Intraday Market Predictability: A Machine Learning Approach

Dillon Huddleston*, Fred Liu†, and Lars Stentoft‡

February 16, 2021

Abstract

Conducting, to our knowledge, the largest study ever of five-minute equity market returns using state-of-the-art machine learning models trained on the cross-section of lagged market index constituent returns, we show that regularized linear models and nonlinear tree-based models yield significant market return predictability. Ensemble models perform the best across time and their predictability translates into economically significant Sharpe ratios of 0.98 after transaction costs. These results provide strong evidence that intraday market returns are predictable during short time horizons, beyond what can be explained by transaction costs. Furthermore, we show that constituent returns hold significant predictive information that is not contained in market returns or in price trend and liquidity characteristics. Consistent with the hypothesis that predictability is driven by slow-moving trader capital, predictability decreased post-decimalization, and market returns are more predictable during the middle of the day, on days with high volatility or illiquidity, and in financial crisis periods.

JEL Classification: G14, G17, C45, C55

Keywords: Machine Learning, Return Prediction, High-Frequency, Equity Market, Big Data, Lasso, Elastic Net, Random Forest, Gradient Boosting, Deep Neural Networks, Fintech

*Department of Economics, University of Western Ontario, Canada, dhuddles@uwo.ca.

†Department of Economics, University of Western Ontario, Canada, fliu227@uwo.ca.

‡Department of Economics and Department of Statistical and Actuarial Sciences, University of Western Ontario, Canada, lars.stentoft@uwo.ca. Corresponding author.

1 Introduction

The predictability of the aggregate market is a central topic in financial economics. While long-horizon (i.e. monthly or quarterly) market predictability has been extensively studied, intraday (i.e. within a trading day) market predictability has received relatively less attention. Traders require time to incorporate new information about cash flows and discount rates, and over short time horizons equity prices can differ from their adjusted fundamental values, particularly when market frictions are high. This process may introduce short-horizon predictability in equity returns and raises several interesting questions. First of all, if markets are predictable at short horizons what is the magnitude of this predictability? Secondly, is intraday return predictability economically profitable, and if so, does this profitability survive transaction costs? Finally, if markets are predictable it is obviously interesting to know which characteristics, like, e.g., lagged liquidity or price trends, are in fact important for predicting intraday returns.

Motivated by these questions, we study intraday market predictability using a cross-section of lagged returns of the market and its constituent stocks as predictor variables.[1] We believe our paper is the first to conduct such a study and speculate that the lack of previous studies on this topic may be due to statistical challenges associated with the high-dimensional inputs and the computational difficulties of estimating models using large panels of high-frequency data. We overcome the first issue by using a variety of cutting-edge machine learning models necessary to accommodate the long list of predictors and rich functional forms. We consider candidate methods from Gu, Kelly, and Xiu (2020b) and Hastie, Tibshirani, and Friedman (2009), including linear models with regularization and dimension reduction using lasso (LAS), elastic net (EN), and principal component regression (PCR), and nonlinear tree based models like random forests (RF) and gradient-boosted regression trees (GBRT) along with artificial neural networks (ANN). We also consider the ensemble mean and median of these models. Our baseline model uses all the five-minute returns in an expanding estimation window. Training machine learning models on such a large data panel is computationally challenging and this second issue is overcome using the Apache Sparkling Water and H2O.ai computing framework, which allows us to efficiently estimate machine learning models on large datasets. For example, in October, 2016, the estimation window covers

[1] While individual stocks are likely more predictable than the market, we begin with market predictability to avoid issues related to data mining and concerns about lack of liquidity.

285 months and contains roughly 450,000 five-minute returns for each S&P constituent, requiring approximately 48 hours to estimate all models.

Our null hypothesis throughout this paper is that markets are not predictable, since any predictability should be removed by active traders. Our alternative hypothesis is that the information in lagged returns is *not* instantly reflected in market prices, and as a result, lagged returns *are* predictive of short-term market returns. These hypotheses are tested by examining the statistical predictability and the economic significance thereof for each of the machine learning models considered. If models can forecast returns and consistently profit in so doing, then markets are likely to be predictable on short-horizons. Our second alternative hypothesis is that there is predictability beyond that explained by transaction costs. This is tested by evaluating the economic significance of model predictions after accounting for realistic transaction costs. If model predictions remain profitable after such costs, then there is likely incorrect pricing beyond that explained by transaction costs.

To examine if there is statistical predictability of intraday market returns over five-minute time intervals, our estimated models are trained on lagged returns with an expanding estimation window from 1993 to 2016. For non-ensemble models, linear as well as nonlinear, out-of-sample R^2 (R^2_{OOS}) values range up to 2.00% for LAS, followed by 1.95% for EN, and 1.71% for the nonlinear RF model. The ensemble mean (median) model yields R^2_{OOS} of 2% (2.01%), illustrating the strength of combined forecasts. However, most of this predictability is concentrated in the pre-decimalization period from 1993 to 2000. In the early post-decimalization period from 2001 to 2004, the LAS, EN, and RF models still have positive R^2_{OOS} values of 0.91%, 0.81%, and 0.85%, respectively. The ensemble mean and median models have respective post-decimalization R^2_{OOS} values of 1.04% and 1.01%. However, during the late decimalization period from 2005 to 2016, we find model predictability significantly decreases due to decreased transaction costs, but remains positive for the LAS, PCR, and ensemble models.

Next, to examine the economic significance of model predictions, a market-timing strategy that buys (sells) the market on positive (negative) predictions is considered. Our results demonstrate that a small intraday R^2_{OOS} value can yield large economic profits, especially given the numerous trading opportunities available at a five-minute interval. Using the baseline model from 1993 to 2016, all models are found to have positive returns. The LAS, EN, RF, and GBRT models have the

highest non-ensemble statistical predictability and so too high annualized returns (Sharpe ratios) of 191%, 188%, 198%, and 192% (2.71, 2.68, 2.90, and 2.84). The ensemble mean and median have annualized returns (Sharpe ratios) of 205% and 204% (2.90 and 2.82). Again, most returns are concentrated in the pre-decimalization period. However, in the late post-decimalization period from 2005 to 2016, all models still earn economically significant profits, with ANN earning the lowest returns (Sharpe ratios) of 16% (0.83). The consistently positive returns and high Sharpe ratios provide further evidence that intraday markets are predictable.

Lastly, economic significance is analyzed after transaction costs. To do this the market-timing strategy is modified to only trade when the signal is strong, i.e., when the model prediction exceeds the transaction cost. Using the baseline model from 1993 to 2016, all models, with the exception of ANN, are found to have positive returns even after accounting for transaction costs. The PCR, RF, and ensemble mean and median models have Sharpe ratios of 0.68, 0.77, 0.67, and 0.98, respectively, vastly exceeding the Sharpe ratio of 0.48 for the benchmark buy-and-hold SPDR S&P 500 (SPY) portfolio. Again, most returns are concentrated in the pre-decimalization period. In the late post-decimalization period from 2005 to 2016, the PCR and RF models have annual returns (Sharpe ratios) of 9% and 8% (0.88 and 0.81) after transaction costs. The ensemble mean and median have respective annual returns (Sharpe ratios) of 4% and 5% (0.35 and 0.68) after transaction costs. As a comparison, from 2005 to 2016 the SPY returned 7% with a Sharpe ratio of 0.46. Thus, in this recent sample PCR, RF, and the ensemble median models continue to earn significantly higher Sharpe ratios than the market does even after transaction costs, providing strong evidence that markets are predictable even after accounting for transaction costs.

We hypothesize that the demonstrated significant predictability of intraday market returns through time is driven by slow-moving trader capital, i.e. by infrequent portfolio rebalancing (Bogousslavsky (2016) and Duffie (2010)), at a high frequency. If some traders rebalance their portfolios infrequently, they may be slow to incorporate shocks to individual stock returns into the aggregate market, particularly when traders face severe volatility or illiquidity. To analyse this further, we examine if our results differ within the trading day, in periods of high versus low volatility or illiquidity, and during periods of financial crisis. First, since traders are most active at the beginning and end of each day, we expect predictability to be low during those times. Our results show that predictability is indeed stronger when traders are less active and exhibits an

inverse-U shape. Second, during periods of high volatility and high illiquidity, traders encounter significant market frictions. Consistent with our hypothesis we find that predictability and its economic significance increases when market volatility and illiquidity are high. Finally, since crisis periods are associated with significant market frictions we expect to find that predictability is also relative high in these periods. Our results demonstrate that predictability is indeed stronger during the Subprime Mortgage crisis and the EU debt crisis.

The baseline predictor variables used here are the cross-section of lagged intraday returns for the market constituents. A natural comparison to these cross-sectional models are autoregressive models for the market return itself, since our predictability results could simply be capturing intraday momentum. For example, Heston, Korajczyk, and Sadka (2010) find significant auto-correlation of half-hour returns at daily intervals and Gao, Han, Li, and Zhou (2018) find that the first half-hour return of the SPY predicts the last. Thus, as an additional analysis we validate our cross-sectional model's results by contrasting them against the results obtained with AR(1), AR(p) where the lag-order p is chosen to minimize the validation error, and AR(500) models estimated using OLS, LAS, EN, and PCR. We confirm that our baseline cross-sectional models significantly outperform the autoregressive models, indicating that the cross-section of lagged constituent returns has significant predictive information that is not contained in lagged market returns.

As a final additional analysis we evaluate intraday market predictability using additional lagged stock characteristics as predictors. In this analysis we include market beta, momentum, illiquidity, extreme returns, trading volume, volatility, skewness, and kurtosis, all estimated over the previous day.[2] We also consider the lagged bid-ask spread. In most cases adding additional variables is found to *decrease* model predictability, indicating that these characteristics do not help lagged returns predict the market portfolio returns. These findings have important implications for the possible economic mechanisms that drive such predictability. For example, could the predictability be caused by intraday momentum, as argued by Heston, Korajczyk, and Sadka (2010) and Gao, Han, Li, and Zhou (2018)? This explanation seems unlikely given that price trend variables fail to improve model predictability. A big picture implication of our findings is that a careful exploration of the economic mechanisms driving predictability is warranted.

Our paper is related to at least three existing strands of literature. First, our results are naturally

[2]Gu, Kelly, and Xiu (2020b) demonstrate that price trend and liquidity have the strongest predictive ability.

related to the recent literature examining intraday return predictability using lagged returns and trading volume. Chordia, Roll, and Subrahmanyam (2005) and Chordia, Roll, and Subrahmanyam (2008) study the predictability of short-run stock returns, finding that intraday returns cannot be predicted by past prices, but that order imbalances do forecast short-horizon returns. Heston, Korajczyk, and Sadka (2010) find significant autocorrelation of returns at daily intervals, for up to 20 days. Gao, Han, Li, and Zhou (2018) demonstrate that the first half-hour return of the SPY market exchange-traded fund (ETF) predicts the last half-hour return. Bogousslavsky (2016) theoretically establishes that seasonality in intraday returns may be caused by traders' infrequent rebalancing. Chinco, Clark-Joseph, and Ye (2019) use a LAS model on the cross-section of NYSE lagged returns to show that one-minute returns are predictable. These studies use *linear* models to forecast returns, whereas our models use information from the entire *cross-section* of lagged returns as well as other characteristics, and permit more flexible functional forms accommodating variable interactions and other *non*linear effects. Furthermore, Ke, Kelly, and Xiu (2019) and Renault (2017) show that text data forecasts intraday returns.

Our paper also relates to the rapidly expanding literature applying machine learning techniques in financial economics.[3] Our paper is most closely related to Gu, Kelly, and Xiu (2020b), which applies an extensive array of machine learning techniques to the problem of predicting equity risk premiums (see also Fischer and Krauss (2018), Long, Lu, and Cui (2019) and Marković, Stojanović, Stanković, and Stanković (2017)).[4] Bianchi, Büchner, and Tamoni (2021) consider the prediction of *bond*, rather than stock, risk premiums, and other financial applications of machine learning include their use for derivatives pricing (Ye and Zhang (2019)), hedge fund selection and return prediction (Chen, Wu, and Tindall (2016)), credit risk management (Barboza, Kimura, and Altman (2017)), portfolio management and optimization (Yun, Lee, Kang, and Seok (2020) and Day and Lin (2019)), cryptocurrency (Dutta, Kumar, and Basu (2020) and Alessandretti, ElBahrawy, Aiello, and Baronchelli (2018)), stochastic discount factors (Korsaye, Quaini, and Trojani (2019)), and factor models (Bryzgalova, Pelger, and Zhu (2019), Chen, Pelger, and Zhu (2019), Feng, Polson, and Xu (2019), Gu, Kelly, and Xiu (2020a), and Kelly, Pruitt, and Su (2019a)).

[3] See Weigand (2019) for a recent concise summary of asset pricing via machine learning and Heston and Sinha (2017) and Hajek and Barushka (2018) for reviews of the history of financial applications of machine learning.

[4] Other papers similar to that of Gu, Kelly, and Xiu (2020b), which are also concerned with the use of machine learning for predicting returns, include those of Chong, Han, and Park (2017), Feng, He, and Polson (2018), Kelly, Pruitt, and Su (2019b), Sutherland, Jung, and Lee (2018), and Xue, Zhou, Liu, Liu, and Yin (2018).

Finally, our study is similar in spirit to the literature on aggregate market return predictability, which focuses on long horizons over months, quarters and years.[5] Fama (1991) established the general view that return predictability at long horizons is driven by time-varying expected returns, consistent with market efficiency. Our research differs fundamentally, since the established predictability and its economic value is driven by market frictions at high frequencies.

The paper is organized as follows: Section 2 introduces the machine learning methods we consider and explains how we evaluate intraday predictability. Section 3 presents results for statistical predictability, economic significance before and after transaction costs, and a robustness check using different training window sizes. Section 4 examines if the results differ across the time of day, across levels of volatility and illiquidity, or during crisis periods, whether autoregressive models can also forecast returns, and whether predictability can be improved by including additional input variables. Section 5 concludes. The appendices contain further details on the data used, variable cleaning, the stock characteristics used as inputs, and on the implementation of the machine learning methods, in general, and hyperparameter optimization, in particular.

2 Methodology

Given T high-frequency price observations, indexed by $1 \leq t \leq T$, and denoting by p_t the natural logarithm of the t^{th} observed price, the corresponding logarithmic return is given by

$$r_t \equiv p_t - p_{t-1}. \tag{1}$$

Of particular interest in this paper is the market return which we denote by r_t^M and whether or not this can be predicted. To assess this, we conduct one of the largest empirical studies of market prediction in the high-frequency literature. The primary dataset is the trade and quotes (TAQ) database, containing intraday transaction data for all stocks on the New York stock exchange (NYSE), American stock exchange (AMEX), NASDAQ, and other American regional exchanges, from February, 1993, to October, 2016. Our goal is to predict five-minute changes in the aggregate market, which we proxy by the SPDR S&P 500 (SPY) ETF. The baseline predictor variables are

[5] See, among others, Ang and Bekaert (2007), Campbell and Thompson (2008), Chen, Da, and Zhao (2013), Kelly and Pruitt (2015), Neely, Rapach, Tu, and Zhou (2014), and Welch and Goyal (2008). See Koijen and Van Nieuwerburgh (2011) and Rapach and Zhou (2013) for recent surveys.

lagged returns of the SPY and S&P 500 constituents. We obtain the list of S&P 500 constituents from the center for research in security prices (CRSP), and update the 500 constituents monthly.[6]

Given a vector of lagged characteristics of the SPY and S&P 500 constituents, $\mathbf{X}_t^\ell$, indexed by integers, $0 \leq \ell \leq 500$, the objective of our paper is to use state of the art machine learning methods to approximate the empirical model given by

$$r_{t+1}^M = f\left(\mathbf{X}_t^\ell\right). \tag{2}$$

In the baseline model, the covariates are lagged returns, such that $\mathbf{X}_t^\ell = r_t^\ell$, resulting in 501 covariates. We consider in addition models that include other firm-level characteristics, such that $\mathbf{X}_t^\ell = [\ r_t^\ell \quad z_t^\ell\ ]$, where [.] denotes matrix concatenation, and z_t^0 (z_t^ℓ with $\ell \geq 1$) is the market beta (of firm $\ell \geq 1$), illiquidity, kurtosis, maximum, minimum, momentum, skewness, volatility, trading volume, or bid-ask spread. These expanded models have 1002 covariates (501 lagged returns and 501 characteristics).[7]

The rest of this section introduces several methods for estimating the function f in the empirical model in Equation (2) above. In this paper we consider specifically 1) linear models, 2) nonlinear models, and 3) so-called ensemble models that weigh together multiple individual models. Finally, we explain how, given an estimator of f, possible predictability can be assessed both statistically and economically.

2.1 Linear models

When restricting the focus to linear models the optimization problem is given by

$$\inf_{\beta \in \mathbb{R}^K} m[\mathbf{y} - \mathbf{X}\beta], \tag{3}$$

where we use notation from the machine learning literature and refer to the market returns as the targets denoted by $\mathbf{y}$, an $n \times 1$ column vector, and where the predictors, e.g. lagged returns or other characteristics, are denoted by $\mathbf{X}$, an $n \times K$ matrix in the case of K predictors. In Equation

[6] See Appendix A.1 for more details on the databases used.

