–2014	0.0034	−0.0012	−0.0067	−0.0008	0.0020	**0.0174**	0.4578
2014–2015	0.0035	−0.0003	−0.0040	0.0020	0.0009	**0.0142**	0.4480
2015–2016	0.0023	−0.0013	−0.0094	−0.0044	0.0038	**0.0128**	0.4549
2016–2017	0.0006	0.0008	−0.0038	−0.0005	0.0032	**0.0083**	0.4516
Panel B: MDA feature importance for realized volatility prediction							
2012–2013	0.0079	0.0001	0.0034	**0.0170**	−0.0015	0.0128	0.5479
2013–2014	0.0117	−0.0027	0.0048	0.0141	−0.0046	**0.0198**	0.5507
2014–2015	0.0001	−0.0003	0.0105	**0.0215**	−0.0139	−0.0003	0.5337
2015–2016	0.0098	0.0012	0.0088	**0.0174**	−0.0092	0.0058	0.5431
2016–2017	−0.0032	0.0023	0.0040	**0.0143**	−0.0105	0.0171	0.5526
Panel C: MDA feature importance for Jarque-Bera test prediction							
2012–2013	−0.0056	−0.0024	−0.0045	−0.0010	−0.0042	**0.0186**	0.5333
2013–2014	0.0040	0.0018	0.0025	0.0080	0.0038	**0.0372**	0.5556
2014–2015	−0.0130	−0.0031	0.0000	0.0033	−0.0111	**0.0131**	0.5319
2015–2016	−0.0049	−0.0039	−0.0045	−0.0008	−0.0162	**0.0269**	0.5382
2016–2017	−0.0155	0.0032	−0.0027	0.0006	−0.0125	**0.0057**	0.5326
Panel D: MDA feature importance for sequential correlation prediction							
2012–2013	−0.0039	−0.0012	0.0057	**0.0313**	−0.0114	−0.0047	0.5435
2013–2014	−0.0069	−0.0018	0.0005	**0.0245**	−0.0091	−0.0056	0.5370
2014–2015	−0.0023	−0.0035	−0.0086	**0.0092**	−0.0077	−0.0158	0.5332
2015–2016	0.0003	−0.0033	0.0054	**0.0200**	−0.0067	−0.0144	0.5332
2016–2017	−0.0033	−0.0055	−0.0054	**0.0194**	−0.0112	−0.0144	0.5322
Panel E: MDA feature importance for absolute skewness prediction							
2012–2013	−0.0015	−0.0035	−0.0030	−0.0021	0.0013	**0.0033**	0.5215
2013–2014	0.0052	0.0033	−0.0012	0.0008	−0.0043	**0.0090**	0.5309
2014–2015	−0.0049	**0.0033**	0.0013	0.0000	−0.0122	−0.0041	0.5137
2015–2016	−0.0028	0.0033	0.0000	−0.0017	−0.0066	**0.0105**	0.5208
2016–2017	−0.0083	**0.0027**	−0.0028	−0.0075	−0.0100	−0.0081	0.5165
Panel F: MDA feature importance for kurtosis prediction							
2012–2013	−0.0032	−0.0012	−0.0023	0.0004	−0.0019	**0.0215**	0.5355
2013–2014	0.0028	−0.0022	−0.0016	0.0031	0.0030	**0.0358**	0.5568
2014–2015	−0.0066	−0.0007	0.0021	0.0055	−0.0070	**0.0269**	0.5378
2015–2016	−0.0062	−0.0012	0.0001	0.0027	−0.0138	**0.0325**	0.5423
2016–2017	−0.0144	0.0004	−0.0010	0.0005	−0.0177	**0.0070**	0.5343

Table 6 provides five panels, one for each variable we predict, showing yearly average MDAs for each of the five years in our sample period. This analysis uses a 250-bar lookback window. In each row, the variable with the largest average MDA for that year is shown in bold type. Abbreviations: MDA, mean decreased accuracy; VIX, Chicago Board Options Exchange volatility index; VPIN, volume-synchronized probability of informed trading.

are not due to bar measurement issues. As noted earlier, we believe a more compelling explanation lies in the construction of this variable.

4.6 Logistic regression

Next, we consider a different classification model, namely, logistic regression for a model-based sensitivity test. Logistic regression models the logarithm of the odds of our two labels with a linear functional form. When the classification

Table 7
MDA feature importance correlation between a 50-bar forward window and a 250-bar forward window

	25	250	50	500	1,000	1,500	2,000
Variable	bars	bars	bars	bars	bars	bars	bars
Kurtosis	0.9984	0.9987	0.9685	0.5405	0.3060	0.8690	0.5939
Bid-ask spread	0.9951	0.9828	0.8229	0.9810	0.9612	0.8946	0.9459
Return variance	0.9911	0.9584	0.4131	0.9197	0.9801	0.9956	0.9864
Sequential correlation	0.9973	0.9979	0.8768	0.9415	0.9032	0.8292	0.9677
Skewness	0.9984	0.9947	0.6333	0.9140	−0.0871	0.3896	0.1941
Jarque-Bera test	0.9946	0.9945	0.8864	0.6871	0.2597	0.8482	0.7291

Table 7 provides the correlations in average MDA feature importance between a 50-bar forecast window and a 250-bar forecast window for each of our six features. Abbreviation: MDA, mean decreased accuracy.

Table 8
Performance comparison between dollar-volume bars and time bars

Variable	Average accuracy	
	DV bar	Time bar
Bid-ask spread	0.4539	**0.4699**
Jarque-Bera test	0.5237	**0.5269**
Kurtosis	**0.5309**	0.5304
Return variance	**0.5663**	0.5609
Sequential correlation	**0.5259**	0.5252
Skewness	**0.5187**	0.5142

Table 8 provides average prediction accuracy using dollar-volume bars and time bars for each of the six variables we predict. The average is taken over all lookback windows (25 bars to 2,000 bars) and all futures contracts. For each variable, the accuracy for the bar type that is greatest is given in bold type. Abbreviation: DV, dollar-volume.

