

Estimating stock market betas via machine learning

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Drobetz, W., Hollstein, F., Otto, T. and Prokopczuk, M. (2025) Estimating stock market betas via machine learning. *Journal of Financial and Quantitative Analysis*, 60 (3). pp. 1074-1110. ISSN 1756-6916 doi: 10.1017/S0022109024000036 Available at <https://centaur.reading.ac.uk/117477/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1017/S0022109024000036>

Publisher: Cambridge University Press

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Estimating Stock Market Betas via Machine Learning

Wolfgang Drobetz

University of Hamburg, Faculty of Business Administration
wolfgang.drobetz@uni-hamburg.de (corresponding author)

Fabian Hollstein 

Saarland University, School of Human and Business Sciences
fabian.hollstein@uni-saarland.de

Tizian Otto

University of Hamburg, Faculty of Business Administration
tizian.otto@uni-hamburg.de

Marcel Prokopczuk

Leibniz University, Hannover School of Economics and Management
prokopczuk@fcm.uni-hannover.de

Abstract

Machine learning-based stock market beta estimators outperform established benchmark models both statistically and economically. Analyzing the predictability of time-varying market betas of U.S. stocks, we document that machine learning-based estimators produce the lowest forecast and hedging errors. They also help to create better market-neutral anomaly strategies and minimum variance portfolios. Among the various techniques, random forests perform the best overall. Model complexity is highly time-varying. Historical stock market betas, turnover, and size are the most important predictors. Compared to linear regressions, allowing for nonlinearity and interactions significantly improves predictive performance.

I. Introduction

In single-factor asset pricing models, such as the capital asset pricing model (CAPM) introduced by Sharpe (1964), Lintner (1965), and Mossin (1966), the expected return of a stock in equilibrium is determined solely by its sensitivity to market risk. While multifactor models that include additional factors can explain the cross-sectional variation in expected returns somewhat better than the CAPM (see, e.g., Fama and French (2008), Harvey, Liu, and Zhu (2016), for extensive evidence), it explains the time series variation in returns well. Moreover, as shown

We thank two anonymous referees, Yakov Amihud, Michael Bauer, Wolfgang Bessler, Axel Cabrol, Mikhail Chernov, Hubert Dichtl, Thierry Foucault (the editor), Kay Giesecke, Lisa Goldberg, Valentin Haddad, Barney Hartman-Glaser, Bernard Herskovic, Bryan Kelly, Serhiy Kozak, Markus Leippold, Martin Lettau, Lars A. Lochstoer, Harald Lohre, Tyler Muir, Andreas Neuhierl, Terrance Odean, Stavros Panageas, Markus Pelger, Tatjana Puhane, Carsten Rother, Boris Valtée, Michael Weber, and Ivo Welch for helpful comments and suggestions.

by Graham and Harvey (2001), Jacobs and Shivdasani (2012), and Graham (2022), the CAPM is widely used in the industry. The vast majority of chief financial officers of large U.S. firms rely on a 1-factor market model to estimate their cost of equity capital. For this application, firms typically estimate market betas as the main component and treat the market risk premium as an almost free parameter (Cochrane (2011), Jacobs and Shivdasani (2012)). Investors, in turn, use the market betas for capital allocation decisions and portfolio risk management (Barber, Huang, and Odean (2016), Berk and van Binsbergen (2016), and Daniel, Mota, Rottke, and Santos (2020)).

However, there are two main problems with using the CAPM, and hence stock market betas, for these applications: Betas i) are not directly observable, which underscores the need for accurate estimates, and ii) are time-varying (Campbell, Lettau, Malkiel, and Xu (2001)). The second problem complicates matters considerably, because most of the above applications require accurate predictions of future betas. Therefore, the main goal of both researchers and practitioners is to find approaches that provide reliable estimates of future betas with minimal prediction error.

Our objective is to examine whether machine learning-based models outperform established approaches in estimating time-varying market betas, and if so, why. In our empirical analysis, we use i) a large universe of U.S. stocks, ii) a long and recent sample period, iii) a broad set of both benchmark and machine learning-based beta estimators, and iv) a comprehensive set of predictor variables. Compared to the existing literature, we go much deeper. Our first contribution is that we significantly expand the scope and rigor in each of these four dimensions. More importantly, our second contribution is to investigate when and how machine learning estimators add value. To the best of our knowledge, we are the first to comprehensively address the “black box” issue in stock market beta estimation by examining the properties and operating scheme of machine learning techniques.