[7] The only exception to this is the expanded model with z_t^ℓ equal to market beta which only has 1001 covariates, since the SPY market beta $z_t^0 = 1$. See Appendix A.2 for details on how these characteristics are calculated.

(3), $m[\cdot]$ denotes a metric or loss for the fit of the model and β denotes the relevant parameters in the model. For example, in the case of OLS regression the metric is taken to be the Euclidean/ℓ^2 norm, i.e. $m \equiv \|\cdot\|_2$, and the solution to the optimization problem in Equation (3) is given by the classical OLS estimator

$$\hat{\beta} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

In the presence of many predictors, the simple linear regression model easily becomes inefficient leading to in-sample over-fitting which is detrimental to our objective of out-of-sample prediction. We present here two approaches to deal with this: 1) regularization and 2) principle component regression.

2.1.1 Regularization

One primary objective of *regularization* techniques (see, e.g., Friedman, Hastie, and Tibshirani (2010)) is to avoid over-fitting in statistical models. This is often accomplished by adding a penalty term to the optimization problem in Equation (3) as follows

$$\inf_{\beta\in\mathbb{R}^K} \left\{m[\mathbf{y} - \mathbf{X}\beta] + \lambda n[\beta]\right\}. \tag{4}$$

Here, the functional $n[\cdot]$, often a norm, penalizes non-zero estimators and the *regularization parameter* λ regulates the penalty's impact as a multiplicative scale. A classical example is ridge regression, in which the optimization problem in Equation (4) is modified to

$$\inf_{\beta\in\mathbb{R}^K} \left\{\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}.$$

The smoothness of ridge regression (see, e.g., Marquaridt (1970)) resulting from using the ℓ^2 norm is computationally advantageous, but may result in many 'near-but-non-zero' coefficients and thus may not reduce the dimensionality of the optimization problem in a sufficient manner. In this paper we instead consider lasso regression (see, e.g., Tibshirani (1996)) and elastic nets (see, e.g., Zou and Hastie (2005)).

In the case of a lasso regression (LAS) the optimization problem in Equation (4) is modified to

$$\inf_{\beta \in \mathbb{R}^K} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

Hence LAS employs the computationally difficult (i.e., non-smooth) ℓ^1 norm, but has the resulting advantage that many coefficients are driven to zero exactly, leaving out only those of sufficient predictive importance. The resulting β, is non-zero only for those predictors which most significantly determine the target and may therefore be of much lower dimension than the original problem which is indeed one of the primary objectives of our use of regularization techniques.

In the case of an elastic net regression (EN) the optimization problem in Equation (4) is modified to

$$\inf_{\beta \in \mathbb{R}^K} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \alpha\lambda_1 \|\beta\|_1 + \frac{1-\alpha}{2}\lambda_2 \|\beta\|_2^2 \right\}.$$

Hence EN convexly combines ridge and lasso penalties to balance these two competing properties; smoothness and perfect elimination of unimportant predictors. Here, $\alpha \in [0, 1]$ is the coefficient of the convex combination of ℓ^1 and ℓ^2 norms of the regression coefficients, β, and in general, each may have its own regularization parameter, respectively, λ_1 and λ_2. In the particular cases of $\alpha \in \{0, 1\}$, elastic nets respectively reduce to ridge and lasso regressions.[8]

2.1.2 Principal component regression

Principal component regression (PCR) is a dimension-reduction technique used to summarize variation within a data set using a small number of linear combinations thereof (see, e.g., Jolliffe (2002)). Given a data set $\mathbf{X}$, consisting of n observations of K predictors, PCR solves the following problem

$$\sup_{\mathbf{w} \in \mathbb{R}^K} \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\mathbf{w}^T \mathbf{w}},$$

where $\mathbf{w} \in \mathbb{R}^K$ are the predictor weight vectors and Σ is the covariance of the predictors. The motivation for PCR is clear from this formulation: since an eigenvector of $\mathbf{\Sigma}$, $\mathbf{w}$, solves the eigenvalue problem, $\mathbf{\Sigma w} = \lambda \mathbf{w}$, for a corresponding eigenvalue of $\mathbf{\Sigma}$, λ, the best *eigenvector* solution of this

[8] Appendix B explains how the hyper-parameters α, λ_1, and λ_2, or in the case of LAS only λ, are selected.

problem is obviously that for which the corresponding eigenvalue, λ, is largest. This follows since the Raleigh quotient to be optimized, $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}/\mathbf{w}^T\mathbf{w}$, which measures normalized variation/variance of the data set along the weight vector, $\mathbf{w}$, simplifies in this case to λ. So in applying PCR, it is necessary only to compute an eigenvalue decomposition of $\mathbf{\Sigma}$, sort its eigenvalues, and take as many corresponding eigenvectors as are desired principal components, to yield the principal directions. Normalizing each by its corresponding square-root eigenvalue, the desired principal components are obtained.

In the context of the optimization problem in Equation (3), PCR is first applied to the predictors, $\mathbf{X}$. Supposing $\kappa \ll K$ components are desired, the projected predictors, $\mathbf{Z}$, are then used in a linear regression yielding the classical OLS estimator given by

$$\hat{\beta} \equiv (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y},$$

in which

$$\mathbf{Z} \equiv (\mathbf{X} - \mu)\mathbf{W},$$

where μ is the mean of the predictors and $\mathbf{W} \equiv [\ \mathbf{w}_{(1)} \quad \mathbf{w}_{(2)} \quad \cdots \quad \mathbf{w}_{(\kappa)}\ ]$ is the matrix of principal components to be used. The resulting (demeaned) projected predictors, $\mathbf{Z}$, are thus only of dimension $\kappa \ll K$, yielding a potentially significant reduction in the number of predictors and resulting regression model complexity, while preserving, to the extent possible, the richness of variation in the original data.[9]

2.2 Nonlinear models

More generally, the fundamental problem to be addressed in this paper is to solve the following optimization problem

$$\inf_{\mathbf{f}\in\mathcal{F}} m[\mathbf{y} - \mathbf{f}(\mathbf{X})], \tag{5}$$

[9]Selecting the value of κ can be a complex problem. Appendix B explains how we set this hyper-parameter.

where $\mathcal{F}$ denotes the set of functions from which candidate predictors are drawn and m denotes a metric or loss for the fit of the model. Above we considered several linear models and we now consider two general examples of nonlinear models: 1) models based on decision trees and 2) models based on artificial neural networks.

2.2.1 Tree-based models

Decision trees are nonparametric, hierarchical sequences of decisions, which optimally construct, based on the training and validation data, sequences of decisions to classify or regress arbitrary input predictors. Individual decision trees train in logarithmic time with the number of training points, but over-fitting is common, as more decisions (greater 'depth') are needed to better model training data, which may not generalize well. Individual trees often behave chaotically, too, in that the optimal structure may change drastically in response to the addition or removal of a handful of training data. Ensembles of trees mitigate the resulting large variance of individual decision trees, and avoid over-fitting by restriction to simple individuals across the ensemble, but inherit from such individuals some degree of the advantages of interpretation and efficient training. The fundamentals and many refinements of decision tree training are provided by, e.g., Breiman, Friedman, Olshen, and Stone (1984).

Random forest (RF) models (see, e.g., Breiman (2001)) independently and pseudorandomly generate decision trees, which are separately trained and whose predictions are then averaged to yield the ensemble prediction. In RF prediction robustness improvements are achieved via variance reductions implicit in the law of large numbers. Given predictors $\mathbf{X}$, denote an individual decision tree's estimator, say that of the i^{th} tree generated in an ensemble, by $\mathbf{g}_i(\mathbf{X})$. Supposing there are N decision trees in the ensemble, the prediction of RF is simply

$$\mathbf{f}_{\text{RF}}(\mathbf{X}) \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i(\mathbf{X}).$$

That is, RF amounts to considering the (arithmetic) average of the predictions of individual trees in the forest.[10]

[10]The crucial hyper-parameters to optimize for RF are the number of trees in the forest, the maximum number of predictors or *features* considered at each node of each tree, and the maximum number of nodes between the leaves and root of any tree, known as the *depth*. Appendix B explains how these are selected.

Each individual decision tree in the ensemble is trained on data drawn pseudorandomly with replacement from the training data. I.e., each individual tree is trained on a bootstrapped sample of the training data, and aggregated in making predictions via the average of those from each individual, and so random forests constitute one instance of bootstrap aggregated (*bagged*) predictors (see, e.g., Breiman (1996)). This *bagging* yields many individuals with uncorrelated prediction errors and less variance on average, whereas individuals tend to have high variance and over-fit. The construction may however introduce estimator bias, which is the problem addressed below by gradient-boosted regression trees. The bagging implicit in random forests also addresses the NP-completeness of the problem of training a *globally* optimal individual decision tree, which necessitates the use of heuristics, including typical greedy implementations, in training such individuals and bagging mitigates much of the bias that such heuristics introduce during individuals' training.

As their name suggests, gradient-boosted regression trees (GBRT) implement *boosting*, a second case of ensemble methods. As opposed to averaging methods, the simple estimators are sequentially (and so not independently) generated in a manner which progressively eliminates bias from the ensemble (a standard reference is Schapire and Freund (2013)). Boosting refers to the notion of developing a strong learner (a predictor with metric, m, approaching zero on arbitrary data) from a weak one (a predictor which performs marginally better than random guessing). Schapire (1990) first affirmatively answered this *hypothesis boosting* question, and the adaptive resampling and combining, or *arcing*, algorithm of Freund and Schapire (1997) is regarded as the canonical method for achieving such boosting in machine learning. *Gradient* boosting refers to the generalization of this and other boosting algorithms to the case of arbitrary differentiable loss functions, first achieved explicitly by Friedman (2001) and Friedman (2002).

In the case of GBRT, the ensemble prediction is a weighted combination of individual predictions, with weights γ_i determined as part of training leading to the following representation

$$\mathbf{f}_{\mathrm{GBRT}}(\mathbf{X}) \equiv \sum_{i=1}^{N} \gamma_i \mathbf{g}_i(\mathbf{X}).$$

Specifying $\mathbf{f}_0 \equiv 0$ and denoting $\mathbf{f}_{\text{GBRT}} \equiv \mathbf{f}_N$, the following holds for $1 \leq i \leq N$

$$\mathbf{f}_i(\mathbf{X}) = \mathbf{f}_{i-1}(\mathbf{X}) + \gamma_i \mathbf{g}_i(\mathbf{X}).$$

At each stage, i, of training, the decision tree predictor, $\mathbf{g}_i$, is greedily chosen to solve

$$\inf_{\mathbf{g} \in \mathcal{G}} \mathcal{L}[\mathbf{y}, \mathbf{f}_{i-1}(\mathbf{X}) + \gamma_i \mathbf{g}_i(\mathbf{X})]. \tag{6}$$

Here, $\mathcal{L}$ is some loss function, typically expressed as a sum of losses, say L, between corresponding targets, $y_j \in \mathbf{y} \equiv \{y_j\}$, and predictors, $\mathbf{x}_{j\cdot} \in \mathbf{X} \equiv \{\mathbf{x}_{j\cdot}\}$, as

$$\mathcal{L}[\mathbf{y}, \mathbf{f}_{i-1}(\mathbf{X}) + \gamma_i \mathbf{g}_i(\mathbf{X})] \equiv \sum_{j=1}^{n} L[y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot}) + \gamma_i \mathbf{g}_i(\mathbf{x}_{j\cdot})]. \tag{7}$$

Note that the solution to the optimization problem in Equation (6) is conditional on the current ensemble, $\mathbf{f}_{i-1}$, as opposed to the independent generation of individual decision tree predictors, $\mathbf{g}_i$, in a RF.

In the case where the loss, L, is differentiable, gradient boosting solves the optimization problem in Equation (6) via gradient descent as

$$\mathbf{f}_i(\mathbf{X}) = \mathbf{f}_{i-1}(\mathbf{X}) - \gamma_i \sum_{j=1}^{n} \nabla_{\mathbf{f}} L[y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot})],$$

where the the weights, γ_i, are chosen to optimize a loss similar to that of Equation (7) given by

$$\inf_{\gamma > 0} \sum_{j=1}^{n} L\left(y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot}) - \gamma \partial_{\mathbf{f}} L[y_j, \mathbf{f}_{i-1}(\mathbf{x}_{j\cdot})]\right).$$

Thus, the weights can be interpreted as step sizes/learning rates in this gradient descent procedure.[11]

[11] In addition to the hyper-parameters discussed for RF, in the case of GBRT the loss function, L, used for training and the additional uniform multiplier/scale, ν, which may be factored out of the step sizes, γ_i, as an explicit learning rate, can also be important. See Appendix B for further details.

2.2.2 Artificial neural networks

As their name suggests, artificial neural networks (ANNs) are motivated by the neural networks found in animal brains, and theoretical neuroscientific models thereof (see Bishop (1995) and Bishop (2006)). The most general-purpose and well-known neural network architecture is feedforward in which an *input layer* consisting of the inputs, $\mathbf{y}^0 \equiv \mathbf{x}_{i\cdot}$, is followed by a sequence of *hidden layers*, each consisting of a number of neurons. Initially, the inputs are weighted by a set of learned parameters and added to another, the so-called *bias*, to yield an input for each neuron in the succeeding layer. Denoting by k_1 the number of neurons in this layer and $k_0 \equiv K$ the number of inputs, there are thus $k_1(k_0 + 1)$ parameters which yield this layer's inputs as

$$(\forall 1 \leq j \leq k_1)\ x_j^1 \equiv \mathbf{y}^0 \mathbf{w}_j^1 + b_j^1.$$

Here, $\mathbf{w}_j^1$ is the learned column vector of weights for the row vector of inputs, $\mathbf{y}^0$, and b_j^1 the corresponding learned additive bias. To introduce non-linearity into the output of a neuron $y_j^1 \equiv \phi_j^1(x_j^1)$ where ϕ_j^1 is the so-called activation function of neuron j.

More concisely we can write the mapping as

$$\begin{aligned} \mathbf{x}^1 &\equiv \mathbf{y}^0 \mathbf{W}^1 + \mathbf{b}^1 \\ \mathbf{y}^1 &= \mathbf{\Phi}^1(\mathbf{x}^1). \end{aligned} \tag{8}$$

Here, $\mathbf{W}^1$ is the matrix with columns $\mathbf{w}_j^1$, $\mathbf{b}^1$ is the row vector with entries b_j^1, and $\mathbf{\Phi}^1$ is the mapping from the row vector $\mathbf{x}^1$, with entries x_j^1, to the row vector $\mathbf{y}^1$, with entries, y_j^1. In succeeding hidden layers, if applicable, the process is iterated as follows

$$\begin{aligned} (\forall 2 \leq \nu \leq N)\ \mathbf{x}^\nu &\equiv \mathbf{y}^{\nu-1} \mathbf{W}^\nu + \mathbf{b}^\nu \\ \mathbf{y}^\nu &= \mathbf{\Phi}^\nu(\mathbf{x}^\nu), \end{aligned} \tag{9}$$

where N is the total number of hidden layers. Finally, the *output layer* yields predictions given by

$$\begin{aligned} \mathbf{x}^{N+1} &\equiv \mathbf{y}^N \mathbf{W}^{N+1} + \mathbf{b}^{N+1} \\ \mathbf{y}^{N+1} &= \mathbf{\Phi}^{N+1}(\mathbf{x}^{N+1}). \end{aligned} \tag{10}$$

In our application, the output layer is linear and the number of neurons, k_{N+1}, is naturally set to one, resulting in

$$\mathbf{f}_{\text{ANN}}(\mathbf{X}) \equiv \mathbf{y}^N \mathbf{W}^{N+1} + \mathbf{b}^{N+1}. \tag{11}$$

Hence, the neurons are simply linearly aggregated into the forecast.