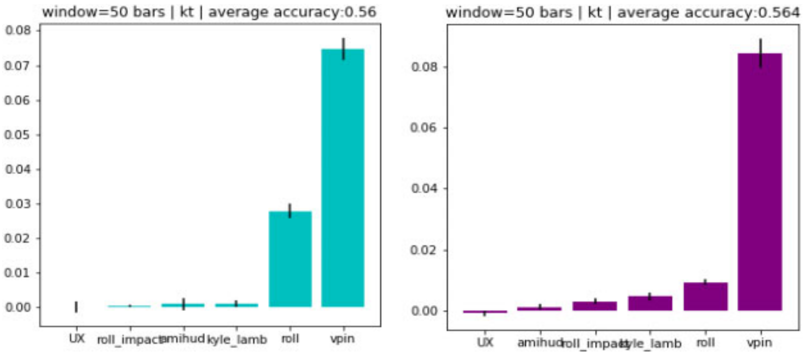


Figure 4
MDA feature importance for kurtosis prediction with window size =50 bars: left: time bar; right: dollar-volume bar
Figure 4 provides bar charts of average MDA feature importance for time bars on the left and dollar-volume bars on the right. The analysis was done with a lookback window of 50 bars. Abbreviations: MDI, mean decreased impurity; UX, VIX front month futures contract; VPIN, volume-synchronized probability of informed trading.

label is binary, denoted as $\{0,1\}$, the prediction probability for the two classes is given by

$$p(0|\vec{x}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}},\ p(1|\vec{x}) = 1 - p(0|\vec{x}), \tag{17}$$

Table 9
MDA feature importance correlation between logistic regression and random forest

Variable	25 bars	250 bars	50 bars	500 bars	1,000 bars	1,500 bars	2,000 bars
Kurtosis	0.9818	0.9779	0.9602	0.9489	−0.1058	0.5929	0.6807
Bid-ask spread	0.6478	0.5756	0.0999	0.1582	−0.0639	0.0177	0.1712
Return variance	0.9312	0.9223	0.8830	0.8264	0.9366	−0.3209	0.9492
Sequential correlation	0.9808	0.9929	0.9719	0.9268	0.9561	0.6887	0.9506
Skewness	0.9983	0.9809	0.8963	0.8058	0.6231	0.2737	0.7509
Jarque-Bera test	0.8611	0.9375	0.9589	0.9377	−0.0275	0.5359	0.7963

Table 9 provides correlations between MDA feature importance computed using the random forest and using logistic regression for each of our six features for each lookback window size ranging from 25 bars to 2,000 bars. Abbreviation: MDA, mean decreased accuracy.

where $\vec{x}$ is the feature vector and the coefficient vector $\vec{w}$ is obtained through a regularized maximum likelihood fit. For each sample, the prediction (+ or −) is the class label with the highest prediction probability. Logistic regression is another commonly used approach because of its simplicity and parametricity. It does not have an in-sample feature importance analysis, like MDI. Nonetheless, we can apply MDA feature importance and compare it to the random forest result. In Table 9, we present the MDA feature importance correlation between logistic regression and random forest.[24] In general, the correlation between two algorithms is high (>0.6), although when the lookback window size is large the correlation declines. Prediction accuracy with logistic regression is also similar to what we obtain with machine learning, with the logistic approach typically being slightly more accurate (Internet Appendix, Table A4).

To shed more light on the difference between the logistic regression and random forest approaches, we provide a scatter plot in Figure 5 for logistic regression's predicted bid-ask spread as a function of the two most important out-of-sample features, VPIN and the Roll measure. Figure 2 provided a scatter plot of predictions for the random forest approach as a function of these two variables. The two plots are similar, with the main difference being the shape of the decision boundary. Both plots illustrate predictions that are in line with our intuition; for example, higher VPIN leads to spread increasing.

We view the similarity of the results obtained with these two quite different approaches as further evidence that the microstructure frictions our features attempt to measure are real and have implications for the process of price adjustment. There is no apparent reason for why the logistic model with log odds given by a linear function of our microstructure features should fit the data reasonably well, but it does, as overall the prediction results are at least as strong as those we obtain with the hierarchical random forest approach. Of course, without having first done the random forest analysis, we would not have known that the logistic model offers a good specification. In other words, the random forest sets a nonparametric benchmark that a classical model can beat by

[24] Detailed results from the logistic approach are provided in Internet Appendix A (Table A4).

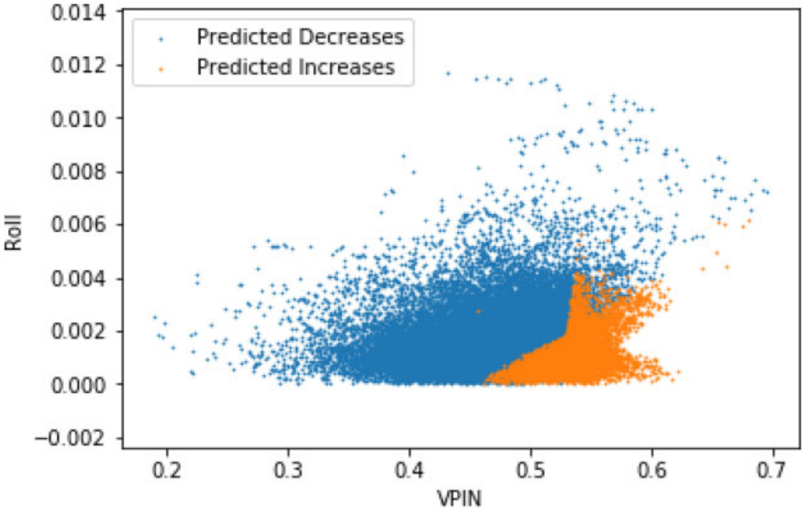


Figure 5
Scatter plot of predicted changes in spread as a function of the VPIN and Roll measures for logistic regression; the plot is for the ES1 Index with a lookback window of 25 bars
Figure 5 provides a scatter plot of predicted changes in the bid-ask spread as a function of the VPIN and Roll measures. The analysis was done using a lookback window of 25 bars. Predicted increases are indicated in blue, and predicted decreases are indicated in orange. Abbreviation: VPIN, volume-synchronized probability of informed trading.

injecting structural information into the forecasting problem. This exemplifies our view that machine learning algorithms do not replace classical methods, but rather complement the use of those classical methods by de-coupling the search for specification from the search for important variables.

It is not surprising that logistic regression predicts approximately as well as random forest when attention is restricted to a small collection of own-asset features (five own-asset market microstructure measures and VIX). The power of the random forest approach is most apparent when many features are considered. This is the case in the next section, where we expand the set of features to include cross-asset features.