We compare the predictive performance of machine learning-based beta estimators (linear regression, tree-based models, and neural networks) with that of established benchmarks (rolling-window approaches as well as shrinkage-based, portfolio-based, and long-memory forecasting models). We consider a comprehensive set of 81 predictors, including sample beta estimates, predictors based on accounting information, technical indicators, macroeconomic indicators, and the industry classification of Fama and French (1997).

Our first main result is that machine learning techniques outperform established approaches both statistically and economically. Random forests perform especially well. They produce the lowest average value-weighted mean squared error (MSE), closely followed by gradient-boosted regression trees (GBRT) and neural networks. These three approaches yield considerably lower average forecast errors than any of the established benchmarks and are included in the vast majority of cases in the model confidence set (MCS) of Hansen, Lunde, and Nason (2011). The corresponding fractions for all benchmark models being in the MCS are substantially smaller. Moreover, the forecast errors for all benchmark approaches are higher for most of the sample period, as indicated by positive and statistically significant Diebold and Mariano (DM) (1995) test statistics. For example, compared to 1-year rolling betas computed from daily returns, the most natural

benchmark model, the average MSE for random forests is 20% lower. They significantly outperform this benchmark approach in almost 70% of all sample months and are more than twice as often in the MCS. Machine learning-based estimators also outperform the benchmarks when evaluated on the basis of ex post hedging errors.

We then examine the differences in forecast errors across beta estimators to identify time periods and types of stocks for which these differences across such estimators are particularly pronounced. The machine learning-based approaches outperform the benchmark models even more in distressed economic environments (during or immediately following most recessions), that is, periods when it is particularly difficult to accurately predict market betas. In contrast to established beta estimators, random forests and other machine learning methods produce less extreme and less volatile forecasts. Such properties avoid systematically underestimating the betas of stocks in low-beta deciles and systematically overestimating those in high-beta deciles, a central problem inherent in the task of forecasting time-varying market betas. We also find that machine learning-based approaches are superior for almost all types of stocks (sorted into portfolios based on firm characteristics or industry) but are particularly beneficial for small stocks, illiquid stocks, value stocks, and loser stocks. Including relevant firm fundamentals as predictors in our forecasting models helps to generate better forecasts for these stocks.

In addition to a statistical comparison, we analyze the economic value of beta forecasts in portfolio construction exercises. We find that machine learning methods again outperform established approaches. In contrast to the benchmark models, they are able to generate anomaly portfolios that are ex post market neutral. Furthermore, the machine learning methods generate market-neutral momentum, idiosyncratic volatility, and betting-against-beta strategies with higher alphas than the benchmark models. Finally, they produce better minimum variance portfolios (MVPs) based on a single-factor parameterization of the covariance matrix.

In a penultimate step, we examine changes in the inherent model complexity over time and decompose predictions into the contributions of individual variables. We find that the degree of model complexity is positively correlated with the overall difficulty of predicting market betas: More complex models are required when market betas are more difficult to predict. Finally, we find that historical betas and technical indicators are the first and second most important groups of predictor variables, respectively. However, the importance of variables varies over time, and unconditionally less informative variables occasionally play important roles.

Our results underscore the systematic relationship between market betas and firm characteristics. An important reason for the outperformance of machine learning methods is their ability to capture the information content of a large set of firm characteristics that appear to affect betas. Importantly, however, random forests, GBRT, and neural networks also outperform linear regressions that include the same set of covariates. We show that much of this outperformance is due to their ability to exploit nonlinear and interactive patterns.

We perform extensive additional tests to demonstrate the robustness of our results (reported in the Supplementary Material). For example, we show that our results extend naturally to the prediction of other factor betas. In particular, for the

Fama and French (1993) size and value betas, the machine learning methods also produce significantly better forecasts than the benchmark models.

Machine learning methods must be properly trained and tuned to avoid overfitting. There are two types of overfitting: model overfitting and backtest overfitting. The former refers to machine learning models with excessively high in-sample fit but poor out-of-sample predictive performance. To avoid model overfitting, we need to control the degree of model complexity by tuning the relevant hyperparameters. These parameters should be determined adaptively from the sample data. Backtest overfitting refers to a researcher's arbitrary choice of firm coverage, sample period, predictors, and tuning parameters. If information from the out-of-sample period is consciously or unconsciously used to fit the models (Schorfheide and Wolpin (2012)), this can lead to exaggerated out-of-sample predictive performance (Bailey, Borwein, de Prado, and Zhu (2014), (2017), Harvey and Liu (2014), (2015), and Harvey et al. (2016)). To avoid backtest overfitting, we use the largest possible firm coverage and sample period.¹ Motivated by prior literature, we use a comprehensive set of 81 predictor variables rather than focusing only on those covariates that have been shown to perform best in similar predictive tasks. Finally, consistent with Gu, Kelly, and Xiu (2020), we choose the time series cross-validation approach to fit the machine learning models. We follow common parameter choices to cover a representative range of possible specifications from which the hyperparameters are selected. This approach helps to mitigate the risk of backtest overfitting.