Collecting all weights and biases across layers, the feedforward network has a total parameter count of $\sum_{\nu=0}^{N} k_{\nu+1}(k_\nu+1)$ as there are $\sum_{\nu=0}^{N} k_{\nu+1}$ biases and $\sum_{\nu=0}^{N} k_{\nu+1}k_\nu$ weights and such networks may indeed be highly parametric. Classically, these are optimized via stochastic gradient descent, but several adaptive/'momentum'-based generalizations have been proposed.[12]

2.3 Ensemble methods

Ensemble methods seek to combine multiple predictors' results, for both variance reduction and 'crowd wisdom' purposes. For example, various measures of central tendency or location parameters, including the median and alternate (weighted) means, such as the harmonic, geometric and arithmetic mean, may act as ensemble aggregation methods and represent an 'average' prediction based on all the predictors' outputs.[13] For regression problems, the two most prominent machine learning ensemble aggregation methods are boosting and bagging, as respectively outlined for GBRT and RF models. Abstractly, given any of these ensemble aggregation methods, say, AM $\in \{\text{Bag}, \text{boost}, \text{(weighted) arithmetic/geometric/harmonic mean}, \text{median}\}$, etc., and a set of predictors' outputs, e.g., $\{\mathbf{f}_{\text{LAS}}(\mathbf{X}), \mathbf{f}_{\text{EN}}(\mathbf{X}), \mathbf{f}_{\text{PCR}}(\mathbf{X}), \mathbf{f}_{\text{RF}}(\mathbf{X}), \mathbf{f}_{\text{GBRT}}(\mathbf{X}), \mathbf{f}_{\text{ANN}}(\mathbf{X})\}$, the ensemble prediction is

$$\mathbf{f}_{\text{AM}}(\mathbf{X}) \equiv \text{AM}\left(\{\mathbf{f}_{\text{LAS}}(\mathbf{X}), \mathbf{f}_{\text{EN}}(\mathbf{X}), \mathbf{f}_{\text{PCR}}(\mathbf{X}), \mathbf{f}_{\text{RF}}(\mathbf{X}), \mathbf{f}_{\text{GBRT}}(\mathbf{X}), \mathbf{f}_{\text{ANN}}(\mathbf{X})\}\right).$$

[12] In ANN, the optimization algorithm employed for training, the loss function it uses, the ℓ^1 and ℓ^2 regularization parameters, the number of epochs and batches, and the level of dropout all constitute *hyperparameters* which impact the quality of predictions from a model with trained *parameters*. However, the actual network architecture, characterized by the number of hidden layers, the number of neurons in each layer, and the activation function associated with each neuron, often have an even greater impact. Appendix B provides further details on these hyperparameters.

[13] Timmermann (2006) provides an overview of forecast combinations and Genre, Kenny, Meyler, and Timmermann (2013) show that forecast combinations using a simple average often outperform methods that rely on estimated combination weights.

In our application we will consider the average and the median as aggregation methods and as such this method does not introduce any additional hyperparameters.

It should be noted that regularization may also be applied to the ensemble aggregation methods just as it may be in the particular cases of GBRT and RF models.[14] The machine learning ensemble aggregation methods are more flexible than straightforward computation of some measure of central tendency or location parameter of the various models' predictions, as the former permit regularization and hyperparameter tuning. Which ensemble aggregation method to use is itself a hyperparameter optimization problem, albeit a small one which may be largely mitigated by simply computing many ensembles, e.g., boosting, bagging, and a variety of 'popular' measures of central tendency or location parameters. The justification for any particular aggregation method may come from either the statistical (variance-reduction and bias) properties of bagging or boosting, or laws of large numbers, central limit theorems, or other asymptotic results applicable to the computed measures of central tendency or location parameters.

2.4 Evaluation criteria

Our objective is to assess the predictability of market returns. We do so using two types of criteria: purely statistical criteria based on a relevant metric for out of sample model fit and economic criteria based on the obtained returns from trading on a given model's predictions.

2.4.1 Statistical significance

To evaluate the predictive performance of the high-frequency market return forecasts, we calculate the out-of-sample R^2 metric proposed by Gu, Kelly, and Xiu (2020b). Given the market return, r^M_{t+1}, and a corresponding model prediction given the history up to time t, $\hat{r}^M_{t+1}$, the out-of-sample R^2 is calculated as

$$R^2_{\text{OOS}} \equiv 1 - \frac{\sum_{t\in\text{Test}} \left(r^M_{t+1} - \hat{r}^M_{t+1}\right)^2}{\sum_{t\in\text{Test}} \left(r^M_{t+1}\right)^2}. \tag{12}$$

Note that the R^2_{OOS} metric is only calculated over the test samples, indexed by times t in the set Test. The denominator of R^2_{OOS} is the squared sum of market returns without demeaning. As

[14] In fact, Koren (2009) aggregated model predictions using GBRT, in the winning solution of the Netflix Prize.

discussed by Gu, Kelly, and Xiu (2020b), the historical mean underperforms a zero forecast. The historical mean return is noisy, resulting in artificially high estimates of R^2. Hence, we benchmark against zero rather than the historical mean.[15]

Following Gao, Han, Li, and Zhou (2018) and Chinco, Clark-Joseph, and Ye (2019), we also consider the estimated coefficients from running a simple predictive regression given by

$$r_{t+1}^M = \alpha + \beta \hat{r}_{t+1}^M + \epsilon_{t+1}, \tag{13}$$

where r_{t+1}^M is the market return and $\hat{r}_{t+1}^M$ is the model prediction for five-minute time interval $t+1$.

2.4.2 Economic significance

In addition to evaluating predictive performance, we assess the ability of each model to time the market. We implement a trading strategy that takes a long (short) position in the market if the model predicts a positive (negative) return. The profitability can be expressed as $\pi \equiv \sum_{t \in \text{Test}} \pi_t$ where the individual daily profits, π_t, are defined as

$$\pi_t\left(\hat{r}_{t+1}^M\right) \equiv \begin{cases} -r_{t+1}^M & \text{if } \hat{r}_{t+1}^M < 0, \\ r_{t+1}^M & \text{if } \hat{r}_{t+1}^M > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Reported are annualized excess arithmetic returns and Sharpe ratios of this trading strategy across the test sample. The Sharpe ratio is calculated as the monthly excess return divided by the corresponding standard deviation and scaled by $\sqrt{12}$.

The economic significance of the trading strategy could be completely driven by small return fluctuations such as the bid-ask bounce discussed by Roll (1984), which is not useful to traders. To assess the models' ability to predict larger returns, we consider a trading strategy that accounts for transaction costs. Given national best bid (ask) price, Bid_t (Ask_t), and midquote $\text{M}_t = \frac{\text{Bid}_t + \text{Ask}_t}{2}$,

[15] When we benchmark against the historical mean, the R^2 increases by approximately 0.01% across methods.

we estimate the transaction cost by the relative national-best bid-offer (NBBO) spread, given by

$$\text{Spread}_t = \frac{\text{Ask}_t - \text{Bid}_t}{\text{M}_t}. \tag{15}$$

Following Chinco, Clark-Joseph, and Ye (2019), we evaluate the economic significance of returns after transaction costs by implementing a trading strategy that only trades when model predictions exceed the transaction costs. The returns of this trading strategy are

$$\phi_t\left(\hat{r}^M_{t+1}\right) \equiv \begin{cases} -r^M_{t+1} - \text{Spread}_t & \text{if } \left|\hat{r}^M_{t+1}\right| > \text{Spread}_t \text{ and } \hat{r}^M_{t+1} < 0, \\ r^M_{t+1} - \text{Spread}_t & \text{if } \left|\hat{r}^M_{t+1}\right| > \text{Spread}_t \text{ and } \hat{r}^M_{t+1} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Again, we report annualized excess returns and Sharpe ratios for this trading strategy. For models to be profitable, their predicted returns must be properly directed and exceed transaction costs, that is $\hat{r}^M_{t+1} > 0$ and $|\hat{r}^M_{t+1}| > \text{Spread}_t$ for a long position to be implemented. This simple trading strategy presents a significant hurdle for model validation and provides a benchmark for returns available to traders. We use a simple strategy to avoid data mining concerns, but more sophisticated strategies can yield higher returns.

3 Empirical results

In this section we present the results for the eight models considered: lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), artificial neural networks (ANN), and the ensemble (arithmetic) average (MEAN) and median (MED). We omit results for OLS from our analysis, since we find that the simple linear model is highly inaccurate, resulting in a negative out-of-sample R^2 in most periods. All models are estimated using Sparkling Water from H2O.ai, which combines the machine learning algorithms of H2O with the big data capabilities of Apache Spark, and permits efficient estimation of machine learning models on much larger datasets than the existing literature. We train the models using all intraday and overnight observations, however we report results from 9:35 to 3:55 only excluding the opening and closing returns to avoid concerns regarding the accuracy of these auction-based

prices and the effect of overnight/weekend effects. For each model, we tune hyperparameters in the validation set using random search on mean squared error.

We evaluate the predictive performance of each model by out-of-sample R^2, R^2_{OOS}, and the predictive regression coefficients and t-statistics. We evaluate the economic significance by the excess returns and Sharpe ratios of the market timing trading strategy with and without transaction costs. In each case, we examine model performance over the entire testing sample from April, 1993, to October, 2016. The results are also presented for significant sub-periods, including the 1/8 tick size sample from 1993 to 1996, the 1/16 tick size sample from 1997 to 2000, the early post-decimalization sample from 2001 to 2004, and the late post-decimalization sample from 2005 to 2016. Finally, we consider the robustness of our results to using models trained on 1, 4, 7, 10, 16, 22, 34, and 58 months of returns instead of our baseline expanding window.

3.1 Market predictability

Table 1 reports coefficients from the predictive regression in Equation (13) along with Newey-West t-statistics with 79 lags, R^2_{OOS} percentages, and the ratio of R^2_{OOS} to mean transaction cost (R^2_{OOS}/(Bid-Ask Spread scaled by 1,000)) for the eight machine learning techniques considered. Panel A presents results for the entire testing sample from April 1993 to October 2016. Across models, slope coefficents are roughly 1 with most t-statistics exceeding 40. The magnitude of the intercept and slope coefficients are not exactly 0 and 1 respectively, since the relationship between realized and predicted returns is non-linear (for example, due to the prevalence of zero return observations in five-minute returns). The table shows that all the models predict the market better than a naive forecast of zero, since each of them have a positive R^2_{OOS}. The LAS and EN models have R^2_{OOS} of 2% and 1.95%, respectively. These regularized linear models have the highest R^2_{OOS} among non-ensemble models. Among non-linear models, RF performs best with an R^2_{OOS} of 1.71%. Surprisingly, ANN performs poorly despite performing well in Gu, Kelly, and Xiu (2020b). This performance difference is likely because they forecast using an ensemble of neural networks, while we are limited by computational time to a single neural network prediction each period, which typically has higher variance forecasts. Across all models, the ensemble mean and median perform best with respective R^2_{OOS} values of 2% and 2.01%. Note that most models have an R^2_{OOS} exceeding 1.6%, which is the R^2_{OOS} documented in Gao, Han, Li, and Zhou (2018) for 30-minute returns and

roughly twice the monthly R^2_{OOS} reported by Gu, Kelly, and Xiu (2020b). Having a similar level of R^2_{OOS} at a higher-frequency (like five-minutes in this paper) than at a lower frequency is interesting since more trades can be carried out based on the predictability. Our findings are different than Chordia, Roll, and Subrahmanyam (2005), which uses linear models and find that lagged returns cannot forecast five-minute equity returns. Our results reveal that by including lagged returns of market constituents using high-dimensional models, market returns are predictable using only lagged returns.

Next, we break down the predictability by time period based on observed structural breaks caused by changes to tick size and transaction costs. Panel B presents results for the early sample from 1993 to 1996, coinciding with a 1/8 tick size. This sample has the strongest predictability, with R^2_{OOS} exceeding 6% across all models and slope t-statistics exceeding 30. This finding makes sense intuitively, since predictability should be high when a large tick size prevents traders from bringing prices to their fundamental values. Also, trading volumes were relatively low and transaction costs were high during this time period, which further increased market frictions. Panel C presents results for the period with 1/16 tick size from 1997 to 2000. This period has the second highest predictability with R^2_{OOS} exceeding 3% for all models except PCR. The R^2_{OOS}/Cost ratio decreased by roughly 90% relative to the previous period, demonstrating that most of the decreases in predictability were due to factors other than the decreases in transaction costs, and possibly due to technological improvements for traders (i.e. faster computer driven trading).[16]

Beginning in 2001, exchanges adopted decimal ($0.01) trading ticks, which is the tick size in the remainder of our sample. For this reason, we refer to 2001 to 2016 as the post-decimalization period. Panel D reports results for the early post-decimalization period from 2001 to 2004. Coefficients are statistically significant at the 1% level across all models, and while the R^2_{OOS} are lower than pre-decimalization, they remain relatively high. Furthermore, the R^2_{OOS}/Cost ratio decreased by roughly 50% relative to the previous period, indicating that much of the decrease in predictability can be explained by decreases in transaction costs. The highest non-ensemble R^2_{OOS} are for the LAS, EN, and RF models at 0.91%, 0.81%, and 0.85%, respectively. The ensemble mean and median reach an R^2_{OOS} of 1.04% and 1.01%, respectively. These R^2_{OOS} are still in line with the

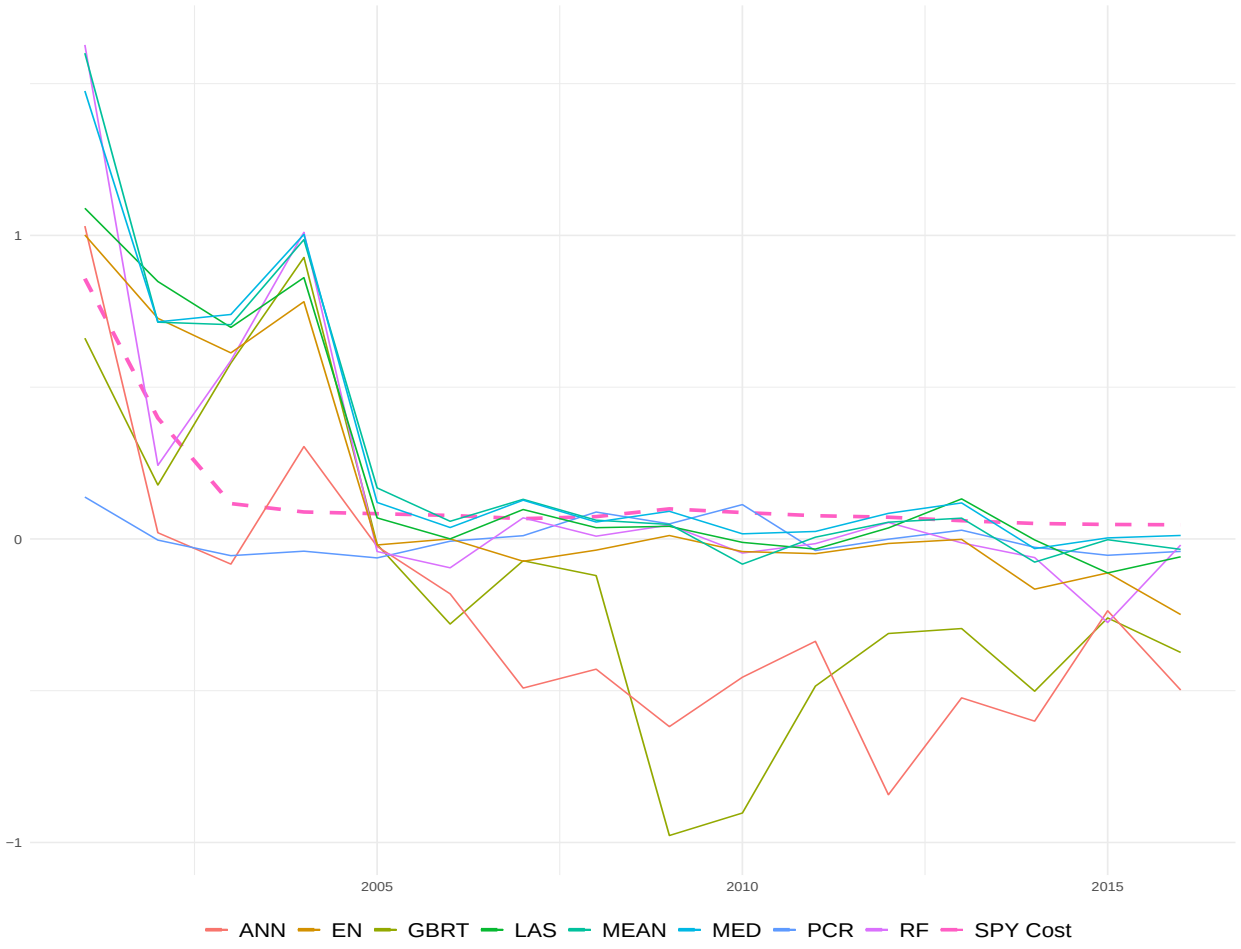
[16]The simple intuition for this is the following: if the decrease in R^2_{OOS} could be fully explained by a decrease in transaction costs, then the R^2_{OOS}/Cost ratio should remain the same across periods.