5. Cross-Futures Effects

The market microstructure literature typically considers each asset in isolation. Thus far, we followed this approach by examining how the various microstructure measures perform on each individual futures contract and then aggregating across all 87 contracts to present our results. Yet, as noted in the introduction, cross-asset activity (and particularly cross-asset market making) is now the norm, suggesting that there could be important cross effects of microstructure measures such as Amihud or VPIN in one asset on other assets.

There is no market microstructure theory of how these cross-asset effects should, or even could, occur, and there are many plausible alternatives. For example, perhaps microstructure measures in the massive E-mini are predictive of future conditions in assets other than the E-mini, as there could be a spillover of trade or information from the E-mini. Or perhaps there are correlations between futures based on similar assets (metals, for example), but not between these futures and futures on financial assets. Or perhaps trade or information from financial futures spills over to other nonfinancials. As theory provides no guidance for which variables should matter, or how they should matter, modern machine learning techniques seem ideally suited to an initial exploration of these questions. That is the challenge we take up in this section.

We repeat the random forest approach outlined in Sections 3 and 4, but now include cross-asset features. This analysis is done using time bars, as discussed in Section 4.5, in order to synchronize observations across markets.[25] For each of our 87 futures contracts, we include its own microstructure features and the microstructure features of a shared subset of 15 futures chosen to represent actively traded commodity, equity, currency, and fixed-income futures.[26] The list of cross-asset futures used is given in Table 10. With cross-asset features included, there are now 81 features for each contract. Given the challenges revealed by our earlier analysis of the constructed bid-ask spread variable, we omit this variable and concentrate on predicting the five other labels. We present results using a 250-bar forecast window.[27]

Figures 6–10 present average (over the 87 futures contracts) MDA results for each of the variables we want to predict: realized volatility, Jarque-Berra statistic, serial correlation of returns, skewness, and kurtosis. In these figures, we show the MDA values only for features with positive values; ones not shown are negative. Table 11 encapsulates these results, showing for each label the top 10 features in terms of predictive power and whether they are own-asset or cross-asset features. We emphasize three general findings: (1) own-asset features continue to play an important, but not always the most important, role in out-of-sample prediction; (2) cross-market features are useful in prediction, with a small number of features exerting an influence across multiple predicted variables; and (3) many of the 81 features have zero or negative MDAs.

Turning to the first result, we find that own-asset Amihud, VPIN, and Roll have the greatest influence, with each of these in the top three features for

[25] We synchronize time bars across all our futures contracts so that there is no potential issue of using measures based on trading in the future in one contract to predict price and liquidity dynamics for another contract.

[26] For each asset class, we selected the most liquid futures within the class to serve as proxies for the class. In the currency class, we selected the dollar, yen, and euro contracts. In the equity class, we selected futures on the S&P, NASDAQ, Euro Stoxx, and Nikkei.

[27] As we did in the analysis using only own microstructure features, we computed results using various window sizes and have reported results for the 250-bar window we focused on previously. Shorter lookback windows yield generally greater prediction accuracy, perhaps because with longer windows there is more noise in the system.

Table 10
Cross-asset futures

- Commodity
  - BO1_Comdty, soybean oil
  - CL1_Comdty, crude oil
  - GC1_Comdty, gold
  - LC1_Comdty, live cattle
  - QS1_Comdty, gasoil
  - SM1_Comdty, soybean meal
- Equity index
  - FA1_Index, S&P
  - NK1_Index, Nikkei
  - NQ1_Index, NASDAQ
  - VG1_Index, Euro Stoxx
- Currency
  - DX1_Curncy, Dollar
  - EC1_Curncy, Euro
  - JY1_Curncy, Japanese Yen
- Fixed income
  - OE1_Comdty, Euro-Bobl 5-Year
  - TY1_Comdty, US Treasury 10-Year

Table 10 provides a list of the 15 cross-asset features used in the cross-asset analysis of Section 5. Each of these features is included as a feature, along with its own-asset features, in the random forest analysis.

four of five predicted variables. Unlike most cross-asset features, own-asset Amihud, Roll, and VPIN always have positive MDAs. However, the exact influence of each feature differs from what we found when creating random forests using only own-asset features. For example, own-asset VPIN is no longer the dominant feature for prediction; VPIN is the top feature for one label, while Roll is the top feature for three labels. Thus, adding cross-asset features influences the role played by own-asset features. We conjecture that this effect may be due, in part, to market microstructure measures in related assets also picking up some of the signal created by information-based trade.[28]

Second, our analysis of the cross-asset features shows the differential importance played by particular contracts. For example, microstructure measures based on trade in financial futures typically have high MDAs; that is, they are valuable in out-of-sample prediction of other futures contracts. Microstructure features based on trade in the 10-Year U.S. Treasury (TY1), either the own VPIN measure or the own Roll measure features, have positive, highly ranked MDAs in predicting other price processes. Similarly, the Euro-bobl future (OE1) consistently ranks as an important feature across assets, and the OE1-VPIN is actually the most important feature for the MDA results related to changes in the Jarque-Bera statistic.[29] In addition, either the own Amihud or own Roll measure features based on trade in the U.S. Dollar Index (DX1) also have high MDAs. Perhaps less expected is that features derived from trade

[28] In Internet Appendix C, we provide histograms of the correlation distributions for (Roll, Amihud), (Roll, VPIN), and (Amihud, VPIN).

[29] The Euro-bobl contract is an interest rate future on a notional medium-term German government bond (Bundesobligation) with a remaining term to maturity of between 4.5 and 5.5 years.