The remainder of this article is organized as follows: [Section II](#) briefly reviews the related literature on beta estimation and machine learning. [Section III](#) describes our data set. [Section IV](#) summarizes the different forecasting models (including the machine learning methods and tuning parameters). The empirical results are presented in [Sections V, VI, and VII](#). In particular, [Sections V and VI](#) show the results of the statistical and economic forecast evaluations, respectively. [Section VII](#) analyzes the properties and operating scheme of the machine learning techniques. [Section VIII](#) concludes the article. The Supplementary Material provides details on the estimation of the benchmark and machine learning models in Sections A and B, together with extensive additional analyses and a battery of robustness checks in Section C.

II. Literature Review

Because the original CAPM is a static, 1-period model, its most natural application is based on the premise that stock market betas are constant over time. However, several studies find evidence of time variation in these betas (Bollerslev, Engle, and Woolridge (1988), Jagannathan and Wang (1996), Ferson and Harvey (1999), Petkova and Zhang (2005), and Ang and Chen (2007)). Therefore, Jagannathan and Wang (1996) propose a conditional version of the CAPM and show that

¹According to Gu et al. (2020), using large data sets mitigates sample selection or data snooping biases (Lo and MacKinlay (1990)) and also helps to avoid model overfitting by increasing the ratio of observation count to parameter count.

it explains the cross-sectional variation in expected returns better than its static counterpart.

To estimate time-varying market betas, traditional approaches focus on historical return information. Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973) use the coefficient estimates from ordinary least squares (OLS) time series regressions of stock-level excess returns on market portfolio excess returns. Their 5-year rolling window of monthly returns accounts for the time variation in beta estimates. Despite being robust to misspecification (no predictors needed), rolling-beta estimates face a bias–variance trade-off with respect to window length and data frequency. In addition, such time series estimators are sensitive to outliers in the return history and often produce extreme and volatile beta forecasts. The literature offers modifications to the basic rolling-window approach to improve this trade-off. For example, Hollstein, Prokopczuk, and Wese Simen (2019) show that a weighted least squares regression approach with exponential weights performs well, while Welch (2022) suggests winsorizing stock-level returns before running the time series OLS regressions. Both studies find significantly reduced forecast errors compared to the baseline rolling-window approach.

Enhancing rolling betas with additional cross-sectional information can also improve beta forecasts. The idea is that a stock's beta estimate should not be too different from that of other stocks with similar characteristics. Vasicek (1973) and Karolyi (1992) find that shrinking rolling-beta estimates toward a prior regarding the true beta reduces estimation noise. In contrast, Cosemans, Frehen, Schotman, and Bauer (2016) argue that shrinkage based on common priors only attenuates part of the noise in rolling-beta estimates. They suggest specifying priors unique to each firm based on a broad set of firm fundamentals as predictors. Kim, Korajczyk, and Neuhierl (2020) and Kelly, Moskowitz, and Pruitt (2021) emphasize that commonly used firm fundamentals (such as size or the book-to-market ratio) can help improve the prediction of time-varying market betas. Other approaches include assigning portfolio beta estimates to individual stocks (Fama and French (1992)) and exploiting the long-memory properties of beta time series (Becker, Hollstein, Prokopczuk, and Sibbertsen (2021)).

Studies using machine learning-based approaches are now abundant in the empirical asset pricing literature. While most of them focus on the predictability of return characteristics, there has been little research on the predictability of risk characteristics.² For example, Christensen, Siggaard, and Veliyev (2023) compare various machine learning algorithms in forecasting stock-level expected volatility and find significant outperformance relative to the well-established heterogeneous autoregressive approach.³ These studies focus almost exclusively on the *total* risk

²Several studies apply machine learning to predict expected stock returns (e.g., Gu et al. (2020), Drobetz and Otto (2021), and Leipold, Wang, and Zhou (2021)), bond risk premia (e.g., Bianchi, Büchner, and Tamoni (2021), Bali, Goyal, Huang, Jiang, and Wen (2022)), and earnings expectations (e.g., van Binsbergen, Han, and Lopez-Lira (2023)), among others.