Table 1: Market predictability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Panel A: Overall Period (1993 - 2016)								
Intercept	-0.64	-0.59	-0.31	-0.45	-0.36	-0.32	-0.64	-0.67
t-stat	-4.02	-3.72	-1.98	-2.84	-2.28	-1.98	-4.04	-4.24
Slope	1.38	1.29	0.65	1.22	0.90	0.83	1.53	1.54
t-stat	47.26	46.06	24.14	54.11	43.23	42.49	58.28	58.97
R^2_{OOS}	2.00	1.95	0.31	1.71	1.64	1.40	2.00	2.01
R^2_{OOS}/Cost	1.16	1.13	0.18	0.99	0.95	0.81	1.15	1.16
Panel B: 1/8 Tick Size Period (1993 - 1996)								
Intercept	-0.22	-0.17	-0.16	-0.14	-0.32	0.16	-0.21	-0.26
t-stat	-1.16	-0.90	-0.95	-0.84	-1.68	0.79	-1.29	-1.59
Slope	1.55	1.45	0.92	1.70	1.18	1.17	1.59	1.65
t-stat	36.53	35.80	44.92	41.48	38.60	41.24	42.96	41.07
R^2_{OOS}	9.27	9.32	6.52	6.78	8.95	7.41	9.16	9.14
R^2_{OOS}/Cost	10.82	10.88	7.61	7.92	10.46	8.66	10.70	10.67
Panel C: 1/16 Tick Size Period (1997 - 2000)								
Intercept	-1.71	-1.67	-0.61	-1.28	-1.22	-1.04	-1.67	-1.73
t-stat	-3.91	-3.82	-1.40	-2.88	-2.71	-2.23	-3.80	-3.93
Slope	1.42	1.38	0.18	1.27	1.05	0.92	1.62	1.56
t-stat	35.52	35.31	2.77	46.73	41.86	38.57	43.56	45.09
R^2_{OOS}	3.72	3.69	-0.76	3.48	3.45	3.21	3.60	3.66
R^2_{OOS}/Cost	0.68	0.67	-0.14	0.64	0.63	0.59	0.66	0.67
Panel D: Early Post-Decimalization Period (2001 - 2004)								
Intercept	-0.65	-0.57	-0.42	-0.25	-0.30	-0.14	-0.55	-0.57
t-stat	-1.40	-1.24	-0.91	-0.55	-0.68	-0.32	-1.19	-1.24
Slope	1.14	0.99	0.59	0.87	0.63	0.61	1.36	1.38
t-stat	10.52	9.67	4.67	16.09	13.46	10.87	15.12	14.65
R^2_{OOS}	0.91	0.81	0.03	0.85	0.48	0.38	1.04	1.01
R^2_{OOS}/Cost	0.30	0.27	0.01	0.28	0.16	0.13	0.35	0.34
Panel E: Late Post-Decimalization Period (2005 - 2016)								
Intercept	-0.10	-0.03	-0.12	-0.07	0.08	0.07	-0.14	-0.20
t-stat	-0.43	-0.11	-0.52	-0.30	0.35	0.29	-0.60	-0.84
Slope	0.58	0.38	0.64	0.48	0.13	0.10	0.68	0.82
t-stat	3.70	3.17	4.80	4.85	1.53	1.25	4.02	4.59
R^2_{OOS}	0.02	-0.04	0.04	-0.01	-0.39	-0.45	0.04	0.06
R^2_{OOS}/Cost	0.07	-0.14	0.13	-0.03	-1.22	-1.42	0.12	0.18

This table reports coefficients from the predictive regression in Equation (13) along with Newey-West t-statistics with 79 lags, R^2_{OOS} percentages, and R^2_{OOS} scaled by transaction cost for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), and ensemble mean (MEAN) and median (MED). The intercept is reported as a percentage scaled by 1,000. The predictors used are lagged returns for the market and all the S&P 500 constituents. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

Figure 1: R^2_{OOS} and trading costs



Plot of the post-decimalization R^2_{OOS} for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). Also plotted is the median transaction cost scaled by 1,000 (SPY Cost).

market predictability results from Gao, Han, Li, and Zhou (2018) and Gu, Kelly, and Xiu (2020b) for the 30-minute and monthly time horizons, respectively. It should be noted that from 2001 to 2004 there was a significant decrease in transaction costs with no change to tick size. Figure 1 plots the median transaction cost (Bid-Ask Spread scaled by 1,000) against the R^2_{OOS} for each model during the post-decimalization period. Consistent with our hypothesis, the decrease in transaction costs in the early post-decimalization period significantly decreased the R^2_{OOS} across all models.

Finally, Panel E reports results for the late post-decimalization period from 2005 to 2016. The

LAS and PCR models have positive R^2_{OOS} of 0.02% and 0.04%, respectively, during this period. Interestingly, the PCR model has a positive R^2_{OOS} during this period despite having the lowest R^2_{OOS} in all previous periods. This is likely because the Subprime crisis and European debt crisis occurred in this period, and as a result of asset correlation increasing to near one in these crisis periods, returns had a strong factor structure.[17] The EN, RF, GBRT, and ANN models have negative R^2_{OOS} of -0.04, -0.01, -0.39, and -0.45, respectively, indicating that they perform worse than a naive constant prediction of 0. The poor performance of the non-linear models suggests that non-linear models may be over-fitting a simpler predictive relationship during this recent period. The ensemble mean and median, however, have the highest R^2_{OOS} of 0.04% and 0.06% respectively, benefiting from PCR's strong performance during crisis periods. This result shows that models that perform poorly on average can still improve ensemble models.

At a first glance, it may appear that after 2005 all of the predictability is gone. However, we argue that this is not the case. In particular, Figure 1 shows that the R^2_{OOS} for the two ensemble methods stay consistently positive, although it is small in magnitude, and the ensemble median R^2_{OOS} is positive in every year except for 2015. Relative to the previous period, the R^2_{OOS} decreased by roughly 95%, but the R^2_{OOS}/Cost ratio only decreased by 33-66%, indicating that some of the decrease in predictability can be explained by decreases in transaction costs. Given the substantial decrease in transaction costs for five-minute returns, a high R^2_{OOS} would be unreasonable, and a small but positive R^2_{OOS} should be expected. As argued by Campbell and Thompson (2008) and Rapach and Zhou (2013), even a small R^2 can generate economically large portfolio returns. This can be especially true given the number of trading opportunities at a five-minute interval.

In summary, our first set of results demonstrate that the market is remarkably predictable at the five-minute frequency. This predictability is highest prior to the decimalization of exchanges in 2001. Post-decimalization, markets became faster in integrating lagged intraday information. However, the R^2_{OOS} remains positive for the LAS, PCR, and ensemble models, demonstrating that some predictability persists even after the decimalization of exchanges and indicating that decreases in transaction costs also decreased market frictions.

[17] In Section 4.3 we confirm that the PCR model indeed performs well during both these crisis periods.

3.2 Economic significance

As we mentioned above, even a small predictability at five-minute intervals can result in large and economically significant returns. In this section we therefore evaluate the economic significance of our models' predictions by implementing the simple trading strategy specified in Equation (14). Since the models are optimized to minimize forecasting error, the economic significance of forecasts provides an indirect evaluation of model performance. Table 2 presents the annualized excess returns, Newey-West t-statistics with 1 lag and Sharpe ratios, denoted SR in the table, of the market timing strategy. Columns (1) - (8) presents results for the eight machine learning models and columns (9) and (10) report the intraday SPY returns as well as the benchmark buy-and-hold SPY returns, respectively, for comparison.

Panel A of Table 2 reports the results for the entire sample from 1993 to 2016. Rankings across methods are mostly consistent with their R^2_{OOS} percentages, and all models have positive returns and Sharpe ratios. Among the non-ensemble models, LAS, EN, RF, and GBRT have the highest R^2_{OOS} and also have high returns (Sharpe ratios) of, respectively, 191%, 188%, 198%, and 192% (2.71, 2.68, 2.90, and 2.84), indicating that dimension reduction is important for predicting returns. In particular, the tree-based RF and GBRT have the highest non-ensemble returns (Sharpe ratios), showing that modeling non-linearities and interaction effects are important for return prediction. The ANN model has lower returns (Sharpe ratios) of 172% (2.66) due to its high variance predictions. PCR has positive but relatively low excess returns of 75% matching its low R^2_{OOS} over the entire sample. Consistent with their R^2_{OOS} percentages, the ensemble mean and median have the highest returns (Sharpe ratios) of, respectively, 205% and 204% (2.90 and 2.82), illustrating that combining forecasts yields higher economic predictability. It is noteworthy that every model significantly out-performs the benchmark buy-and-hold SPY strategy, which yields excess returns (Sharpe ratios) of 7% (0.48).

Next, Panel B of Table 2 reports results for the 1/8 tick period from 1993 to 1996. The excess returns (Sharpe ratios) of all models exceed 300% (5). The large economic returns are consistent with the large R^2_{OOS} observed for this period. Panel C reports results for the 1/16 tick period from 1997 to 2000. While Table 1 showed that the R^2_{OOS} decreased relative to the previous period, surprisingly returns and Sharpe ratios increased during this period for nearly all models. This

Table 2: Excess Returns and Sharpe ratios

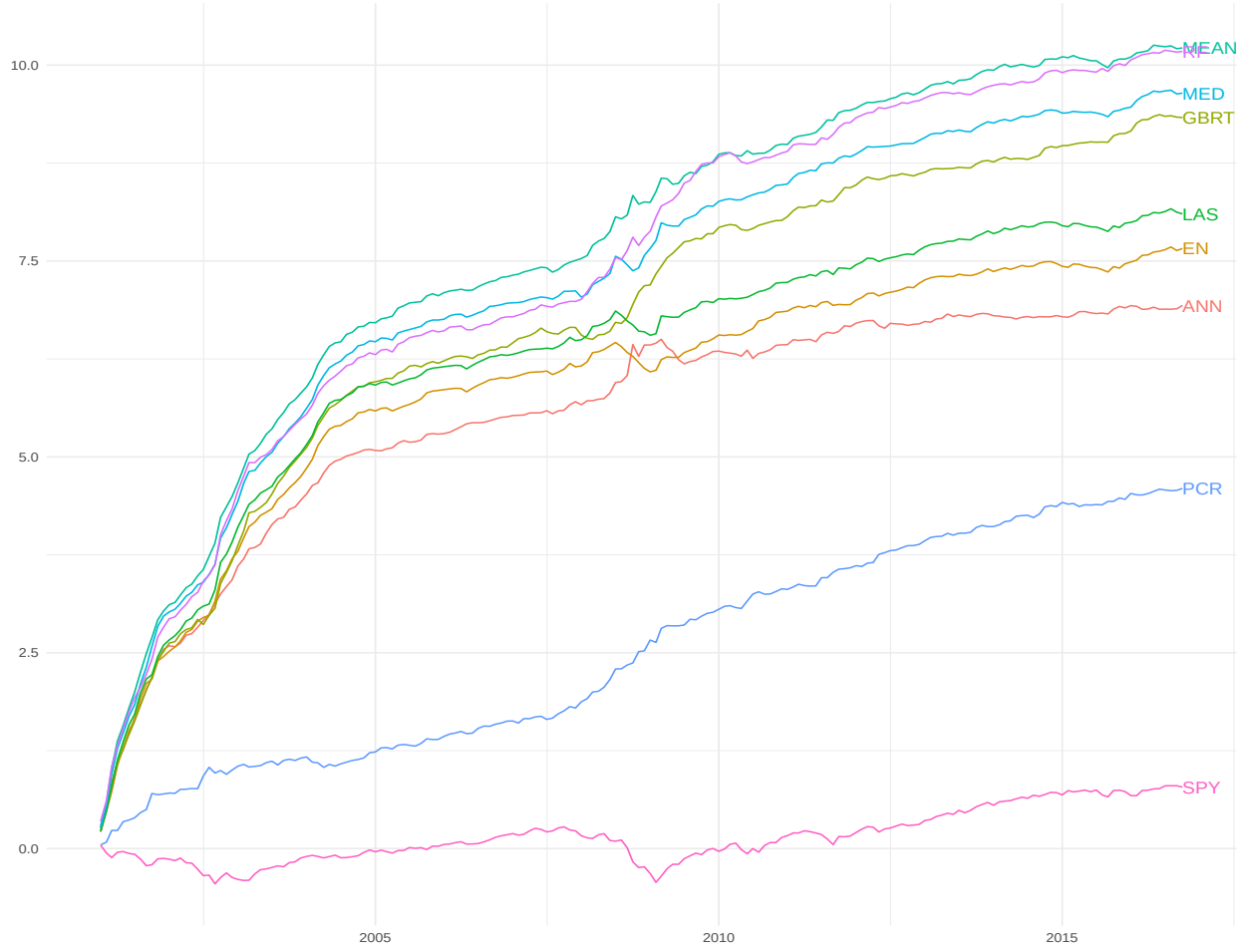
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
Panel A: Overall Period (1993-2016)										
Return	1.91	1.88	0.75	1.98	1.92	1.72	2.05	2.04	-0.03	0.07
t-stat	4.81	4.92	4.62	4.50	4.26	5.11	4.60	4.91	-1.11	2.19
SR	2.71	2.68	1.97	2.90	2.84	2.66	2.90	2.82	-0.24	0.48
Panel B: 1/8 Tick Size Period (1993 - 1996)										
Return	3.74	3.73	3.15	3.47	3.47	3.39	3.83	3.88	-0.01	0.11
t-stat	5.73	6.00	10.73	6.43	6.35	8.13	6.77	6.72	-0.37	2.62
SR	5.63	5.60	8.49	6.81	5.73	6.90	6.63	6.19	-0.17	1.19
Panel C: 1/16 Tick Size Period (1997 - 2000)										
Return	5.61	5.57	0.33	5.72	5.61	5.11	5.79	5.83	-0.14	0.12
t-stat	25.49	26.00	1.35	16.05	18.06	18.92	17.92	20.74	-1.78	1.62
SR	12.35	12.20	1.01	11.30	12.73	10.35	11.73	11.28	-0.89	0.71
Panel D: Early Post-Decimalization Period (2001-2004)										
Return	1.61	1.51	0.31	1.74	1.61	1.37	1.85	1.77	-0.04	-0.01
t-stat	5.12	5.74	3.32	5.37	6.31	4.06	4.46	4.60	-0.63	-0.12
SR	4.55	4.62	1.56	4.29	4.84	4.42	4.67	4.64	-0.35	-0.07
Panel E: Late Post-Decimalization Period (2005 - 2016)										
Return	0.18	0.17	0.29	0.33	0.29	0.16	0.30	0.27	0.01	0.07
t-stat	4.48	3.89	7.75	6.02	4.84	3.27	6.22	4.96	0.32	1.37
SR	1.36	1.41	2.10	2.06	2.03	0.83	1.80	1.73	0.09	0.46

Reported are the annualized excess returns, t-statistics, and Sharpe ratios for the SPY market-timing strategy. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). The independent variables are lagged returns of S&P 500 constituents. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

volatile period contains both the Asian crisis and the dot-com bust, and in Section 4.3 we show that the economic significance of model predictions increase during financial crises. PCR has a highly negative R^2_{OOS} during this period resulting in a low predictive return, suggesting that there is not a strong factor structure during this period.

Finally, we consider the results in the post-decimalization period. Panel D reports results for the early post-decimalization period from 2001 to 2004 and shows that, consistent with the finding that the R^2_{OOS} decreased post-decimalization, excess returns and Sharpe ratios, although significant and larger than the benchmark buy-and-hold SPY returns, are lower than in previous periods. Panel E reports results for the late post-decimalization period from 2005 to 2016. Table 1 showed that this period had a substantial decrease in R^2_{OOS}, and consequently returns and Sharpe ratios are significantly lower than in previous periods. However, returns and Sharpe ratios remain

Figure 2: Cumulative returns by model



Plot of the post-decimalization cumulative log returns of the market-timing strategy for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED).

high relative to the benchmark buy-and-hold SPY. In particular, all models have returns exceeding 16% and most models have Sharpe ratios exceeding 1. In comparison, the buy-and-hold SPY returns (Sharpe ratios) were 7% (0.46).

Figure 2 plots the cumulative log returns for each model and of the SPY from 2001 to 2016. Every model has higher cumulative returns than the market portfolio with the RF and ensemble models having significantly higher cumulative returns. In summary, the consistency of the results across models supports the hypothesis that the intraday market is predictable and demonstrates that the predictability is economically significant. Next we show that profitability remains even

with (large) transaction costs.

3.3 Economic significance after transaction costs

The economic significance established above does not account for transaction costs, which are substantial when trading at a five-minute frequency. If models are only accurate for small return fluctuations but cannot forecast large returns, then they are not useful to traders. This section shows that the predictability of the market portfolios is economically significant even after accounting for transaction costs. Additionally, we documented above that after 2005, R^2_{OOS}, excess returns, and transaction costs significantly decreased at the same time, so it is not obvious if model predictions remain profitable after accounting for transaction costs in this recent sample.

Table 3 presents the annualized excess returns and Sharpe ratios of the market timing strategy with transaction costs along with the average percentage of executed trades. Panel A of Table 3 reports the results for the entire sample from 1993 to 2016. Columns (1) - (8) presents results for the eight machine learning models and columns (9) and (10) report the intraday SPY and the benchmark buy-and-hold SPY returns respectively. As expected, all returns are significantly lower due to transaction costs and infrequent trading. However, even after accounting for transaction costs, every model (with the exception of ANN) has positive returns. Among non-ensemble models, PCR and RF have the highest returns (Sharpe ratios), respectively yielding 5% and 6% (0.68 and 0.77). The two ensemble models have high returns (Sharpe ratios), both yielding 6% (0.67 and 0.98). As a benchmark, a buy-and-hold SPY strategy has 7% return and a Sharpe ratio of 0.48. Thus, even after transaction costs, the PCR, RF, and ensemble models outperform holding the market.[18] These findings demonstrate that such models can predict large returns that exceed the transaction costs very well. Our results are similar in magnitude to the strategies in Gao, Han, Li, and Zhou (2018) and Chinco, Clark-Joseph, and Ye (2019) that have after-transaction cost annualized returns of 4.46% (Sharpe ratio of 0.98) and 4.92%, respectively. Consistent with the previous sections, ANN predictions have a high variance and hence fail to forecast large returns. Similarly, the linear LAS and EN models had high R^2_{OOS}, but the statistical predictability did not

[18] When regressing the returns after transaction costs of our trading strategy onto the benchmark buy-and-hold SPY we find that the alphas are positive and statistically significant for the PCR, RF, and ensemble models. The betas, on the other hand, are mostly small, and if they are large and significant, they tend to be negative indicating that our trading strategy if anything hedges systematic risk. Thus, there is strong evidence that our trading strategy out-performs the benchmark portfolio and little evidence that our models are simply buying systematic risk.