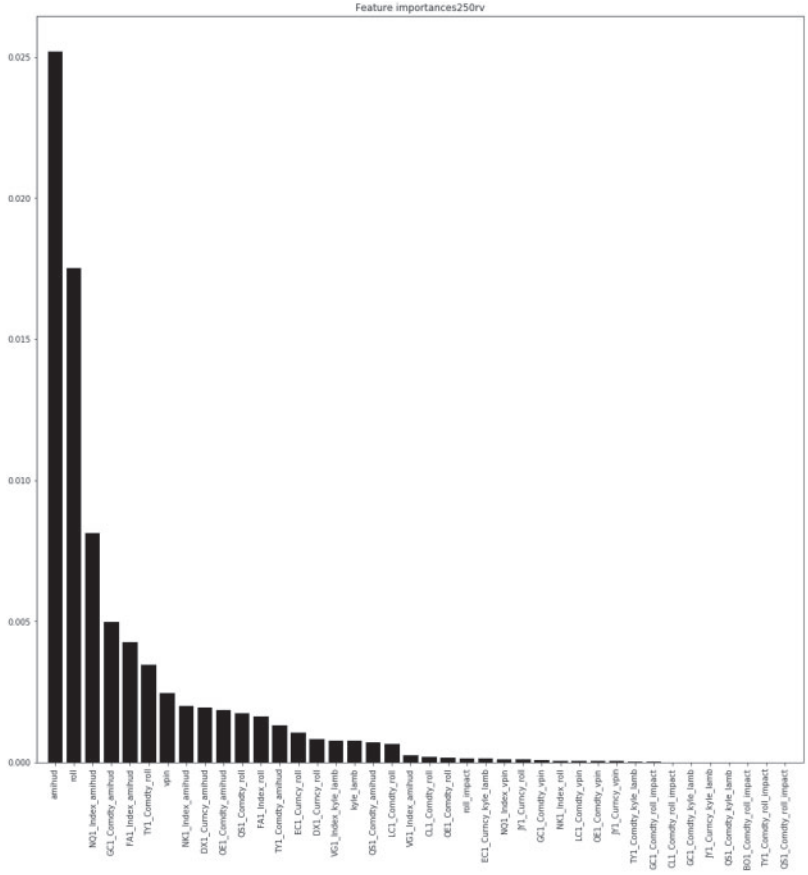


Figure 6
MDA results for predicting the sign of realized volatility using a 250–time bar lookback window

in the E-mini are not particularly important. MDA measures for these features are typically positive, but they are not among the most important features in predicting any of our variables. Features based on trade in the NASDAQ futures contract are typically more important than those based on trade in the E-mini.

Finally, our third result, that many cross effects have zero or negative MDAs, should not be unexpected. The information gleaned from microstructure features of relatively unconnected or uncorrelated assets should have negligible impacts on the price processes of other such assets. What is useful to recognize is that the random forest machine learning approach is capable of discerning which features matter and which do not. Such discernment may be helpful for researchers hoping to build better models of market behavior.

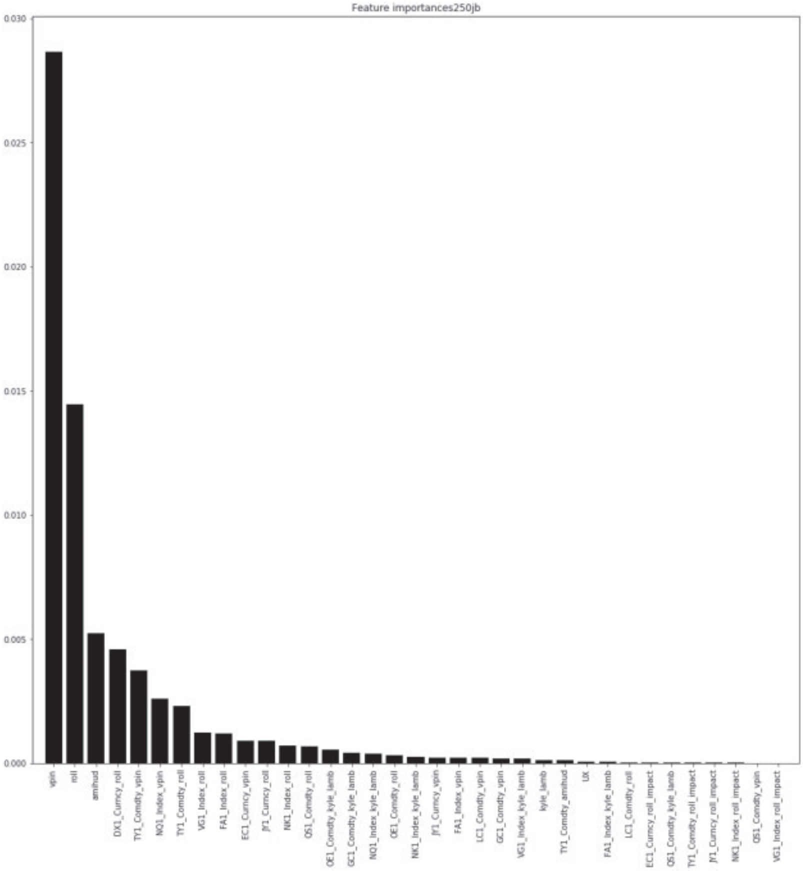


Figure 7
MDA results for predicting signs of change in the J-B test using a 250–time bar lookback window

6. Prediction Accuracy

Our analysis is focused primarily on evaluating the usefulness of standard microstructure measures in modern financial markets. The primary criterion we use for this purpose is MDA, which is based on the effect of including or excluding a particular feature (for us, a microstructure measure) on prediction accuracy. In this section, we discuss these accuracy results, their usefulness, and their sensitivity to various specifications.

The overall accuracy levels in Table 3 (last column of each panel) suggest that our machine learning algorithm is capturing something of value. For binary financial time series classification, a classifier often gives accuracy around 0.5. This standard inability to do better than random guessing is consistent with the efficient market hypothesis: for liquid markets, the market is efficient most of