³Other studies, which also apply machine learning techniques to predict future volatility, each focus on a specific method. For example, Mitnik, Robinzonov, and Spindler (2015) and Luong and Doku-chayev (2018) consider tree-based models, while Donaldson and Kamstra (1997), Hillebrand and Medeiros (2010), Fernandes, Medeiros, and Scharth (2014), Bucci (2020), and Rahimikia and Poon (2020) explore neural networks.

of a stock. However, an estimate of its *systematic* risk, that is, its CAPM beta, is at least as important to companies and investors.

For most studies in the literature, the ultimate goal is to model expected stock returns. As an intermittent tool, several studies use firm characteristics to capture the time variation in multifactor betas (e.g., Connor and Linton (2007), Connor, Hagmann, and Linton (2012), Fan, Liao, and Wang (2016), and Kelly, Pruitt, and Su (2019)). Kozak, Nagel, and Santosh (2020) and Gu, Kelly, and Xiu (2021) use machine learning settings for this task. We contribute to this literature by focusing explicitly and directly on the estimation of CAPM market betas using machine learning methods. Our analysis is important for two main applications: i) equity cost of capital estimation, for which it is industry practice to use estimated market betas rather than expected returns, and ii) portfolio risk management, which requires direct knowledge of stock-level systematic risk characteristics (i.e., stock market betas).

The study most closely related to ours is Jourovski, Dubikovskyy, Adell, Ramakrishnan, and Kosowski (2020). The authors use estimates from linear regressions and tree-based models to predict realized betas. To do so, they analyze the MSCI U.S. stock universe (on average 540 mostly large-cap stocks) over the sample period from Jan. 1999 to Dec. 2019. They show that regression trees generally outperform rolling-window estimation and linear regression both statistically and economically. However, they do not compare the machine learning methods with the best benchmark models documented in the recent literature. Furthermore, the authors only examine the economic value of machine learning methods for betting-against-beta portfolios. While they also analyze the importance of each predictor variable, the authors do not examine changes in the inherent model complexity over time or explore patterns of nonlinear and interactive effects in the relationship between predictor variables and beta estimates.

Therefore, our empirical analysis goes much deeper: We i) comprehensively compare the performance of machine learning estimators (including neural networks) with the best benchmark models documented in the recent literature, ii) analyze *when* and *how* machine learning techniques outperform by comparing the time series of forecast errors and by analyzing the forecast errors of cross-sectional portfolio sorts, iii) document the economic value for MVPs and a large set of anomaly portfolios, iv) analyze both model complexity and variable importance as well as nonlinear and interactive effects, and v) examine a much larger and longer sample along with a much more comprehensive set of predictors.⁴

III. Data

Our market and fundamental data come from CRSP and Compustat, respectively, and consist of daily and monthly returns and various firm characteristics. They are aggregated on a monthly basis and denominated in U.S. dollars when

⁴A comparison with the best benchmark models is important because, as described in the text, several methods exist that outperform simple rolling-window estimators. Without further analysis, it remains unclear whether machine learning techniques also outperform more sophisticated, and thus more conservative, benchmark models and indeed provide the best beta estimates.

currency-related.⁵ Our sample is free of survivorship bias and includes all firms that were or are listed on the New York Stock Exchange (NYSE), the American Stock Exchange (AMEX), or the National Association of Securities Dealers Automated Quotations (NASDAQ). To calculate excess returns, we use the 3-month U.S. T-bill rate, scaled to the daily or monthly horizon, as the risk-free rate. The value-weighted portfolio of all stocks serves as a proxy for the market portfolio.

In Table 1, we present our comprehensive set of 81 predictors in detail. It is an extension of the set used by Cosemans et al. (2016), which includes five fundamental covariates (size, book-to-market ratio, financial leverage, operating leverage, and momentum), one macroeconomic covariate (default spread), and 47 dummies that correspond to the industry classification of Fama and French (1997).⁶ We augment this base set with 28 variables that have been shown to explain the cross-sectional variation in future market betas (Beaver, Kettler, and Scholes (1970), Amihud and Mendelson (2000), Jacoby, Fowler, and Gottesman (2000), Chincarini, Kim, and Moneta (2020), and Kelly et al. (2021)).⁷ In particular, we include 25 additional fundamental covariates (e.g., age, illiquidity, or turnover), which we classify into 18 predictors based on accounting information and 7 technical indicators. To capture the time series dynamics in betas, we further include three predictors based on sample estimates of beta obtained from rolling regressions. We use 3-month and 1-year historical windows of daily returns (OLS_1Y_D and OLS_3M_D, respectively) as well as a 5-year historical window of monthly returns (OLS_5Y_M) to obtain information about short-, medium-, and long-term trends in the beta time series. The inclusion of historical betas based on three different horizons allows for a heterogeneous autoregressive forecast structure. As documented by Becker et al. (2021), this helps to capture the long-memory properties of market beta time series.