Table 3: Excess Returns and Sharpe ratios with transaction costs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
Panel A: Overall Period (1993 - 2016)										
Return	0.03	0.01	0.05	0.06	0.02	-0.07	0.06	0.06	-0.03	0.07
t-stat	1.71	0.77	3.29	2.70	1.10	-3.87	3.38	4.27	-1.11	2.19
SR	0.42	0.20	0.68	0.77	0.24	-0.66	0.67	0.98	-0.24	0.48
% Trades	11.63	12.73	11.45	12.39	14.86	19.26	11.21	11.00		
Panel B: 1/8 Tick Size Period (1993 - 1996)										
Return	0.04	0.05	0.04	0.01	0.07	0.03	0.05	0.04	-0.01	0.11
t-stat	4.79	4.86	2.68	1.22	2.82	2.77	4.98	4.28	-0.37	2.62
SR	2.50	2.60	1.48	0.80	2.27	1.20	2.70	2.34	-0.17	1.19
% Trades	0.47	0.60	1.53	0.27	1.25	0.75	0.52	0.48		
Panel C: 1/16 Tick Size Period (1997 - 2000)										
Return	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	-0.14	0.12
t-stat	1.83	1.67	1.48	1.88	0.46	1.38	2.15	2.42	-1.78	1.62
SR	1.01	0.95	0.75	0.97	0.21	1.14	1.04	0.98	-0.89	0.71
% Trades	0.11	0.12	0.07	0.13	0.20	0.25	0.08	0.10		
Panel D: Early Post-Decimalization Period (2001 - 2004)										
Return	0.14	0.13	0.00	0.11	0.06	-0.06	0.17	0.17	-0.04	-0.01
t-stat	3.17	3.05	0.05	2.09	0.85	-1.29	3.17	3.70	-0.63	-0.12
SR	2.00	1.97	0.02	1.20	0.49	-0.61	2.20	2.19	-0.35	-0.07
% Trades	10.57	11.31	7.23	12.76	15.01	17.54	10.05	10.05		
Panel E: Late Post-Decimalization Period (2005 - 2016)										
Return	-0.01	-0.04	0.09	0.08	0.00	-0.14	0.04	0.05	0.01	0.07
t-stat	-0.58	-1.41	3.07	1.95	-0.18	-4.44	1.41	2.32	0.32	1.37
SR	-0.18	-0.49	0.88	0.81	-0.05	-0.98	0.35	0.68	0.09	0.46
% Trades	19.45	21.35	19.88	20.29	24.12	32.18	18.78	18.37		

Reported are the annualized excess returns, t-statistics, and Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). The independent variables are lagged returns of S&P 500 constituents. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

translate well into economic profitability, earning returns (Sharpe ratios) after transaction costs of 3% and 1% (0.42 and 0.20), respectively. We show in Section 4.3 that LAS and EN had highly negative returns during the Subprime Mortgage crisis, since these models generally perform worse when predictors are highly correlated (Wang, Nan, Rosset, and Zhu (2011)).[19]

[19] In unreported results we confirm that when correlations among constituent returns increase LAS and EN perform worse while PCR performs better. Intuitively, LAS and EN may encounter difficulties when stocks are highly correlated. For example, if several highly correlated stocks are relevant for prediction, then LAS may only select one from the group and shrinks the rest to zero (Zou and Hastie (2005)). EN mitigates this issue by using the ridge penalty. However, the ridge penalty forces the estimated coefficients of highly correlated predictors to be close together, which is problematic since the coefficients on our predictor stocks likely have different magnitudes or different signs (Wang, Nan, Rosset, and Zhu (2011)). Conversely, PCR has stronger predictability when correlations increase, since the model creates new predictors that summarize the variation of the constituent stocks.

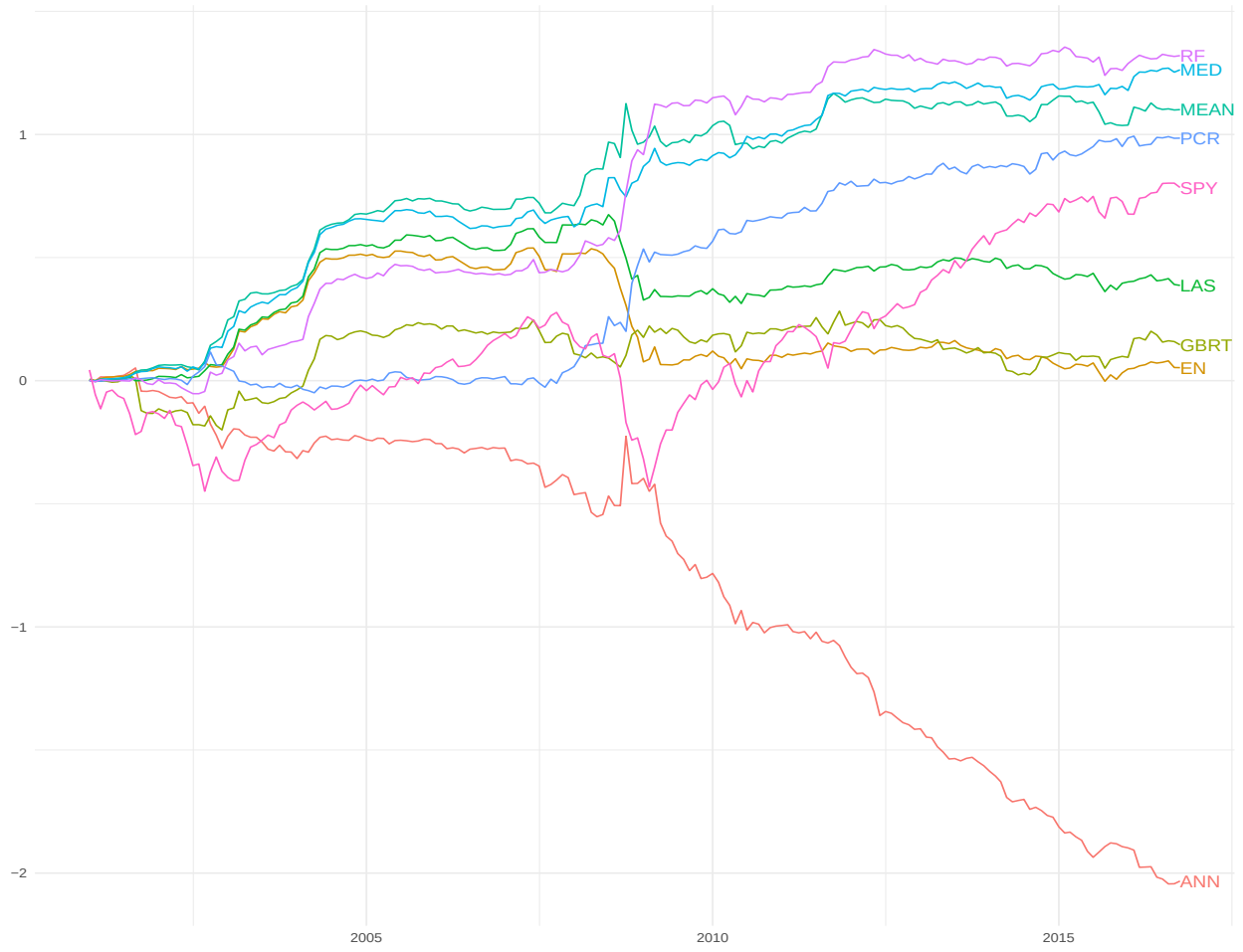
Next, Panel B of Table 3 reports results for the 1/8 tick period from 1993 to 1996. The models returns are low since they only trade roughly 1% of the time. However, the large and positive Sharpe ratios of at least 0.8 indicate model predictions are economically significant even after paying transaction costs. The LAS, EN, GBRT, and ensemble models have Sharpe ratios exceeding 2 and exceeding the benchmark buy-and-hold SPY Sharpe ratio of 1.6. Panel C of Table 3 reports results for the 1/16 tick period from 1997 to 2000, where models trade less than 0.25% of the time due to large transaction costs. Whereas Table 2 showed that pre-transaction cost returns are higher than in the previous period, Table 3 shows that after accounting for transaction costs returns (Sharpe ratios) are now lower than in the previous period. LAS, EN, PCR, RF, ANN, and the ensemble models, though, continue to have higher Sharpe ratios than the buy-and-hold benchmark Sharpe ratio of 0.71.

Finally, Panel D and E reports results for the post-decimalization period from 2001 to 2016, where models trade much more frequently due to lower transaction costs. Panel D reports results for the early post-decimalization period from 2001 to 2004 and shows that excess returns (Sharpe ratios) increase for the LAS, EN, RF, GBRT, and ensemble models relative to the period before 2001 due to the significantly lower transaction costs. All models, except ANN, earn much higher returns (Sharpe ratios) than the benchmark -0.01% (-0.07) for the buy-and-hold SPY. Panel E of Table 3 reports results for the late post-decimalization period from 2005 to 2016. In the previous sections, we documented that this period had substantially lower R^2_{OOS}, returns, and Sharpe ratios relative to previous periods. However, due to the decrease in transaction costs, the after-transaction cost returns and Sharpe ratios are still high. During this period, the benchmark buy-and-hold SPY earned 7% returns with a 0.46 Sharpe ratio. The PCR and RF models out-perform the benchmark with returns (Sharpe ratios) of 9% and 8% (0.88 and 0.81), respectively, which we in Section 4.3 show is partially due to their strong performance during the Subprime mortgage crisis. The ensemble median also beat the buy-and-hold SPY with returns (Sharpe ratios) of 5% (0.68).[20]

Figure 3 plots the cumulative log returns after transaction costs for each model and of the SPY from 2001 to 2016. After accounting for transaction costs, the PCR, RF and ensemble models

[20]These results are robust to assuming fixed transaction costs of, say, 0.01% or 0.1% instead of the bid-ask spread (the median post-decimalization spread was 0.008%). In particular, with 0.01% fixed transaction costs economic profitability generally increases across models and though the models nearly never trade with a 0.1% fixed transaction cost Sharpe ratios remain positive for the LAS, EN, and ensemble models. This verifies that our results are not driven by models that only trade when the spread is low.

Figure 3: Cumulative returns after transaction costs by model



Plot of the post-decimalization cumulative log returns of the market-timing strategy with transaction costs for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED).

have higher cumulative returns relative to the market, despite only trading infrequently. Even accounting for the drop in R^2_{OOS} after 2005, the after-transaction cost returns remain consistently large. That is, even after accounting for transaction costs, the considered models earn economically significant returns with low variance. We demonstrate economic gains available to traders using such model forecasts supporting the hypothesis that markets are predictable at the five-minute frequency. However, the market is notably less predictable post-decimalization, particularly after 2005, as expected, evidenced by the lower returns and Sharpe ratios. Furthermore, the LAS and EN models outperformed PCR and RF prior to the 2008 crisis, but performed poorly during and

after the crisis due to the higher level of correlation among stocks.

3.4 Robustness to training sample size

The baseline model uses an expanding window for training and one month each for validation and for out-of-sample testing. If the predictive relationship that we document is stable, then forecasting accuracy should be increasing in training sample size, since more observations yield lower variance forecasts. However, financial time series are notorious for containing structural breaks, time-varying volatility, and other nonstationarities (Timmermann (2008)). A shorter training sample may therefore outperform a longer one if the empirical model in Equation (2) changes due to this nonstationarity, in which case using a longer training sample may yield biased forecasts. We test the importance of various training periods by evaluating R^2_{OOS} and after-transaction cost Sharpe ratios using 58-, 34-, 22-, 16-, 10-, 7-, 4-, and 1-month samples for training and compare the results with the baseline training duration.

Panel A of Table 4 reports the post-decimalization R^2_{OOS} of the eight machine learning models using different training window lengths. This sample includes several possible structural breaks, including the Subprime mortgage crisis and EU debt crisis. Across nearly all models, R^2_{OOS} is increasing in training size, indicating that the predictive relationship is mostly stable. However, the R^2_{OOS} do not increase monotonically, which demonstrates that non-stationarities do have some effect on forecasts. This is particularly apparent comparing the 58-month estimation window to expanding, suggesting that results could be improved by starting the expanding window later in the sample. Furthermore, the R^2_{OOS} are positive for the ensemble models across nearly all training periods and positive for most individual models. Panel B of Table 4 reports the after-transaction cost Sharpe ratios using different training window lengths for the post-decimalization period. Consistent with the results for the R^2_{OOS}, model Sharpe ratios are mostly increasing in training size across models.

The results in Table 4 first of all show that our predictability findings for intraday market returns are largely robust to using different training window specifications. Secondly, they importantly show that predictability increases with the training window size. Chinco, Clark-Joseph, and Ye (2019) theorize that market predictability could be driven by very short-term sparse signals. Our results indicate instead that predictability may be consistently exploiting inefficiencies across time and is

Table 4: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by training duration post-decimalization

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Post-Decimalization Out-of-Sample R^2								
Training months	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Expanding	0.30	0.22	0.04	0.26	-0.12	-0.19	0.35	0.35
58 month	0.33	0.30	-0.06	0.25	-0.13	-0.15	0.38	0.39
34 month	0.23	0.26	0.09	0.23	-0.29	-0.20	0.36	0.38
22 month	0.29	0.27	0.07	0.22	-0.42	-0.30	0.33	0.34
16 month	0.19	0.18	0.08	-0.01	-0.39	-0.36	0.28	0.30
10 month	0.07	0.05	0.07	-0.06	-0.36	-0.29	0.22	0.24
7 month	0.05	0.05	0.00	-0.03	-0.36	-0.42	0.20	0.21
4 month	0.05	0.05	-0.08	-0.22	-0.63	-0.65	0.12	0.14
1 month	-0.09	-0.09	-0.33	-0.68	-1.35	-0.40	-0.08	-0.04
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs								
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Expanding	0.34	0.06	0.72	0.90	0.12	-0.90	0.69	1.04
58 month	0.26	0.21	-0.16	0.85	0.23	-0.73	0.80	0.81
34 month	-0.07	-0.21	0.19	0.69	-0.30	-0.72	0.45	0.56
22 month	-0.13	-0.26	0.43	0.19	-0.30	-1.27	0.50	0.37
16 month	-0.10	-0.28	0.41	0.06	-0.24	-1.04	0.41	0.57
10 month	-0.11	-0.14	0.27	-0.43	-0.64	-0.54	0.50	0.55
7 month	-0.04	-0.06	0.28	-0.03	-0.22	-0.63	0.59	0.61
4 month	-0.41	-0.40	0.02	-0.30	-0.66	-0.89	0.34	0.15
1 month	-0.16	-0.15	-0.40	-1.50	-2.13	-1.29	-0.29	0.09

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare R^2 values across 1-, 4-, 7-, 10-, 16-, 22-, 34-, and 58-month training windows. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

not necessarily driven by infrequent signals.

4 Additional analysis

Our hypothesis for intraday predictability is based on slow-moving trader capital. In this section we examine if our results differ within the trading day, in periods of high versus low volatility or illiquidity, and during periods of financial crisis. Finally, we compare our baseline model performance to that of autoregressive models for the market return and we consider whether forecasting accuracy may be improved by including additional lagged variables.