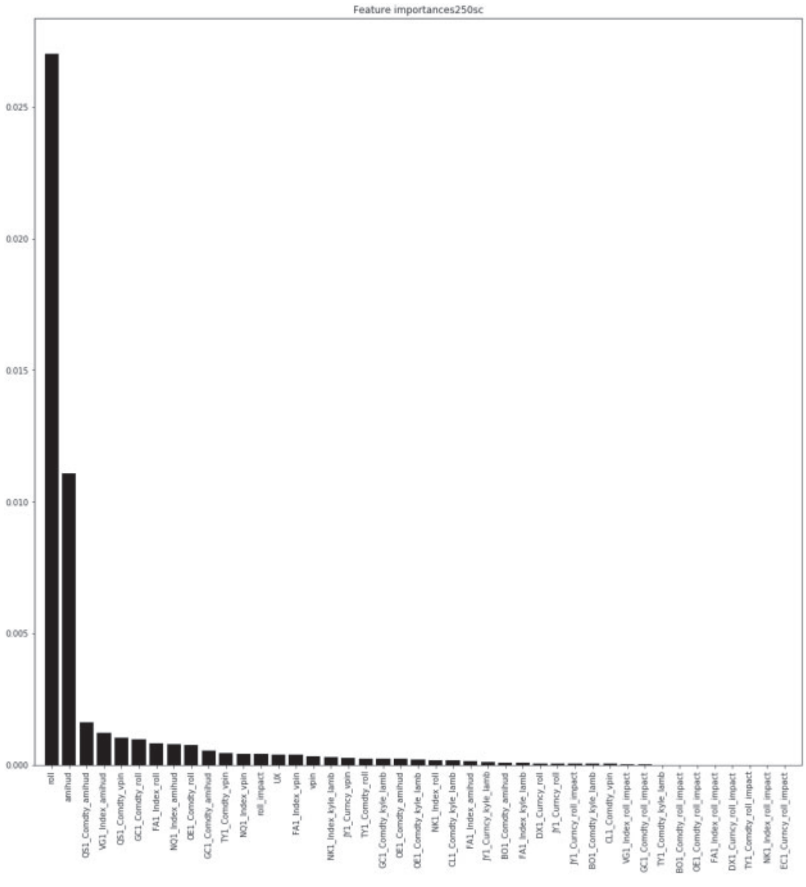


Figure 8
MDA results for predicting signs of change in serial correlation of returns using a 250–time bar lookback window

the time and acts like a random walk. Thus, anything above 0.5 can be viewed as capturing a potential inefficiency of the market and so is a positive result. With the exception of the bid-ask spread estimation, our out-of-sample accuracy levels reach highs ranging from 0.54 to 0.61 (depending on the lookback windows), which by financial machine learning standards is very high.[30] The bid-ask spread accuracy is not as good. We conjecture that this is due to the lack of an observable bid-ask spread in futures; we impute one using the Corwin-Schultz estimator. It may be that the errors in the estimation as applied to futures make prediction via the random forest methodology ineffective.

[30] For example, see Krauss, Do, and Huck (2017), who in a similar binary classification problem obtain accuracy levels between 0.50 and 0.55.

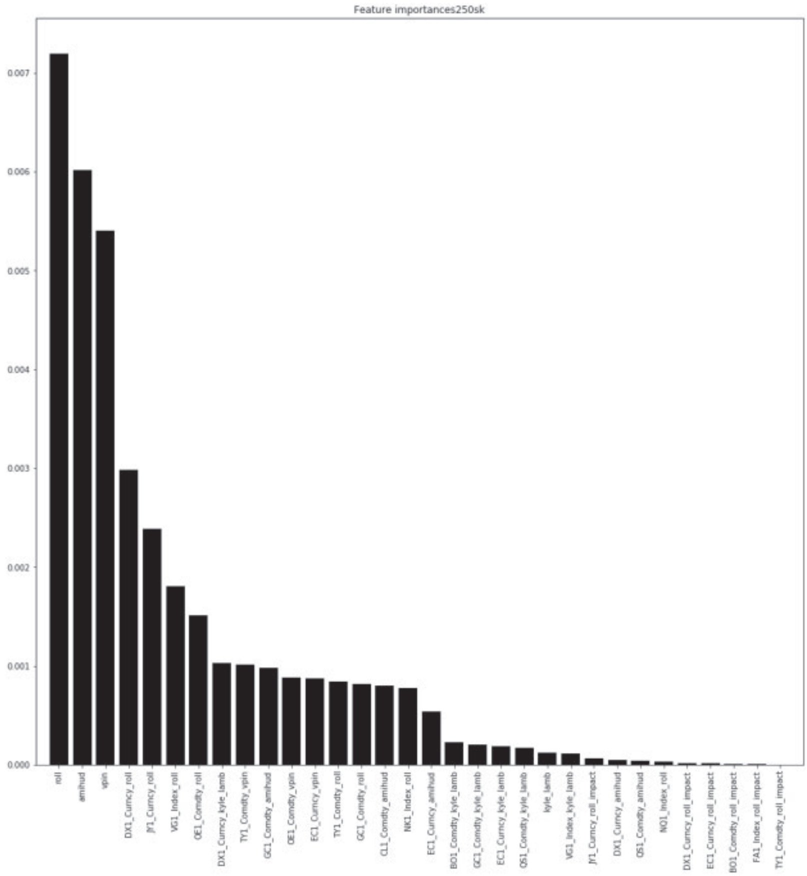


Figure 9
MDA results for predicting signs of change in skewness of returns using a 250–time bar lookback window

These predictions are made once for every bar for every day, and as a result, even a small improvement in predictive power can lead to a substantial improvement in returns to a strategy employing these predictions. For example, consider a strategy that produces n IID bets per year, where the outcome X_i of a bet $i \in [1,n]$ is a profit $\pi > 0$ with probability $P[X_i = \pi] = p$ and a loss $-\pi$ with probability $P[X_i = -\pi] = 1 - p$. Here p represents the precision of our binary classifier, where a positive value results from a correctly predicted outcome (predictive accuracy) and a negative value comes from an incorrectly predicted outcome. As the outcomes $\{X_i\}_{1,\dots,n}$ are independent, the relevant statistics to use in evaluating these bets are the expected moments per bet. The expected profit from one bet is $E[X_i] = \pi(2p-1)$, and its variance is $V[X_i] = 4\pi^p(1-p)$.