For many firms, some of the firm characteristics are missing. In these cases, the entire firm-month observations would have to be omitted because the econometric models require data sets without missing data. To avoid losing these data, we use the approach of Freyberger, Höppner, Neuhierl, and Weber (2024) to impute missing accounting-based firm characteristics and technical indicators. Specifically, we first always take 60 months of data jointly. We impute the first 60 months together and then use the current month and the 59 previous months to estimate the parameters. Second, we identify all missing value patterns (combinations of missing and non-missing variables) in the data. Third, for each of these patterns, we use the largest

⁵Market data are assumed to become public immediately, and fundamental data are assumed to be published 4 months after the end of the fiscal year.

⁶Cosemans et al. (2016) follow Gulen, Xing, and Zhang (2011) in measuring a firm's operating leverage as the ratio of change in operating income before depreciation to change in net sales. We opt for the Novy-Marx (2011) definition, which is another well-established measurement approach in the literature. This choice increases consistency across predictors, especially with respect to financial leverage. The main results of our empirical analysis are qualitatively similar for other operating leverage definitions.

⁷From the extensive list of predictors that significantly predict future market betas in the Kelly et al. (2021) study, we omit only the bid-ask spread because the data are largely unavailable until the mid-1980s. The main findings of our empirical analysis are qualitatively similar when including the bid-ask spread.

TABLE 1
Variable Descriptions and Definitions

Table 1 presents the descriptions and definitions of the 81 predictors used in the empirical analysis. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020. Data from CRSP and Compustat are aggregated on a monthly basis and denominated in U.S. dollars when currency-related. Market data are assumed to be immediately available, and fundamental data are assumed to be available 4 months after the end of the fiscal year.

| # | Predictor | Description | Definition |
|---|---------------|--------------------------------|---|
| <i>Predictors Based on Sample Estimates of Beta</i> | | | |
| 1 | OLS_3M_D | Short-term beta | Sample beta estimate from rolling regressions using a 3-month window of daily returns |
| 2 | OLS_1Y_D | Medium-term beta | Sample beta estimate from rolling regressions using a 1-year window of daily returns |
| 3 | OLS_5Y_M | Long-term beta | Sample beta estimate from rolling regressions using a 5-year window of monthly returns |
| <i>Predictors Based on Accounting Information</i> | | | |
| 4 | AGE | Age | Log number of years since first inclusion in CRSP |
| 5 | AT | Total assets | Log book value of total assets |
| 6 | BM | Book-to-market ratio | Log ratio of book and market value of equity |
| 7 | CAPTURN | Capital turnover | Log ratio of net sales to lagged book value of total assets |
| 8 | DIVPAY | Dividend payout ratio | Ratio of dividends paid during the last fiscal year to net income |
| 9 | EP_COVAR | Covariability in earnings | Coefficient estimate in the time series ordinary least squares regression of monthly earnings-to-price ratios on the market's monthly earnings-to-price ratio over the last 3 years |
| 10 | EP_VAR | Variability in earnings | Log standard deviation of monthly earnings-to-price ratios over the last 3 years |
| 11 | FINLEV | Financial leverage | Log ratio of book value of total assets to market value of equity |
| 12 | FXDCOS | Fixed cost of sales | Log ratio of selling, general, and administrative expenses plus research and development expenses plus advertising expenses to net sales |
| 13 | INVEST | Investment | Year-on-year growth of book value of total assets |
| 14 | NOA | Net operating assets | Ratio of operating assets minus operating liabilities to book value of total assets |
| 15 | OPACCR | Operating accruals | Ratio of changes in noncash working capital minus depreciation to book value of total assets |
| 16 | OPLEV | Operating leverage | Log ratio of operating costs (i.e., the sum of costs of goods sold and selling, general, and administrative expense) to market value of total assets |
| 17 | PPE | PPE change-to-assets ratio | Ratio of changes in property, plants, and equipment (PPE) to lagged book value of total assets |
| 18 | PROF | Profitability | Ratio of gross profits to book value of equity |
| 19 | ROA | Return on assets | Ratio of income before extraordinary items to book value of total assets |
| 20 | ROE | Return on equity | Ratio of income before extraordinary items to book value of equity |
| 21 | RON | Return on net operating assets | Ratio of operating income after depreciation to lagged net operating assets |
| 22 | SALESTOASSETS | Sales-to-assets ratio | Log ratio of net sales to book value of total assets |
| 23 | SALESTOPRICE | Sales-to-price ratio | Log ratio of net sales to market value of equity |
| 24 | SGATOSALES | SGA-to-sales ratio | Log ratio of selling, general, and administrative (SGA) expenses to net sales |
| <i>Technical Indicators</i> | | | |
| 25 | ILLIQ | Illiquidity | Ratio of monthly absolute return to monthly dollar trading volume |
| 26 | INTERMOM | Intermediate momentum | Excess return from month -12 to month -7 |
| 27 | IVOL | Idiosyncratic volatility | Log standard deviation of daily residuals from Fama and French's (1992) 3-factor model within the current month |
| 28 | LTREV | Long-term reversal | Excess return from month -36 to month -13 |
| 29 | ME | Size | Log market value of equity |
| 30 | MOM | Momentum | Excess return from month -12 to month -2 |
| 31 | RELPRC | Relative price | Ratio of end-of-month price to its highest daily price during the last year |
| 32 | STREV | Short-term reversal | Excess return from the current month |
| 33 | TO | Turnover | Log monthly dollar trading volume |
| <i>Macroeconomic Indicators</i> | | | |
| 34 | DFY | Default spread | Yield differential between Moody's Baa- and Aaa-rated corporate bonds |
| <i>Industry Classifiers</i> | | | |
| 35–81 | IND | Industry classification | Fama and French's (1997) industry classification, resulting in $48 - 1 = 47$ industry dummies |