Table 5: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by time post-decimalization

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Post-Decimalization Out-of-Sample R^2									
Time	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	
9:35 - 10:00	-0.06	-0.21	0.21	-0.20	-1.26	-0.72	0.12	0.19	
10:00 - 10:30	0.34	0.26	0.05	0.09	-0.27	-0.17	0.33	0.36	
10:30 - 11:00	0.38	0.40	0.10	0.55	0.29	-0.02	0.56	0.49	
11:00 - 11:30	0.41	0.32	0.03	0.35	0.20	-0.07	0.49	0.47	
11:30 - 12:00	0.32	0.24	-0.02	0.29	-0.22	-0.11	0.38	0.38	
12:00 - 12:30	0.47	0.41	-0.01	0.48	0.32	-0.21	0.54	0.51	
12:30 - 13:00	0.39	0.26	0.24	0.48	-0.20	-0.46	0.40	0.45	
13:00 - 13:30	0.31	0.33	0.10	0.43	-0.05	-0.28	0.38	0.31	
13:30 - 14:00	0.37	0.26	0.01	0.34	-0.04	-0.07	0.38	0.39	
14:00 - 14:30	0.99	0.96	0.19	0.62	0.65	0.62	0.90	0.84	
14:30 - 15:00	0.42	0.35	0.17	0.47	0.01	0.12	0.44	0.47	
15:00 - 15:30	-0.16	-0.27	-0.26	-0.08	-0.32	-0.77	-0.13	-0.09	
15:30 - 15:55	0.10	-0.01	-0.26	0.13	0.11	-0.11	0.17	0.14	
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs									
Time	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY
9:35 - 10:00	-0.20	-0.35	0.23	0.26	-0.64	-0.63	0.55	0.21	-0.25
10:00 - 10:30	0.57	0.47	0.60	0.08	0.19	-0.19	0.50	0.69	-0.02
10:30 - 11:00	0.18	0.23	0.35	0.07	0.17	0.03	0.62	0.31	-0.21
11:00 - 11:30	-0.11	-0.30	-0.08	0.00	0.18	-0.17	0.08	-0.08	-0.29
11:30 - 12:00	0.13	0.15	0.21	0.52	0.08	-0.23	0.40	0.32	0.12
12:00 - 12:30	0.00	-0.09	0.44	0.38	0.20	-0.50	0.28	0.24	-0.04
12:30 - 13:00	0.22	0.32	0.40	0.39	-0.52	-0.89	0.18	0.62	0.47
13:00 - 13:30	0.13	0.11	0.55	0.74	0.23	-0.29	0.18	0.57	0.16
13:30 - 14:00	0.17	0.00	0.09	0.39	-0.26	-0.59	0.34	0.34	-0.31
14:00 - 14:30	0.01	0.19	0.51	0.44	0.64	-0.17	0.27	0.34	-0.23
14:30 - 15:00	0.31	0.17	0.49	0.52	0.24	-0.30	0.35	0.59	0.43
15:00 - 15:30	-0.13	-0.30	-0.30	-0.34	-0.51	-0.99	-0.59	-0.39	0.27
15:30 - 15:55	-0.11	-0.28	-0.44	0.43	0.40	0.27	-0.05	-0.11	-0.09

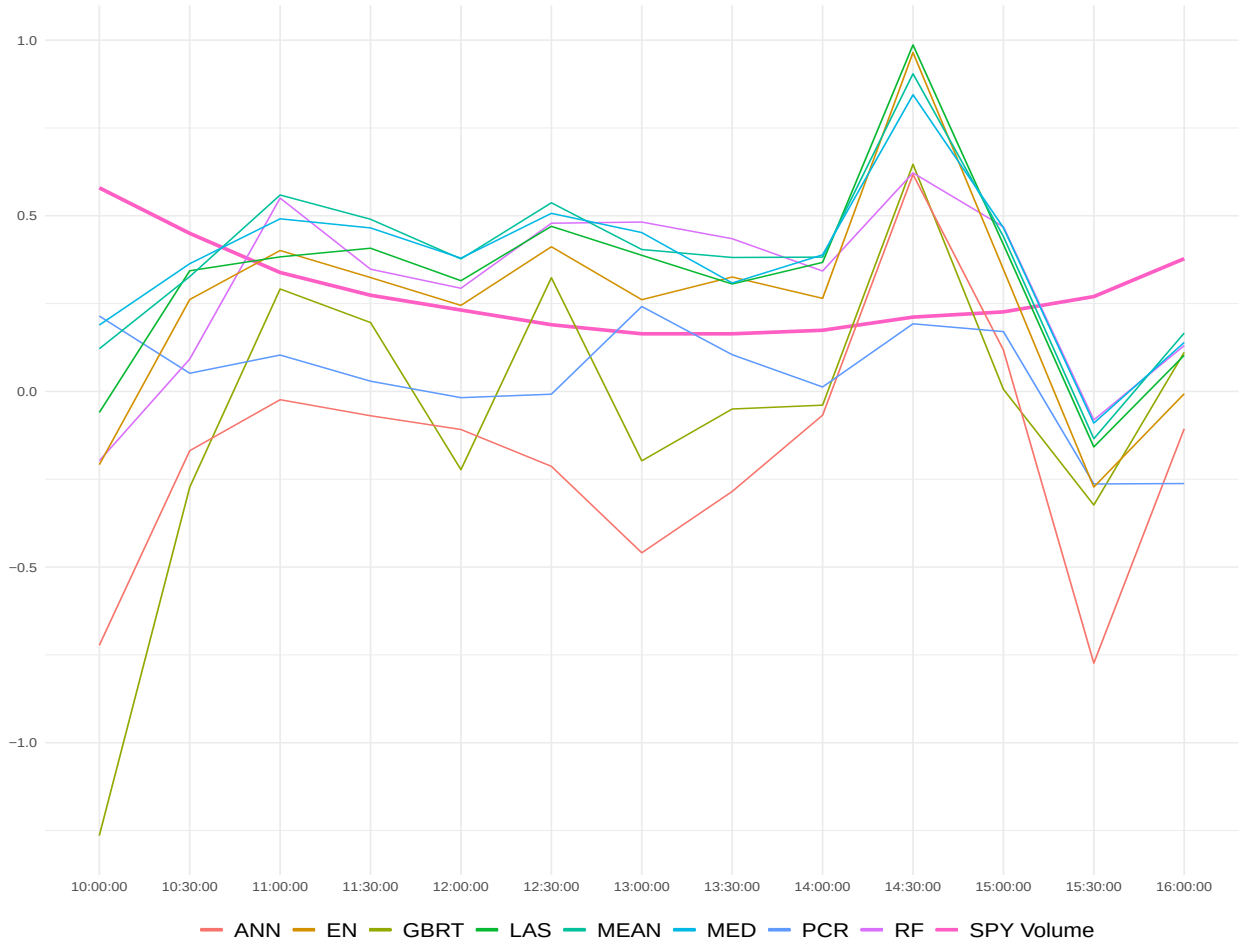
Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare results across 30-minute windows throughout the trading day. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

4.1 Time-of-day patterns

This section tests the intraday implications of slow traders on intraday predictability. Since traders are most active at the beginning and end of each day, we expect predictability to be low during those times. Also, since intraday trading volume exhibits a "U" shape, we expect predictability to exhibit an inverse-U shape.

Panel A of Table 5 reports the R^2_{OOS} of each half-hour interval in the post-decimalization period.

Figure 4: R^2_{OOS} and trading volume



Plot by time of day of the post-decimalization R^2_{OOS} for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). Also plotted is the median SPY trading volume in millions.

For every model, the R^2_{OOS} is low in the first half hour and last hour of the day, when traders are most active. On the other hand, these models have high R^2_{OOS} between 10:00 - 15:00, demonstrating that nearly all of the predictability occurs during the middle of the day when traders are less active. Panel B of Table 5 reports after-transaction cost Sharpe ratios of each half-hour interval in the post-decimalization period. For most models, the Sharpe ratios are highest between 10:00 - 15:00, demonstrating that the increased R^2_{OOS} translates into economically meaningful returns after transaction costs. The Sharpe ratio for the intraday SPY in column (9) does not display the same pattern, showing that our models are not simply buying the SPY portfolio.

Figure 4 plots the R^2_{OOS} and median trading volume in each half-hour interval in the post-decimalization period. Across all models (except ANN), there is a remarkably similar inverse-U pattern, suggesting that these models are approximating the same function. Every model's R^2_{OOS} jumps up between 14:00 - 14:30, suggesting that this is the least active time for traders. Every model's R^2_{OOS} jumps down between 15:00 - 15:30, suggesting that traders are active at this time, which is consistent with Lou, Polk, and Skouras (2019)'s finding that institutional investors tend to initiate trades near the close. Our discovery that predictability is stronger when traders are less active is consistent with our hypothesis that the predictability is driven by slow traders.

Gao, Han, Li, and Zhou (2018) find that returns in the first half hour forecast returns in the last half hour. The results in Table 5 show that the predictability of our models are orthogonal to their findings and complement their paper meaningfully. Economically, there are likely several sources of risk and trading behaviors driving intraday patterns in returns.

4.2 Volatility and illiquidity effects

During periods of high volatility or illiquidity (i.e. periods of low liquidity), traders encounter higher market frictions. According to our slow trader hypothesis, predictability should be higher when volatility or illiquidity is higher. We test this hypothesis by first sorting days into 3 equal groups (tertiles) based on their daily realized volatility and Amihud (2002) measure of illiquidity, respectively. We then study the predictability individually for each tertile.

Panel A of Table 6 reports the R^2_{OOS} of each volatility group during the post-decimalization period. Consistent with our hypothesis, R^2_{OOS} is strictly increasing in volatility across every model (except PCR). Likewise, Panel B of Table 6 shows that after-transaction cost Sharpe ratios are increasing in volatility for most models. Note that column (9) shows that the intraday SPY Sharpe ratio decreases in volatility, so model Sharpe ratios relative to the benchmark are strongly increasing across the volatility tertiles.[21]

Panel A of Table 7 reports the R^2_{OOS} of each illiquidity group during the post-decimalization period. Interestingly, R^2_{OOS} is small for low and mid illiquidity days, but extremely large only for

[21] To examine this further, we considered a volatility timing strategy that trades only when the previous day's volatility is in a given tertile group using only past information to create the groups, i.e. with no look-ahead bias. The results show that Sharpe ratios are increasing in volatility for all models except LAS and EN, demonstrating that our ex-post volatility analysis could be converted into a tradeable volatility timing strategy.

Table 6: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by volatility in the post-decimalization period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Post-Decimalization Out-of-Sample R^2									
Volatility	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	
Low	0.09	-0.02	-0.03	0.02	-0.41	-0.76	0.12	0.17	
Mid	0.19	0.11	-0.09	0.11	-0.11	-0.37	0.24	0.25	
High	0.34	0.26	0.07	0.31	-0.09	-0.11	0.40	0.39	
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs									
Volatility	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY
Low	0.47	0.09	0.20	0.23	-0.81	-1.79	-0.03	0.44	2.88
Mid	0.40	0.39	0.43	0.28	0.26	-1.08	0.81	0.89	-0.20
High	0.14	-0.15	0.73	1.04	0.23	-0.33	0.60	0.87	-0.93

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare results across 3 groups sorted on realized volatility. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

Table 7: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by illiquidity in the post-decimalization period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Post-Decimalization Out-of-Sample R^2									
Illiquidity	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	
Low	0.00	-0.08	0.06	-0.06	-0.41	-0.46	0.04	0.04	
Mid	0.01	-0.04	0.02	0.02	-0.43	-0.34	0.05	0.07	
High	0.76	0.66	0.04	0.69	0.37	0.11	0.85	0.83	
Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs									
Illiquidity	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY
Low	-0.25	-0.43	0.42	0.11	-0.34	-1.08	0.24	0.26	0.33
Mid	-0.02	-0.21	0.93	0.83	0.11	-0.47	0.23	0.59	-0.22
High	1.03	0.85	0.18	1.03	0.47	-0.56	1.50	1.22	-0.07

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the market-timing strategy with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). We compare results across 3 groups sorted on Amihud (2002) illiquidity. The independent variables are lagged returns of S&P 500 constituents. Reported are results for the *post-decimalization* subsamples.

high illiquidity days. This suggests that most of the predictability occurs during days when the market is illiquid and when it's more expensive for traders to rebalance their portfolios. Panel B of Table 7 reports after-transaction cost Sharpe ratios of each illiquidity group during the post-decimalization period. For most models, Sharpe ratios are increasing in illiquidity.

In summary, predictability and economic significance increase when market volatility and illiquidity are high. These findings are similar to Gao, Han, Li, and Zhou (2018) who show that 30-minute predictability is higher on high volatility and illiquidity days. Our results are consistent with their findings and with our hypothesis that market predictability is driven by slow-moving capital from traders that face market frictions. These findings also support the argument in Chordia, Roll, and Subrahmanyam (2005) and Chordia, Roll, and Subrahmanyam (2008) that prices can be separated from fundamentals and are predictable in short horizons due to insufficient liquidity.

4.3 Impact of financial crisis

The recent 2005 to 2016 period contains the Subprime mortgage and European debt crises. These crisis periods are associated with significant market frictions. According to our hypothesis, predictability should be higher during these crisis periods. This is tested in this section by studying several interesting sub-periods of the late post-decimalization period illustrating crisis and non-crisis periods.

Panel A of Table 8 reports the R^2_{OOS} of crisis and non-crisis periods. Across most models, the R^2_{OOS} during the Subprime and EU crises are higher than during the 2010 - 2011 and 2014 - 2016 periods. However, the R^2_{OOS} was high during the earliest 2005-2007 period, possibly due to technological factors (i.e. less sophisticated trading). PCR had the highest R^2_{OOS} during the Subprime mortgage crisis and a positive R^2_{OOS} during the EU debt crisis. This finding confirms our assertion in Section 3.1 that PCR has stronger predictability when equities are more correlated and follows a strong factor structure, as is the case during a financial crises. From 2014 - 2016, model predictability is negative across models. This may be because the period was relatively stable, but could also be caused by reduced trader frictions.[22] These results demonstrate that the predictability is higher during a financial crisis. In the slow moving theory of capital, traders may not invest in arbitrage opportunities if they have insufficient capital or can invest in assets with higher expected returns (Duffie (2010)). These results suggest that during the financial crisis, the market portfolio may have been particularly predictable due to slow moving capital. These findings are similar to Gao, Han, Li, and Zhou (2018), who show 30-minute predictability is higher during

[22] It will be interesting to study if predictability increases during the ongoing Coronavirus crisis once high-frequency data becomes available for this period.

Table 8: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) by crisis period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Post-Decimalization Out-of-Sample R^2										
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED		
2005-2007	0.07	-0.04	-0.01	0.00	-0.11	-0.30	0.12	0.10		
Subprime	0.04	-0.02	0.08	0.02	-0.37	-0.48	0.06	0.07		
2010-2011	-0.02	-0.05	0.03	-0.03	-0.67	-0.39	-0.03	0.02		
EU Debt	0.08	-0.01	0.01	0.03	-0.30	-0.70	0.06	0.10		
2014-2016	-0.07	-0.17	-0.04	-0.14	-0.36	-0.41	-0.03	0.00		
Panel B: Post-Decimalization Sharpe Ratio before Transaction Costs										
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
2005-2007	1.64	1.48	1.65	1.82	2.08	2.06	2.38	1.97	-1.09	0.56
Subprime	1.06	0.96	3.11	3.78	2.67	0.94	2.22	1.91	0.07	-0.40
2010-2011	2.18	2.07	2.22	1.87	2.32	1.14	2.42	3.22	0.20	0.54
EU crisis	3.07	2.88	2.82	3.08	2.30	0.84	3.52	3.08	1.34	2.29
2014-2016	0.79	0.94	1.70	1.68	1.88	0.51	0.91	1.35	0.61	0.68
Panel C: Post-Decimalization Sharpe Ratio after Transaction Costs										
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED	Intraday SPY	Hold SPY
2005-2007	0.35	-0.06	0.17	0.08	-0.21	-0.92	0.16	-0.04	-1.09	0.56
Subprime	-1.09	-1.55	1.38	2.00	-0.12	-0.52	0.73	0.87	0.07	-0.40
2010-2011	0.77	0.30	1.97	1.07	0.38	-1.35	0.69	1.82	0.20	0.54
EU crisis	0.63	-0.22	0.68	0.14	-1.05	-2.68	-0.09	0.57	1.34	2.29
2014-2016	-0.58	-0.38	0.64	0.12	0.17	-2.18	-0.06	0.40	0.61	0.68

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy without and with transaction costs. Models include the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). The independent variables are lagged returns of S&P 500 constituents. Reported are results for the late post-decimalization subsample which include the Subprime Mortgage crisis, during 2008-2009, and the European Debt crisis, during 2012-2013.

the Subprime mortgage crisis.

The economic significance results of crisis and non-crisis periods are somewhat mixed. Panel B of Table 8 reports the pre-transaction cost Sharpe ratios of crisis and non-crisis periods. The pre-transaction cost Sharpe ratios are mostly consistent with the R^2_{OOS} results, with the exception of the 2010 - 2011 period which earned higher Sharpe ratios than expected given the low R^2_{OOS}. Panel C of Table 8 reports the after-transaction cost Sharpe ratios. Surprisingly, the results differ from the findings in Panel A and B for the 2005 - 2007 period and the EU debt crisis period. It appears that high transaction costs during these periods removed most of the predictive profits. Similarly, the 2010 - 2011 period had the highest after-transaction cost Sharpe ratios, potentially due to reduced transaction costs after the Subprime crisis. Interestingly, the LAS and EN models had extremely low returns during the financial crisis, likely because these models often perform

poorly when predictors are highly correlated as explained in Section 3.3.

4.4 Comparison to autoregressive models

A natural comparison to our cross-sectional models with lagged intraday returns for the market constituents are autoregressive models for the market return itself, since our predictability results could simply be capturing intraday momentum. For example, Heston, Korajczyk, and Sadka (2010) find significant auto-correlation of half-hour returns at daily intervals and Gao, Han, Li, and Zhou (2018) find that the first half-hour return of the SPY predicts the last. This section evaluates the predictability of linear models estimated on up to 500 lagged SPY returns using the same specifications as our baseline machine learning models. We consider a simple AR(1) that uses the SPY return with 1 lag (i.e. our baseline model without constituent returns) and an AR(p) model where the number of lagged returns (up to 500) are chosen to minimize the validation mean squared error. We also consider the linear LAS, EN, and PCR models to perform dimension reduction on the 500 lagged SPY returns.