To provide some intuition about how valuable this strategy can be, we compute the Sharpe ratio for n IID bets per year employing this strategy. The

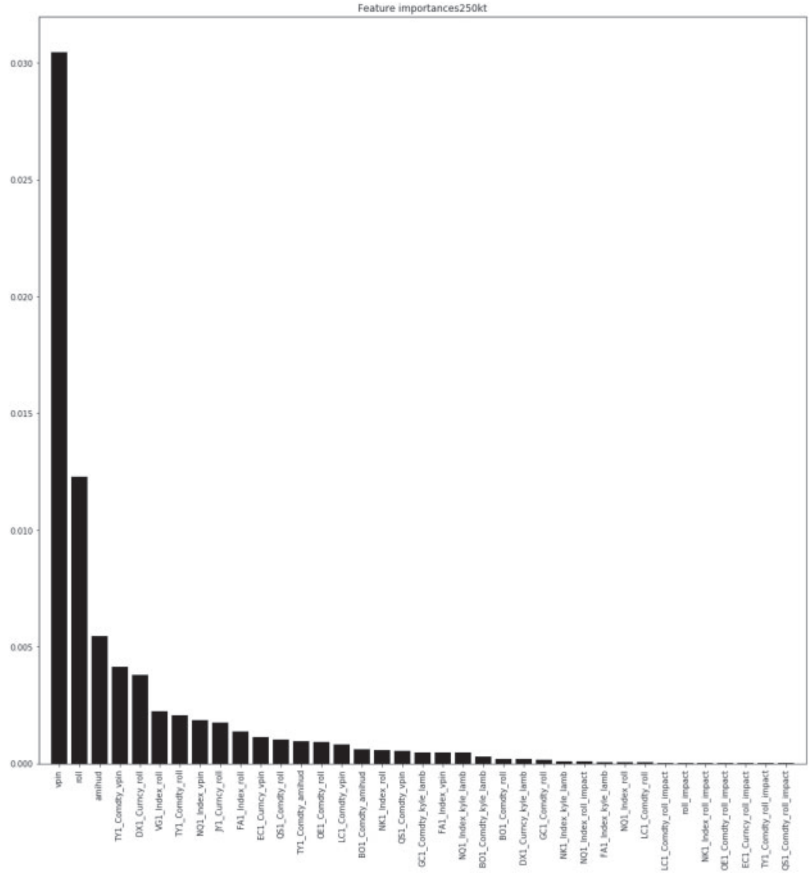


Figure 10
MDA results for predicting signs of change in kurtosis of returns using a 250–time bar lookback window

annualized Sharpe ratio (θ) of these bets in a zero risk-free-rate environment is[31]

$$\theta[p,n] = \frac{n\mathrm{E}[X_i]}{\sqrt{n\mathrm{V}[X_i]}} = \underbrace{\frac{2p-1}{2\sqrt{p(1-p)}}}\sqrt{n}. \tag{18}$$

On average, we have 50 bars per day, so the number of bets per year, n, is approximately 13,000. At $\approx .5$ we obtain $\theta[p,n] \approx 0$. At $p \approx .52$, we obtain $\theta[p,n] \approx 2.04$, which is not only statistically significant at a 5% significance level, but is also considered a sizeable Sharpe ratio among practitioners. So the

[31] Note that profit per trade π cancels out of the above equation, because the payouts are symmetric. Just as in the Gaussian case, $\theta[p,n]$ can be understood as a rescaled t-value.

Table 11
Most important features in cross-asset features analysis

Factors by MDA Importance	Δ Sign Realized Volatility	Δ Sign Jarque-Bera Test	Δ Sign Serial Correlation	Δ Sign Skewness	Δ Sign Kurtosis
1	Own	Cross	Own	Own	Own
2	Own	Own	Own	Own	Own
3	Own	Cross	Cross	Own	Own
4	Own	Own	Cross	Cross	Cross
5	Cross	Cross	Cross	Cross	Cross
6	Cross	Cross	Cross	Cross	Cross
7	Cross	Cross	Cross	Cross	Cross
8	Cross	Cross	Cross	Cross	Cross
9	Cross	Cross	Cross	Cross	Cross
10	Cross	Cross	Cross	Cross	Cross

Table 11 provides a classification of the most important features, ranked by average MDA importance, for each of the five variables we predict in our cross-asset analysis. Features are characterized as own asset if the feature is one of the six own-asset features we consider or cross-asset if it is one of the 15 cross-asset features we consider. Abbreviation: MDA, mean decreased accuracy.

prediction accuracies presented in Table 3 and the apparently small differences between prediction accuracies of the random forest and logistic regression approaches using cross-asset features in Table 14 matter for investment efficiency.[32]

6.1 Alternative specifications and accuracy measures

Prediction accuracy depends on both the variables we want to predict and the features used to predict these variables. The six variables we are interested in predicting are fixed, but the horizon at which we predict them is not fixed. Table 3 provides accuracy results for a forecast horizon of 250 bars. Table 12 provides average accuracy results for both the random forest and the logistic regression for shorter forecast horizons of 25 and 50 bars. Generally, for these shorter horizons, we have similar but slightly lower accuracy results. Note, however, that even for these short horizons our prediction accuracy results are above 0.5 for all labels other than bid-ask spread.

Second, because our data are time series, it is possible that additional features involving lagged values of returns or return volatility could change the accuracy results. To address this issue, we included these variables and reran our random forest algorithm. The variables we included at each bar t are the returns from bar t-h to t for each of h = 25, 50, 250, and 1,000; that is, returns over the last 25, 50, 250, and 1,000 bars and lagged realized bar volatility with 250-bar and 1,000-bar horizons. The results are presented in Table 13. Comparing these results with column 1 of Table 8 shows that inclusion of these lagged return and volatility features improves prediction accuracy, indicating that even with our purging of data between the training and test sets there is some residual time series structure

[32] A trader who wanted to use this approach might want to consider a different classification of the variables we predict and consider only large positive and large negative changes, consider individual futures rather the average we focus on, and tune the random forest to generate the sharpest possible predictions for those assets.

Table 12
Comparison of average accuracy for logistic regression and random forest for shorter forecast horizons

Panel A: 50-bar forecast horizon

Variable	Average accuracy	
	Logistic regression	Random forest
Bid-ask spread	0.4508	0.4425
Jarque-Bera test	0.5416	0.5366
Kurtosis	0.5549	0.5477
Return variance	0.5915	0.5821
Sequential correlation	0.5287	0.5255
Skewness	0.5236	0.5192

Panel B: 25-bar forecast horizon

Variable	Logistic regression	Random forest
Bid-ask spread	0.4495	0.4435
Jarque-Bera test	0.5383	0.5350
Kurtosis	0.5486	0.5432
Return variance	0.5737	0.5667
Sequential correlation	0.5202	0.5170
Skewness	0.5194	0.5173

Table 12 provides average accuracy results for each of the six variables we predict for both the random forest approach and logistic regression using shorter forecast horizons of 50 bars in panel A and 25 bars in panel B.