possible sample of complete cases, that is, firm-month observations with no missing data, and regress each missing characteristic on all others that are available for that pattern. This procedure gives us the best conditional expectation for the missing variables as a function of the nonmissing variables. Finally, we use the nonmissing characteristics of each stock along with the coefficient estimates from this regression to impute the missing values.

We follow Cosemans et al. (2016) in cleaning the data set. We include a stock in the empirical analysis for month t only if it meets the following criteria: First, its book value of equity (according to Fama and French (1992)) must be nonnegative, and both its net sales and monthly dollar trading volume must be positive. Second, its return in the current month t and over the previous 36 months must be available. Third, after the imputation procedure, it must provide complete information on historical and realized betas as well as all predictor variables used in the empirical analysis. These requirements limit our sample period to Mar. 1970 to Dec. 2020, for which we have an average of 1,806 stocks per month.

Following Cosemans et al. (2016), we winsorize outliers in all firm characteristics to the 0.5th and 99.5th percentile values of their cross-sectional distributions. In addition, as in Kelly et al. (2019) and Freyberger, Neuhierl, and Weber (2020), we cross-sectionally rank all firm characteristics each month and then map the ranks to the $(-1, +1)$ interval.

An important caveat is that many of the predictors are constructed similarly, such as sample beta estimates based on different rolling windows, or contain similar information, such as valuation ratios measured relative to the market value of the stock, resulting in relatively high correlations. However, as discussed in the study by Lewellen (2015), any resulting multicollinearity is not a major concern because we are primarily interested in the overall predictive power of the machine learning-based forecasting models rather than the marginal effects of each individual predictor. Moreover, most of the machine learning techniques we use are suitable for solving the multicollinearity problem, either by their nature (tree-based models) or by applying different types of regularization, such as lasso-based penalization of the weights (neural networks).

IV. Forecast Models

The main objective of our empirical analysis is to investigate whether machine learning techniques outperform established beta estimators in terms of predictive performance, and if so, why. In particular, we are interested in whether incorporating nonlinearity and interactions in the relationship between predictors and future market betas adds incremental predictive power. Therefore, we run a horse race between traditional and machine learning-based beta estimators.

Following Cosemans et al. (2016) and Hollstein and Prokopczuk (2016), we estimate and evaluate the forecasts at the individual stock level. The setup in our analysis is as follows: Out-of-sample beta estimates are obtained at the firm level and on a monthly basis following an iterative procedure. In the first iteration step, we use data up to the end of month t and obtain forecasts for the beta of each stock i

during the out-of-sample forecast period (from the beginning of month $t + 1$ to the end of month $t + k$): $\beta_{i,t+k|t}^F$ (or abbreviated $\beta_{i,t}^F$). We set k equal to 12 and focus on a 1-year forecast horizon.⁸ In the next iteration step, we use data up to the end of month $t + 1$ and obtain forecasts of stock-level betas during the subsequent out-of-sample forecast period (from the beginning of month $t + 1 + 1$ to the end of month $t + 1 + k$). By iterating through the data set, we obtain time series of overlapping out-of-sample beta estimates, which we then compare to realized betas. Andersen, Bollerslev, Diebold, and Wu (2006) document that a realized beta measure constructed from high-frequency returns is a consistent estimator of the true integrated beta. Therefore, we measure future realized betas using daily returns over the 1-year forecast horizon as $\beta_{i,t+k}^R = \frac{\text{Cov}_{iM,t+k}^R}{\text{Var}_{M,t+k}^R}$, where $\text{Cov}_{iM,t+k}^R$ is the realized covariance between stock i and the market portfolio M , and $\text{Var}_{M,t+k}^R$ is the realized market variance. Both moments are computed from continuously compounded returns.