Panel A of Table 9 reports the R^2_{OOS} of the linear autoregressive models together with the linear baseline models for comparison. During the overall period from 1993 to 2016, LAS and EN generally have the highest R^2_{OOS} among AR models of 0.6% and 0.58%, suggesting that certain market return lags contain predictive information. However, these regularized models only slightly improve on the simplest AR(1) model's R^2_{OOS} of 0.55%, indicating that the most recent lag is the most important for prediction. This is reinforced by the AR(p), which uses a median of 29 lags and performs worse than the AR(1) model with an R^2_{OOS} of 0.53%. The AR(500) model without dimension reduction, i.e. estimated with OLS, performs poorly due to the high-dimensional inputs as expected. In comparison, our baseline LAS and EN models significantly outperform every AR model with an R^2_{OOS} of 2% and 1.95% respectively. Additionally, the baseline PCR's R^2_{OOS} of 0.31% is higher than the AR(500) PCR model's R^2_{OOS} of -0.07%. These results demonstrate that there is significant predictive information embedded in the lagged returns of the S&P 500 constituents.

Panel B of Table 9 reports the after-transaction cost Sharpe ratios of the linear autoregressive models together with the linear baseline models for comparison. During the overall period from 1993 to 2016, every AR model has a negative Sharpe ratio except for the AR(1), which has a Sharpe ratio of 0. In contrast, the baseline linear models all have positive Sharpe ratios, with the

Table 9: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) for autoregressive models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Out-of-Sample R^2									
	Autoregressive Models						Baseline Models		
	AR(1)	AR(p)	OLS	LAS	EN	PCR	LAS	EN	PCR
1993-2016	0.55	0.53	-0.49	0.58	0.60	-0.07	2.00	1.95	0.31
1993-1996	0.22	0.37	-4.51	0.12	0.40	-0.04	9.27	9.32	6.52
1997-2000	1.87	1.95	1.07	1.78	1.79	-0.01	3.72	3.69	-0.76
2001-2004	0.17	0.08	-0.81	0.30	0.27	-0.13	0.91	0.81	0.03
2005-2016	-0.10	-0.20	-0.67	-0.02	-0.03	-0.08	0.02	-0.04	0.04
Panel B: Sharpe Ratio after Transaction Costs									
	Autoregressive Models						Baseline Models		
	AR(1)	AR(p)	OLS	LAS	EN	PCR	LAS	EN	PCR
1993-2016	0.00	-0.63	-1.25	-0.27	-0.38	-0.03	0.42	0.20	0.68
1993-1996	-0.49	-0.71	-0.56	-0.68	-0.68	0.14	2.50	2.60	1.48
1997-2000	0.72	0.73	0.41	0.49	0.65	0.33	1.01	0.95	0.75
2001-2004	0.10	-0.39	-2.08	0.63	0.27	0.38	2.00	1.97	0.02
2005-2016	-0.04	-0.91	-1.70	-0.50	-0.62	-0.10	-0.18	-0.49	0.88

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY market-timing strategy with transaction costs. Models include the AR(1), AR(p), lasso (LAS), elastic net (EN), and principal component regression (PCR). The independent variables are 500 lagged returns of S&P 500. Results are reported for the full, 1/16 tick size, 1/8 tick size, early post-decimalization, and late post-decimalization samples.

PCR's Sharpe ratio of 0.68 even outperforming the buy-and-hold S&P 500's Sharpe ratio of 0.48. These results demonstrate that using the lagged returns of the S&P 500 constituents is necessary for improving the economic significance of our predictions. The panel also shows that this holds true for all the sub-periods considered in this paper.

4.5 Effect of additional variables

Finally, we consider whether including additional lagged characteristics of the S&P 500 constituents may improve forecasting accuracy. Due to computational constraints (both in terms of memory requirements and training time), we only consider results using the four-month training sample during the post-decimalization period. We consider characteristics that proxy short-term changes in liquidity and trading trends. The characteristics include firm-level market beta, momentum, maximum, minimum, volatility, illiquidity, trading volume, kurtosis, and skewness calculated over preceding days as well as the lagged observed bid-ask spread.

Table 10: Market predictability (percentage R^2_{OOS}) and profitability (Sharpe ratio) with additional characteristics in the post-decimalization period

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Post-Decimalization Out-of-Sample R^2							
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Base	0.05	0.05	-0.08	-0.22	-0.63	-0.65	0.12	0.14
BETA	-0.98	-1.41	-2.44	-1.92	-1.54	-4.10	-0.39	-0.03
ILLIQ	0.09	-0.44	-8.73	-2.70	-7.74	-3.22	-0.62	0.05
KURT	-0.64	-2.27	-1.1E+18	-2.12	-1.62	-4.25E+23	-1.18E+22	-0.01
MAX	0.03	-0.50	-0.10	-0.56	-1.14	-1.27	-0.02	0.09
MIN	-0.37	-0.36	-0.10	-5.53	-0.68	-1.38	-0.25	0.01
MOM	0.07	-1.24	-0.07	-0.42	-2.59	-1.05	-0.08	0.08
SKEW	0.06	-1.13	-4.8E+12	-0.72	-1.59	-1.8E+15	-5.5E+13	-0.01
VOL	-0.02	-0.35	-0.16	-1.69	-2.26	-1.09	-0.07	0.07
VOLUME	-1.95	-2.24	-0.06	-3.05	-7.49	-1.35	-0.47	-0.24
SPREAD	-1.51	-1.20	-3.67	-0.12	-0.42	-1.40	-0.08	0.13
	Panel B: Post-Decimalization Sharpe Ratio after Transaction Costs							
	LAS	EN	PCR	RF	GBRT	ANN	MEAN	MED
Base	-0.41	-0.40	0.02	-0.30	-0.66	-0.89	0.34	0.15
BETA	-0.27	-0.30	0.14	-0.80	-2.18	-1.38	-0.17	-0.08
ILLIQ	-0.29	-0.36	-0.48	-1.12	-2.77	-1.29	-0.95	-0.18
KURT	-0.47	-0.46	0.22	-0.90	-1.74	-1.19	-0.07	0.11
MAX	-0.21	-0.31	-0.27	-0.71	-1.76	-1.00	-0.53	0.02
MIN	-0.18	-0.18	-0.06	-0.47	-1.28	-0.99	-0.20	0.06
MOM	-0.29	-0.32	0.12	-0.80	-1.72	-0.92	-0.18	0.07
SKEW	-0.26	-0.37	0.27	-0.62	-1.58	-1.00	-0.10	0.32
VOL	-0.53	-0.61	-0.21	-0.94	-2.33	-1.14	-0.17	-0.12
VOLUME	-0.24	-0.27	0.28	-1.02	-2.75	-1.11	-0.68	0.08
SPREAD	-0.21	-0.25	-0.03	0.16	-0.56	-1.00	0.17	0.44

Reported are the out-of-sample predictive R^2 percentages and annualized Sharpe ratios for the SPY using the lasso (LAS), elastic net (EN), principal component regression (PCR), random forest (RF), gradient-boosted regression trees (GBRT), neural network (ANN), mean ensemble (MEAN), and median ensemble (MED). In addition to lagged returns (LAG), the independent variables include market beta (BETA), illiquidity (ILLIQ), kurtosis (KURT), maximum (MAX), minimum (MIN), momentum (MOM), skewness (SKEW), volatility (VOL), or volume (VOLUME) of S&P 500 constituents calculated over the previous day as well as the percent bid-ask spread (SPREAD). See Appendix A.2 for details on how these characteristics are calculated.

Table 10 reports R^2_{OOS} when including different characteristics during the post-decimalization period. The first row reports R^2_{OOS} for the baseline model using lagged returns only. The panel shows that the LAS model is improved by including illiquidity, momentum, or skewness, indicating that there may be some index constituents with useful characteristics for predicting the market return. PCR is also slightly improved by adding momentum and volume. However, for most models including any single characteristic reduces the R^2_{OOS}. One possibility for the poor performance of the price trend and liquidity characteristics is that they are estimated over the preceding day and

potentially noisy. However when we include the lagged bid-ask spread, which is not estimated, the R^2_{OOS} also decreases across most models. These result may suggest that including additional characteristics simply adds noise to the model without providing additional predictive information. While adding these characteristics generally does not improve market forecasts beyond the baseline model, it remains possible that such characteristics may help predict individual stock returns.

Panel B of Table 10 reports the after-transaction cost Sharpe ratios when including different characteristics during the post-decimalization period. This panel shows, similarly to Panel A, that including some characteristics may slightly improve the Sharpe ratios of certain models. For example, LAS, EN, PCR, RF, ANN, and the ensemble median can be slightly improved by including some characteristics. In particular, the median ensemble after including the bid-ask spread achieves a Sharpe ratio of 0.44. However, these results are not robust and could just be due to random chance.

One concern may be that including other characteristics may increase the dimensionality of the data and increase estimation error. However, we also analyzed the predictability of only the characteristics, removing lagged returns from the model, and again found no evidence of predictability. In summary, we find little evidence that including other characteristics can improve predictions. Among our tested characteristics, only lagged returns consistently predict market returns.

5 Conclusion

This paper conducts, to our knowledge, the largest study ever of intraday market return predictability using state-of-the-art machine learning models trained on the cross-section of lagged market index constituent returns and other characteristics to forecast five-minute market returns over the longest possible time period. The paper demonstrates that there is significant statistical predictability of intraday market returns and establishes that this return predictability translates into economically significant profits even after accounting for transaction costs. Furthermore, we show that the lagged constituent returns holds significant predictive information that is not contained in lagged market returns or in lagged price trend and liquidity characteristics.

Specifically, the paper shows that regularized linear models such as lasso and elastic nets and nonlinear tree-based models such as random forests yield the largest positive out-of-sample R^2s.

Linear models such as principal component analysis had a high out-of-sample R^2s during the Subprime crisis and EU debt crisis, providing returns that hedge these crisis states. Ensemble models that combine individual model predictions perform the best across time and the return predictability from these models translates into economically significant profits with Sharpe ratios after transaction costs of 0.98. This Sharpe ratio is much higher than what is obtained from holding the index intraday and significantly exceeds the Sharpe ratio of the benchmark buy-and-hold strategy.

Across time, we show that the statistical predictability has suffered somewhat as transaction costs were reduced post-decimalization. We argue that this strongly suggests that predictability could be a result of slow-moving trader capital. Consistent with the hypothesis of slow traders, market returns are also shown to be more predictable during the middle of the day when trading activity is lower, on days with high volatility or high illiquidity where prices can be driven further away from their fundamental values, and during years of financial crisis which are plagued by market frictions. Nevertheless, the best ensemble models retain some predictability and trading based on the model's signals remain profitable throughout the sample, even after adjusting for transaction costs.

Our results provide strong evidence that market returns are predictable over short-horizons. Our empirical findings suggest that further investigation into the economic mechanisms driving such short-horizon predictability is warranted. The late-informed investor explanation in Gao, Han, Li, and Zhou (2018) is supported by our evidence. Another explanation in Chinco and Fos (2019) theorizes that computational complexity of traders' rebalancing introduces noise. We believe this can explain some of our predictability results, wherein models appear able to capture the systematic behaviour of traders' rebalancing. In particular, this could explain the persistent profitability of our models in recent years. However, verifying these economic mechanism requires further investigation which we leave for future research.

References

Alessandretti, L., A. ElBahrawy, L. M. Aiello, and A. Baronchelli (2018): "Anticipating Cryptocurrency Prices Using Machine Learning," *Complexity*, 2018.

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez (2015): "Does Realized Skewness Predict the Cross-Section of Equity Returns?," *Journal of Financial Economics*, 118(1), 135–167.

Amihud, Y. (2002): "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects," *Journal of Financial Markets*, 5(1), 31–56.

Ang, A., and G. Bekaert (2007): "Stock Return Predictability: Is It There?," *The Review of Financial Studies*, 20(3), 651–707.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006): "The Cross-Section of Volatility and Expected Returns," *The Journal of Finance*, 61(1), 259–299.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011): "Maxing Out: Stocks as Lotteries and the Cross-Section of Expected Returns," *Journal of Financial Economics*, 99(2), 427–446.

Barboza, F., H. Kimura, and E. Altman (2017): "Machine Learning Models and Bankruptcy Prediction," *Expert Systems with Applications*, 83, 405–417.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008): "Designing Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76(6), 1481–1536.

Bergstra, J., and Y. Bengio (2012): "Random Search for Hyper-Parameter Optimization," *The Journal of Machine Learning Research*, 13(1), 281–305.

Bianchi, D., M. Büchner, and A. Tamoni (2021): "Bond Risk Premia with Machine Learning," *The Review of Financial Studies*, 34, 1046–1089.

Bishop, C. M. (1995): *Neural Networks for Pattern Recognition*. Oxford University Press.

——— (2006): *Pattern Recognition and Machine Learning*. Springer.

Bogousslavsky, V. (2016): "Infrequent Rebalancing, Return Autocorrelation, and Seasonality," *The Journal of Finance*, 71(6), 2967–3006.

Breiman, L. (1996): "Bagging Predictors," *Machine Learning*, 24(2), 123–140.

——— (2001): "Random Forests," *Machine Learning*, 45(1), 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984): "Classification and Regression Trees," *International Group*, 432, 151–166.

Bryzgalova, S., M. Pelger, and J. Zhu (2019): "Forest Through the Trees: Building Cross-Sections of Stock Returns," *Available at SSRN 3493458*.

Campbell, J. Y., and S. B. Thompson (2008): "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *The Review of Financial Studies*, 21(4), 1509–1531.

Chen, J., W. Wu, and M. L. Tindall (2016): "Hedge Fund Return Prediction and Fund Selection: A Machine-Learning Approach," *Ocasional Papers - Dallas Federal Reserve*, 16, 04.

Chen, L., Z. Da, and X. Zhao (2013): "What Drives Stock Price Movements?," *The Review of Financial Studies*, 26(4), 841–876.

Chen, L., M. Pelger, and J. Zhu (2019): "Deep Learning in Asset Pricing," *Available at SSRN 3350138*.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019): "Sparse Signals in the Cross-Section of Returns," *The Journal of Finance*, 74(1), 449–492.

Chinco, A., and V. Fos (2019): "The Sound of Many Funds Rebalancing," *Available at SSRN 3346352*.

Chong, E., C. Han, and F. C. Park (2017): "Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies," *Expert Systems with Applications*, 83, 187–205.

Chordia, T., R. Roll, and A. Subrahmanyam (2005): "Evidence on the Speed of Convergence to Market Efficiency," *Journal of Financial Economics*, 76(2), 271–292.

——— (2008): "Liquidity and Market Efficiency," *Journal of Financial Economics*, 87(2), 249–268.

Chordia, T., A. Subrahmanyam, and V. R. Anshuman (2001): "Trading Activity and Expected Stock Returns," *Journal of Financial Economics*, 59(1), 3–32.

Day, M.-Y., and J.-T. Lin (2019): "Artificial Intelligence for ETF Market Prediction and Portfolio Optimization," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1026–1033.

Duffie, D. (2010): "Presidential Address: Asset Price Dynamics with Slow-Moving Capital," *The Journal of Finance*, 65(4), 1237–1267.

Dutta, A., S. Kumar, and M. Basu (2020): "A Gated Recurrent Unit Approach to Bitcoin Price Prediction," *Journal of Risk and Financial Management*, 13(2), 23.

Fama, E. F. (1991): "Efficient Capital Markets: II," *The Journal of Finance*, 46(5), 1575–1617.

Feng, G., J. He, and N. G. Polson (2018): "Deep Learning for Predicting Asset Returns," *arXiv preprint arXiv:1804.09314*.

Feng, G., N. Polson, and J. Xu (2019): "Deep Learning in Characteristics-Sorted Factor Models," *Available at SSRN 3243683*.

Fischer, T., and C. Krauss (2018): "Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions," *European Journal of Operational Research*, 270(2), 654–669.

Freund, Y., and R. E. Schapire (1997): "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55(1), 119–139.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1), 1.

Friedman, J. H. (2001): "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, pp. 1189–1232.

——— (2002): "Stochastic Gradient Boosting," *Computational Statistics & Data Analysis*, 38(4), 367–378.

Gao, L., Y. Han, S. Z. Li, and G. Zhou (2018): "Market Intraday Momentum," *Journal of Financial Economics*, 129(2), 394–414.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013): "Combining Expert Forecasts: Can Anything Beat the Simple Average?," *International Journal of Forecasting*, 29(1), 108–121.

Goodfellow, I. J., D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio (2013): "Maxout Networks," *arXiv preprint arXiv:1302.4389*.

Gu, S., B. Kelly, and D. Xiu (2020a): "Autoencoder Asset Pricing Models," *Journal of Econometrics*.

——— (2020b): "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33(5), 2223–2273.