Table 13
Average prediction accuracy for random forest and logistic regression with lagged returns and return volatility included as features

Variable	Average accuracy	
	Logistic regression	Random forest
Bid-ask spread	0.4685	0.4629
Jarque-Bera test	0.5534	0.5440
Kurtosis	0.5685	0.5563
Return variance	0.6341	0.6152
Sequential correlation	0.5447	0.5387
Skewness	0.5315	0.5261

Table 13 provides average prediction accuracy results with four lagged returns (with lags of 25, 50, 250 and 1,000 bars) and two lagged realized bar volatilities (with lags of 250 and 1,000 bars) for each of the six variables we predict for each of the random forest approach and logistic regression.

in the data.[33] However, inclusion of lagged returns and volatility does not change our results about own and cross-asset feature importance significantly.

Third, inclusion of cross-asset features not only affects which microstructure features are important (as discussed in Section 6) but also affects overall accuracy. Summary accuracy results are presented in Table 14. The average accuracy results for this more complex random forest estimation range from 0.5169 to 0.5743, demonstrating as before that successful prediction is attainable with machine learning. What is more intriguing to consider is how these results compare to within-asset prediction accuracy levels. Comparison of

[33] This analysis of the effect of lagged returns is based on dollar-volume bars, and the results in Table 13 are averaged over all lookback windows and futures contracts, making the appropriate comparison column 1 of Table 8, which also presents average results using dollar-volume bars.

Table 14
Average prediction accuracy for random forest and logistic regression with cross-asset features included

Variable	Average accuracy	
	Logistic regression	Random forest
Jarque-Bera test	0.5328	0.5353
Kurtosis	0.5388	0.5400
Return variance	0.5581	0.5743
Sequential correlation	0.5180	0.5201
Skewness	0.5165	0.5169

Table 14 provides average prediction accuracy using the random forest approach and logistic regression for each of the five variables we predict. These accuracy measures are averaged over all lookback windows (25 bars to 2,000 bars) and all futures contracts.

the accuracy results in column 2 of Table 8 and those in column 2 of Table 14 shows an improvement in accuracy for all but one variable, resulting from inclusion of the cross-asset features.[34] Table 14 also provides a comparison of the accuracy of logistic regression prediction and random forest predictions when including the cross-asset features. Unlike the results obtained with a smaller number of own-asset features (Internet Appendix Table A4), we see that now the random forest approach consistently yields better predictions. We conjecture that this results from the flexibility of the random forest, particularly its nonlinear capabilities, in dealing with a large number of noisy explanatory variables.

Finally, we consider the robustness of our results with respect to an alternative accuracy measure: ROC-AUC. This overall accuracy measure can be interpreted as the probability that the procedure (random forest or logistic regression) will rank higher a randomly drawn positive than a randomly drawn negative. ROC-AUC has the advantage that it is not biased by class skewness. The accuracy results with this measure are reported in Table 15. They are similar to those we report with our prediction accuracy measure (the fraction of correct predictions).[35]

7. Conclusion and Future Directions

In this study, we attempted to shed light on the importance of various microstructure features for explanatory and forecasting purposes. The six variables we wish to explain and predict are highly relevant to market makers, portfolio managers, regulators, and researchers: bid-ask spread, realized volatility, normality, skewness, kurtosis and serial correlation. We apply

[34] This analysis of the effect of including cross-asset features is based on time bars, and the results in Table 14 are averaged over all lookback windows and futures contracts, making the appropriate comparison column 2 of Table 8, which also presents average results using time bars.

[35] There are also other measures of accuracy, in particular, ones that treat true and false positives or negatives asymmetrically. Making use of our approach in an investment strategy such alternative measures (such as precision, recall, or F1 score) may be more useful than the simple accuracy measure we use here. Our goal is more generic and not tied to a particular investment strategy, so we treat errors symmetrically.

Table 15
Results for ROC-AUC accuracy measure

- 250-bar forecast horizon, idiosyncratic microstructure features

Variable	Average accuracy	
	Logistic regression	Random forest
Bid-ask spread	0.5583	0.5471
Jarque-Bera test	0.5592	0.5484
Kurtosis	0.5830	0.5710
Return variance	0.6506	0.6215
Sequential correlation	0.5714	0.5531
Skewness	0.5447	0.5352

- 250–time bar forecast horizon, cross-asset microstructure features

Variable	Average accuracy	
Jarque-Bera test	0.5491	0.5486
Kurtosis	0.5664	0.5510
Return variance	0.6025	0.6305
Sequential correlation	0.5284	0.5406
Skewness	0.5289	0.5337

Table 15 provides average prediction accuracy results using the AUC measure of accuracy. Results are averages over all lookback windows (25 bars to 2,000 bars) and all futures contracts. Results are provided for both the random forest approach and logistic regression. Abbreviation: AUC, area under the curve.

machine learning methods in order to capture the complexity inherent to high-frequency data, without concerning ourselves at this point with determining a parametric structure to characterize the complex relationship between variables. We provide clear evidence that some extant microstructure variables have value for predicting the new dynamics of market behavior. At the same time, however, we find that other popular microstructure variables can have high explanatory power (in-sample) and yet fail to provide forecasting power (out-of-sample).

We believe these findings have important implications for future microstructure research. Foremost among these implications is good news: our results clearly show that market frictions continue to play an important role in affecting market dynamics and that extant microstructure measures capture (to varying extents) these dynamic effects. Thus, despite the complexity of current markets, frictions such as asymmetric information, illiquidity arising from constraints on market maker risk-bearing, or endogenous patterns arising from algorithms programmed to hide in particular market structures all continue to affect price dynamics as predicted by microstructure research. More good news is that the efficacy of these microstructure variables in capturing these effects appears to be remarkably robust. Our out-of-sample forecasting results are virtually the same whether we use time clocks or volume clocks, shorter samples or longer, regularized or unregularized forests, even simple logistic models versus hierarchal machine learning—the rankings of which variables matter most generally stay the same. These findings should be helpful in thinking about the type of models (and measures) we need to work on to capture better market dynamics.

There are other implications to consider as well. As most empirical research in the market microstructure literature follows an in-sample procedure, without out-of-sample cross-validation, it is possible that some established empirical results are artificial. To determine this, however, requires more extensive study and new empirical analytics. The machine learning approach taken here is one such direction, but there are many new approaches that seem well suited to analyses of complex market structures.

At a more fundamental level, the high out-of-sample accuracy we have achieved appears to indicate that markets are less efficient than is generally believed. For microstructure researchers, efficiency has long been a problematic concept; over short intervals, prices are not random walks, and even the concept of a price is tricky given that it may differ depending on whether you want large or small amounts, are a buyer or a seller, etc. Our findings here, however, are more concrete and troubling. Using machine learning techniques, successful forecasting of price process dynamics using simply past data on market microstructure features is both feasible and accurate. From a practical perspective, this suggests increased research on ways to exploit this information in profitable trading strategies. From a broader perspective, these results highlight the changing role played by trading and trading strategies in affecting asset price dynamics. Recognizing these trading dynamics may be particularly useful for asset pricing research.[36]

Recognizing these dynamics may also be a fruitful path for future microstructure research. Traditional microstructure research has focused on the idiosyncratic; that is, market makers cared about their own inventory in a particular stock, and information was asset-specific. But cross-asset trading activity (and, particularly, cross-asset market making) is now the norm, and electronic market making algorithms exhibit a complexity not captured by simple single-asset models. Our machine learning analysis of cross-asset effects shows the importance of these broader effects. While we continue to find predictive power from own-asset microstructure measures, we find strong evidence of cross-asset effects.

What is particularly intriguing is the broad influence that a small set of such cross-asset microstructure metrics has in predicting market variables—suggesting that these metrics may be capturing systemic effects. Microstructure research (Chordia et al. 2000; Hasbrouck and Seppi 2001; Malceniece et al. 2019) has investigated commonality in spread movements across markets, finding evidence of at least some co-movement in spreads. Our findings

[36] The information arrivals perspective has proven useful in explaining the time series of return premiums. For example, Savor and Wilson (2014, 2016) show that the bulk of the equity return premium is earned on a handful of days that are systematically important in terms of information arrivals, such as macroeconomic news days and earnings announcements. See also Easley et al. (2019).

here may explain why such co-movement can occur by identifying measures capturing such potential underlying microstructure systemic effects. However, before we can reach such a conclusion, we need more research, both empirical to specifically address this systemic issue and theoretical to develop models capable of capturing and explaining these broader influences. We think machine learning will play a role in helping both research agendas.

References

Abadie, A., and M. Kasy. Forthcoming. The risk of machine learning. *Review of Economics and Statistics*

Andersen, T., and O. Bondarenko. 2015. Assessing measures of toxic order flow and early warning signals for market turbulence. *Review of Finance* 19:1–54.

Amihud, A. 2002. Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets* 5:35–56.

Barardehi, V., D. Bernhardt, and R. Davies. 2019. Trade-time measures of liquidity. *Review of Financial Studies* 32:126–79.

Cavallo, A., and R. Rigobon. 2016. The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives* 30:171–98.

Chakrabarty, B., P. Moulton, and A. Shkilko. 2012. Short sales, long sales, and the Lee–Ready trade classification algorithm revisited. *Journal of Financial Markets* 15:467–91.

Chinco, A., A. Clark-Joseph, and M. Ye. 2018. Sparse signals in the cross-section of returns. *Journal of Finance* 74:449–92.

Chordia, T., R. Roll, and A. Subrahmanyam. 2000. Commonality in liquidity. *Journal of Financial Economics* 56:3–28.

Corwin, S., and P. Schultz. 2012. A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance* 67:719–59.

Easley, D., N. Kiefer, M. O'Hara, and J. Paperman. 1996. Liquidity, information, and infrequently traded stocks. *Journal of Finance* 51:1405–36.

Easley, D., M. Lopez de Prado, and M. O'Hara. 2012a. The volume clock: Insights into the high frequency paradigm. *Journal of Portfolio Management* 39:19–29.

———. 2012b. Flow toxicity and liquidity in a high frequency world. *Review of Financial Studies* 25:1457–93.

———. 2015. Optimal execution horizon. *Mathematical Finance* 25:640–72.

———. 2016. Discerning information from trade data. *Journal of Financial Economics* 120:269–86.

Easley, D., and M. O'Hara. 1992. Time and the process of security price adjustment. *Journal of Finance* 47:577–607.

Easley, D, D. Michaelyk, M. O'Hara, and T. Putnins. 2019. Information flows and asset pricing. Working Paper, University of Technology Sydney.

Engle, R., and J. Lange. 2001. Measuring, Forecasting and Explaining Time Varying Liquidity in the Stock Market. *Journal of Financial Markets* 4:113–42.

S. Gu., B. Kelly, and D. Xiu. 2018. Empirical asset pricing via machine learning. Chicago Booth Research Paper No. 18-04.

Hasbrouck, J. 2009. Trading costs and returns for U.S. equities: Estimating effective costs from daily data. *Journal of Finance* 63:1445–74.

Hasbrouck, J., and D.J. Seppi. 2001. Common factors in prices, order flows, and liquidity. *Journal of Financial Economics* 59:383–411.

C. Krauss, X. A. Do, and N. Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research* 259:689–702.

Kyle, A. P. 1985. Continuous auctions and insider trading. *Econometrica* 53:1315–35.

Lopez de Prado, M. 2018. *Advances in Financial Machine Learning.* New York: Wiley.

Low, R., T. Li, and T. Marsh. 2018. BV-VPIN: Measuring the impact of order flow toxicity and liquidity on international equities markets. Working Paper, University of Queensland Business School.

Malceniece, L., K. Malcenieks, and T. Putnins. 2019. High frequency trading and comovement in financial markets. *Journal of Financial Economics* 134:381–99.

Mullainathan, S., and J. Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31:87–106.

O'Hara, M. 2015. High frequency market microstructure. *Journal of Financial Economics* 116:257–70.

Philip, R. 2020. Estimating permanent price impact via machine learning. *Journal of Econometrics* 215:414–49.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–30.

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39:1127–39.

Rossi, A., 2018. Predicting stock market returns with machine learning. Working Paper, University of Maryland.

Savor, P., and M. Wilson. 2014. Asset pricing: A tale of two days. *Journal of Financial Economics* 113:171–201.

———. 2016. Earnings announcements and systematic risk. *Journal of Finance* 71:83–138.

Simon, H. A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* 106:467–82.

Varian, H. R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28:3–28.