Hajek, P., and A. Barushka (2018): "Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction," in *Proceedings of the 2nd International Conference on Business and Information Management*, pp. 158–162.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Heston, S. L., R. A. Korajczyk, and R. Sadka (2010): "Intraday Patterns in the Cross-Section of Stock Returns," *The Journal of Finance*, 65(4), 1369–1407.

Heston, S. L., and N. R. Sinha (2017): "News vs. Sentiment: Predicting Stock Returns From News Stories," *Financial Analysts Journal*, 73(3), 67–83.

Jegadeesh, N., and S. Titman (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *The Journal of Finance*, 48(1), 65–91.

Jolliffe, I. (2002): *Principal Component Analysis*, Springer Series in Statistics. Springer.

Ke, Z. T., B. T. Kelly, and D. Xiu (2019): "Predicting Returns with Text Data," Discussion paper.

Kelly, B., and S. Pruitt (2015): "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors," *Journal of Econometrics*, 186(2), 294–316.

Kelly, B., S. Pruitt, and Y. Su (2019a): "Instrumented Principal Component Analysis," *Available at SSRN 2983919*.

Kelly, B. T., S. Pruitt, and Y. Su (2019b): "Characteristics are Covariances: A Unified Model of Risk and Return," *Journal of Financial Economics*, 134(3), 501–524.

Koijen, R. S., and S. Van Nieuwerburgh (2011): "Predictability of Returns and Cash Flows," *Annual Review of Financial Economics*, 3(1), 467–491.

Koren, Y. (2009): "The Bellkor Solution to the Netflix Grand Prize," *Netflix Prize Documentation*, 81(2009), 1–10.

Korsaye, S. A., A. Quaini, and F. Trojani (2019): "Smart SDFs," *Available at SSRN 3475451*.

Lintner, J. (1965): "Security Prices, Risk, and Maximal Gains from Diversification," *The Journal of Finance*, 20(4), 587–615.

Long, W., Z. Lu, and L. Cui (2019): "Deep Learning-Based Feature Engineering for Stock Price Movement Prediction," *Knowledge-Based Systems*, 164, 163–173.

Lou, D., C. Polk, and S. Skouras (2019): "A Tug of War: Overnight Versus Intraday Expected Returns," *Journal of Financial Economics*, 134(1), 192–213.

Marković, I., M. Stojanović, J. Stanković, and M. Stanković (2017): "Stock Market Trend Prediction Using AHP and Weighted Kernel LS-SVM," *Soft Computing*, 21(18), 5387–5398.

Marquaridt, D. W. (1970): "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, 12(3), 591–612.

Nair, V., and G. E. Hinton (2010): "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou (2014): "Forecasting the Equity Risk Premium: The Role of Technical Indicators," *Management Science*, 60(7), 1772–1791.

Rapach, D., and G. Zhou (2013): "Forecasting Stock Returns," in *Handbook of Economic Forecasting*, vol. 2, pp. 328–383. Elsevier.

Renault, T. (2017): "Intraday Online Investor Sentiment and Return Patterns in the US Stock Market," *Journal of Banking & Finance*, 84, 25–40.

Roll, R. (1984): "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market," *The Journal of Finance*, 39(4), 1127–1139.

Schapire, R. E. (1990): "The Strength of Weak Learnability," *Machine Learning*, 5(2), 197–227.

Schapire, R. E., and Y. Freund (2013): "Boosting: Foundations and Algorithms," *Kybernetes*.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *The Journal of Finance*, 19(3), 425–442.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, 15(1), 1929–1958.

Sutherland, I., Y. Jung, and G. Lee (2018): "Statistical Arbitrage on the KOSPI 200: An Exploratory Analysis of Classification and Prediction Machine Learning Algorithms for Day Trading," *Journal of Economics and International Business Management*, 6(1), 10–19.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Timmermann, A. (2006): "Forecast Combinations," *Handbook of Economic Forecasting*, 1, 135–196.

——— (2008): "Elusive Return Predictability," *International Journal of Forecasting*, 24(1), 1–18.

Wang, S., B. Nan, S. Rosset, and J. Zhu (2011): "Random Lasso," *The Annals of Applied Statistics*, 5(1), 468.

Weigand, A. (2019): "Machine Learning in Empirical Asset Pricing," *Financial Markets and Portfolio Management*, 33(1), 93–104.

Welch, I., and A. Goyal (2008): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *The Review of Financial Studies*, 21(4), 1455–1508.

Xue, J., S. Zhou, Q. Liu, X. Liu, and J. Yin (2018): "Financial Time Series Prediction Using L2, 1RF-ELM," *Neurocomputing*, 277, 176–186.

Ye, T., and L. Zhang (2019): "Derivatives Pricing via Machine Learning," *Available at SSRN 3352688*.

Yun, H., M. Lee, Y. S. Kang, and J. Seok (2020): "Portfolio Management via Two-Stage Deep Learning with a Joint Cost," *Expert Systems with Applications*, 143, 113041.

Zeiler, M. D. (2012): "Adadelta: An Adaptive Learning Rate Method," *arXiv preprint arXiv:1212.5701*.

Zou, H., and T. Hastie (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

A Data

This appendix provides further details on our trade and quote (TAQ) data cleaning procedures and describes how the additional characteristics used in the empirical analysis are calculated.

A.1 Trade and quotes cleaning

The trade and quotes (TAQ) data require substantial cleaning due to contamination from market microstructure noise. We use the Monthly TAQ Second database up to 2003 and the Daily TAQ Millisecond/Microsecond database from 2004-2016 and filter noisy observations following the procedures similar to Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008). Returns are primarily calculated using prices from the trades database. Bid and ask observations from the quotes database are also used for cleaning trades and calculating transaction costs. For both trades and quotes, entries with time stamps outside of the trading day (9:30 AM to 4:00 PM) are removed. Entries with a zero or negative bid, ask, or price are also removed. For each stock, we keep only entries from the exchange with the highest volume in the month, and delete entries from other exchanges.

For only the quotes database, we construct the national-best bid-ask (NBBO) following procedures from (https://support.sas.com/resources/papers/proceedings16/11201-2016.pdf) using quotes

from all exchanges at a second interval. We remove all entries with a negative spread (ask minus bid) and quotes larger than 50 times the median spread on each day.

For only the trades data, entries with corrected trades (CORR $\neq 0$) are deleted. Entries with an abnormal sale condition are removed, only keeping entries with COND equal to E, F, @, *, @E, @F, *E and *F. Multiple trades with the same second timestamp are replaced with an entry with the median price. Finally, quotes are used to discipline the trade prices: Entries are removed if their price is above the ask plus spread or below the bid minus spread.

The cleaned second-by-second price, bid and ask data is then aggregated into five-minute intervals and merged with the daily center for research in security prices (CRSP) data file by the CUSIP key. Merging the TAQ and CRSP data is challenging, because their tickers often differ. Additionally, tickers change in time due to mergers, acquisitions, and other corporate events. We instead merge the TAQ and CRSP databases using CUSIPs. The CUSIP identifier of each stock is obtained by merging trades with the TAQ master files. Finally, each stock is indexed by CRSP PERMINOs, which are unique and do not change in time.

A.2 Additional firm characteristics

In addition to returns with one lag, we consider other lagged characteristics of the SPY and S&P 500 constituents. These characteristics proxy short-term changes in liquidity and price trends. We create high-frequency analogs of characteristics from the asset pricing literature that have been shown to predict returns, including firm-level market beta, momentum, illiquidity, maximum, minimum, and trading volume. We also consider higher order moments, including volatility, skewness, and kurtosis. These characteristics are all calculated over preceding days. The particular variables used are the following:

Market beta: The market beta is motivated by the capital asset pricing model of Sharpe (1964) and Lintner (1965). At time $1 \leq t \leq T$, given (SPY market) return of stock $1 \leq \ell \leq 500$, r_t^ℓ (r_t^m), it is based on a single factor model given by

$$r_t^\ell = \alpha^\ell + \beta^\ell(r_t^m) + \epsilon_t^\ell,$$

which is computed for each five-minute interval using a rolling window of preceding days, yielding

one overnight and 78 intraday observations. BETA of stock ℓ during each five-minute interval, t, is the least-squares estimate, $\hat{\beta}^{\ell}$.

Momentum: Motivated by Jegadeesh and Titman (1993), the MOM of stock ℓ during each five-minute interval, t, is the cumulative return for the preceding day, yielding one overnight and 78 intraday observations; that is, $\sum_{t=1}^{T} r_t^{\ell}$ with $T = 79$.

Illiquidity: Motivated by Amihud (2002), the ILLIQ of stock ℓ, during each five-minute interval, t, is the ratio of the absolute stock return, $|r_t^{\ell}|$, to the dollar trading volume, averaged in a day excluding the overnight volume, i.e., over the $T = 78$ previous trading intervals in the day. Given the corresponding five-minute trading volume (price times number of shares traded) in dollars at time t, VOLD_t^{ℓ}, this yields

$$\text{ILLIQ}^{\ell} \equiv \frac{1}{T} \sum_{t=1}^{T} \frac{|r_t^{\ell}|}{\text{VOLD}_t^{\ell}}.$$

Maximum (minimum) return: Motivated by Bali, Cakici, and Whitelaw (2011), the MAX (MIN) of stock ℓ during each five-minute interval, t, is defined as the maximum (minimum) five-minute return within the preceding day (i.e., the previous 79 observations), that is, $\max_{1 \le t \le T} r_t^{\ell}$ ($\min_{1 \le t \le T} r_t^{\ell}$).

Trading volume: Motivated by Chordia, Subrahmanyam, and Anshuman (2001), the VOLUME of stock ℓ during each five-minute interval, t, is defined as the average number of shares traded within the preceding day excluding the overnight volume (i.e., the previous 78 observations):

$$\text{VOLUME}^{\ell} \equiv \frac{1}{T} \sum_{t=1}^{T} \text{SharesTraded}_t^{\ell}.$$

Volatility: Motivated by Ang, Hodrick, Xing, and Zhang (2006), the VOL of stock ℓ during each five-minute interval, t, is the standard deviation of five-minute returns within the preceding day (i.e., the previous 79 observations):

$$\text{VOL}^{\ell} \equiv \sqrt{\frac{1}{T} \sum_{t=1}^{T} (r_t^{\ell} - \bar{r})^2} \equiv \text{sd}(r_t^{\ell}).$$

Here, $\bar{r} \equiv \frac{1}{T} \sum_{t=1}^{T} r_t^{\ell}$ denotes the corresponding arithmetic mean.

Skewness (kurtosis): Motivated by Amaya, Christoffersen, Jacobs, and Vasquez (2015), the SKEW (KURT) of stock ℓ during each five-minute interval, t, is the skewness (kurtosis) of five-minute returns within the preceding day (i.e., the previous 79 observations):

$$\text{SKEW}^{\ell} \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{(r_t^{\ell}-\bar{r})^3}{\text{sd}(r_t^{\ell})^3} \qquad \left(\text{KURT}^{\ell} \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{(r_t^{\ell}-\bar{r})^4}{\text{sd}(r_t^{\ell})^4}\right).$$

B Hyperparameters

This appendix discusses in more detail the most relevant hyperparameters that are tuned for each of the models used in the current paper. First, and applicable to all models, inherent in regularization is the particular binary problem of *whether* to regularize or not. Generally, the hyperparameter optimization problem is that of tuning the regularization parameters, λ in the case of lasso (LAS) regularization, with the binary problem of whether this parameters is zero or not. In elastic net (EN) regularization, the convex combination coefficient $\alpha \in [0,1]$ is another hyperparameter to be optimized, with the endpoints $\{0,1\}$ respectively corresponding to the discrete choices of only ridge or only lasso regularization, respectively. In principal component regression (PCR) the number of principal components, $1 \leq \kappa \leq K$, can be considered a hyperparameter to be optimized.[23]

There are two crucial hyperparameters for random forest (RF) models: (1) the number of decision trees (or base learners) in the forest, and (2) the maximum depth of each of these trees, i.e., the maximum path length from any tree root to any of its leaves. More trees in the forest yield greater prediction variance reduction at increased computational cost. Restricting maximum tree depth limits the complexity, both saving computational time and avoiding over-fitting. Similar goals may be achieved less directly, e.g., by limiting the number of training samples per leaf, which generally decreases with increased tree size. The same hyperparameters are important for gradient-boosted regression tree (GBRT) models, in addition to the explicit *learning rate*, ν, which may be factored out from the step sizes or weights γ_i. A smaller learning rate typically leads to better testing performance, but requires more training steps/additional decision trees in the ensemble, to maintain a given training error. Though these are the principal hyperparameters to consider for RF and GBRT models, several other, generally less important ones do exist. For example, bounding

[23] Other heuristics for determining κ is to accept the smallest value which achieves some level of total explained variance or to increase κ so long as the increase in variance explained exceeds a given threshold.

above the number of predictors determining node splits in base learners is relatively important for RF models in particular. Greater bounds permit more complexity, resulting in bias reduction at the expense of increased prediction variance and computational cost. Note that base learner complexity may be directly limited by bounding above the number of node splits or the total number of tree nodes, i.e., the tree size, itself.

For artificial neural networks (ANN), in particular, the loss function used for training is important. Standard choices include mean squared error, used in the current paper, as well as mean absolute error, logarithmic hyperbolic cosine, and a variety of others adapted to more specific scenarios.[24] We guard against model overfitting by adding ℓ^1 and ℓ^2 regularization to the loss function. Other important factors involved in training include the number of iterations, or epochs, and the level of *dropout*, the percentage of training data discarded in each epoch to avoid over-fitting and regularize the optimization problem (see, e.g., Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014) for a discussion).

While the optimization algorithm employed for training, the loss function it uses, the number of epochs, and the level of dropout, all constitute important *hyperparameters* impacting the quality of predictions from ANN models with trained *parameters*, the actual network architecture, i.e. the number of hidden layers, the number of neurons in each hidden layer, and the particular activation function associated with each neuron, may often have an even greater impact on the results. The activation functions used in this paper are the Rectified linear unit (ReLU) given by $\phi(x) = \max\{0, x\}$ (see Nair and Hinton (2010)), the Maxout given by $\phi(x_1, x_2) = \max\{x_1, x_2\}$ (see Goodfellow, Warde-Farley, Mirza, Courville, and Bengio (2013)), and the Hyperbolic tangent given by $\phi(x) = \tanh x$. We use a hyperparameter grid holding the number of hidden layers fixed at 3. Models are estimated using stochastic gradient descent optimization with the ADADELTA (Zeiler (2012)) adaptive learning rate method for faster estimation. We use dropout and early stopping as regularization techniques to prevent overfitting.

Table 11 summarizes the set of hyperparameters that are tuned for each of the models considered in the current paper. To tune these at a particular time, the sample of data is partitioned into training, testing, and validation sets and used specifically in the following way: the optimization

[24] Of particular interest is the Huber loss function, a combination of the ℓ^1 and ℓ^2 norms that permits control of the sensitivity to outliers, used in Gu, Kelly, and Xiu (2020b). Preliminary results show that this metric may outperform the MSE metric used for both the ANN and the GBRT models.

Table 11: Model Hyperparameters

LAS	λ
EN	λ, $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$
PCR	κ = 1-99 by 2
RF	#trees = 100-700 by 100, tree depth = 1-40 by 5, #minimum rows = {1, 25, 50}, columns randomly selected = floor(#number of features * {.05, .15, .25, .333, .4})
GBRT	#trees = 100-700 by 100, tree depth = 2-40 by 3, learning rate = 0.01 - 0.1 by 0.03, sample rate = 0.5-1 by 0.2, column sample rate = 0.1 - 1 by 0.2
ANN	Activation Functions = Relu, Maxout, Tanh; with and without dropout, Hidden nodes = {1000,500,10}, {100,50,10}, {16,14,12},{20,15,10},{25,17,10},{15,10,5}, epochs = {50,100}, dropout ratio = {0, 0.1, 0.2}, max w^2 = {10, 100, 1000, 3.4028235e+38}, ℓ^1 = {0, 0.00001, 0.0001}, ℓ^2 = {0, 0.00001, 0.0001}, rho (rate time decay) = {0.9, 0.95, 0.99, 0.999}, epsilon (rate time smoothing) = {1e-10, 1e-8, 1e-6, 1e-4}

This table describes the set of hyperparameters that are tuned for each individual model considered in the paper.

problem in Equation (5) is repeatedly solved using training predictors and targets, validating at each training iteration to determine whether to terminate training and minimize over-fitting using validation predictors and targets. When it comes to training, all decisions are made according to the goal of minimizing the metric, $m[\cdot]$, modifying internal parameters, like β in the case of the linear models, according to a specific optimization algorithm, e.g., classical stochastic gradient descent in the case of ANN models. When it comes to validation of the hyperparameters the mean squared error from the validation predictors and targets is used. For LAS and EN, λ is selected using coordinate descent (see Friedman, Hastie, and Tibshirani (2010)). For PCR grid search is used. For all other algorithms, random search is used with 50 models. Random search has been shown to be more efficient than grid search, see, e.g., Bergstra and Bengio (2012). In contrast to Gu, Kelly, and Xiu (2020b), we do not use ensembles of neural networks due to computational limitations.