For the sake of brevity, we introduce the models used to estimate time-varying market betas in Section A of the Supplementary Material (see Table A1 in the Supplementary Material for an overview). These models differ in their overall approach and complexity, but they all aim to minimize the forecast error, which we compute as the *value-weighted* MSE at the end of each month t :

$$(1) \quad \text{MSE}_{t+k|t} = \sum_{i=1}^{N_t} w_{i,t} \left(\beta_{i,t+k}^R - \beta_{i,t+k|t}^F \right)^2 \text{ with } k = 12,$$

where N_t is the number of stocks in the sample at the end of month t , and $w_{i,t}$ is the market capitalization-based weight of stock i . It is important to note that realized betas are themselves estimates. However, evaluating forecasts based on future realized betas is an approach that works well statistically (see, e.g., Hansen and Lunde (2006), for the theoretical framework and empirical evidence). Moreover, in the context of volatility forecasting, Patton (2011) shows that the MSE criterion is robust to mean-zero noise in the evaluation proxy.⁹

A. Benchmark Estimators

From the extensive literature on beta estimation, we select a representative set of established forecasting models, which we classify into four model families based on methodology (see Section A of the Supplementary Material for details, implementation choices, and references). The first model family consists of rolling-window estimators, for which we consider two basic historical betas obtained from

⁸Alternatively, 1-month and 5-year forecast horizons (with $k = 1$ and $k = 60$, respectively) are also common in the literature. However, both alternatives have shortcomings. First, realized betas computed from 1-month rolling windows of daily returns are very noisy, making it difficult to evaluate forecast errors. Second, forecast horizons much longer than 12 months are less common in the industry due to the underlying nature of fiscal years. We demonstrate the robustness of our results to different forecast horizons in Section C of the Supplementary Material (see Table C7 in the Supplementary Material).

⁹In Section C of the Supplementary Material, we examine the robustness of our results to changes in the forecast error measure. In particular, the results for an equal-weighted mean squared error (see Table C4 in the Supplementary Material) and a value-weighted mean absolute error (see Table C5 in the Supplementary Material) are qualitatively similar to our baseline results in Section V.

rolling regressions using a 5-year window of monthly returns (OLS_5Y_M) and a 1-year window of daily returns (OLS_1Y_D) as well as two common modifications, exponentially weighted betas based on short (EWMA_S) and long (EWMA_L) half-lives and slope-winsorized betas (BSW). The second model family consists of shrinkage-based estimators, for which we include three shrinkage betas that shrink OLS_1Y_D toward the average beta within the stock universe (VASICEK), an industry portfolio (KAROLYI), and a firm-specific beta prior (HYBRID). The third and fourth model families are portfolio-based and long-memory estimators, respectively, for which we include portfolio betas assigned to individual stocks (FAMA–FRENCH) and long-memory betas that exploit the long-memory properties of beta time series (LONG-MEMO).

B. Machine Learning Estimators

The machine learning-based approaches follow a different, more rigid path to capture the cross-sectional variation in future betas. For example, shrinkage-based estimators derive prior beliefs and sample estimates of beta separately before aggregating these two sources of information into shrinkage betas. Rather than taking this “detour,” machine learning techniques focus directly on the goal of predicting market betas. Realized betas enter directly into the regressive framework as dependent variables, while sample estimates of beta, firm characteristics, etc., serve as predictors. This approach keeps the forecasting objective in mind when training the model and uses multiple sources of information, potentially leading to incremental predictive power. We adapt the additive prediction error model outlined in the study by Gu et al. (2020) to describe a stock’s beta:

$$(2) \quad \beta_{i,t+k}^R = E_t(\beta_{i,t+k}^R) + \varepsilon_{i,t+k},$$

where $\beta_{i,t+k}^R$ is the realized beta of stock i over the 1-year forecast horizon starting at the beginning of month $t + 1$, and $\varepsilon_{i,t+k}$ is an error term. The expected beta, $E_t(\beta_{i,t+k}^R)$, is estimated as a function of the predictor variables and is described by the “true” model $g^*(z_{i,t})$, where $z_{i,t}$ represents the P -dimensional set of predictors:

$$(3) \quad E_t(\beta_{i,t+k}^R) = g^*(z_{i,t}).$$

Although our machine learning-based prediction models belong to different families (linear regression, tree-based models, and neural networks), they are all designed to approximate the true prediction model by minimizing the out-of-sample MSE. Approximations of the conditional expectation $g^*(z_{i,t})$ are flexible and family-specific. The approximation function $g(\cdot)$ can be linear or nonlinear, as well as parametric based on $g(z_{i,t}, \theta)$, where θ is the set of true parameters, or nonparametric, with $g(z_{i,t})$.

1. Sample Splitting

Machine learning methods are designed to i) simultaneously incorporate a large number of variables and ii) account for both nonlinearity and interactions in the relationship between predictor variables and beta estimates. However, they are

prone to overfitting, and thus we need to control model complexity by tuning the relevant hyperparameters (e.g., the number and/or depth of trees in tree-based models or the number of layers and/or nodes in neural networks). The hyperparameters should be determined adaptively from the sample data and selected from an extensive set of parameter specifications (see Panel B of Table A1 in Section A of the Supplementary Material for more details). The parameter tuning approach iteratively reduces the in-sample fit by searching for a level of model complexity that produces reliable out-of-sample predictive performance. To this end, following Gu et al. (2020), we apply the time series cross-validation approach, which preserves the temporal order of the data and divides the sample into three distinct subsamples: a training sample, a validation sample, and a test sample.

We use the training sample to estimate the model for multiple parameter specifications, while we use the validation sample to tune the parameters. That is, based on the models estimated from the training sample, we compute the time series mean of the monthly *value-weighted* MSEs within the validation sample for each parameter specification. The model with the parameter specification that minimizes the validation error is used for out-of-sample testing. The test sample is not used for either model estimation or parameter tuning. Therefore, it is truly out-of-sample and appropriate for evaluating the out-of-sample predictive power of a model.

In portfolio management applications, where new data emerge over time, some sample splitting scheme must be applied that periodically incorporates more recent data (see, e.g., West (2006), for an overview). We follow Gu et al. (2020) and refit the models once a year. We use a rolling-window approach. Each year, we roll forward the training and validation samples by 1 year, keeping the length of each sample constant. We always select 10 years of data for training and validation, that is, 9 years for training and 1 year for validation, and 1 year for testing. Starting in Dec. 1979, we obtain the last beta estimates in Dec. 2019 using 10 years of data for training and validation (Jan. 2009 to Dec. 2017 and Jan. 2018 to Dec. 2018, respectively), which we compare to the realized betas over the following year.¹⁰ In total, we use 40 years and 1 month of data for testing.

2. Machine Learning Techniques

In our empirical analysis, we analyze three different families of predictive models that differ in their overall approach and complexity (see Section B of the Supplementary Material for details, implementation choices, and references). The first family of models consists of *linear regressions*, where we use the training sample to run pooled OLS regressions of future realized betas $\beta_{i,t+k}^R$ on the set of 81 predictors. Specifically, we use either the OLS loss function (LM) or an elastic net penalization (ELANET). The latter is the most common machine learning technique to overcome the overfitting problem in a high-dimensional regression setting. Unless explicitly included as predetermined terms, pooled regressions cannot capture nonlinear or interactive effects. We use linear regressions as a

¹⁰Because we focus on a 1-year forecasting horizon, there is a 1-year gap between the end of the sample used for training and validation (Dec. 2018) and the estimation date (Dec. 2019).

benchmark to determine whether such effects, in addition to the interaction between firm characteristics and the default spread, lead to additional predictive power.

The second model family consists of tree-based models, for which we include random forests and GBRT, the most common representatives within this subcategory. The third family of models are neural networks (NN_1–NN_5), for which we consider specifications with up to 5 hidden layers and 32 neurons.¹¹ Both tree-based models and neural networks inherently account for nonlinearity and multiway interactions without the need to add new predictors that capture these effects in advance.

V. Statistical Analysis of Market Beta Forecasts

A. Forecast Errors

In the first step of our empirical analysis, we evaluate the models' abilities to predict out-of-sample market betas.¹² In particular, we run a horse race between established and machine learning-based beta estimators, comparing their predictive performance from a statistical perspective. Panel A of [Table 2](#) reports the time series averages of the monthly value-weighted MSEs (based on a 1-year forecast horizon), calculated as in [equation \(1\)](#). Ignoring any cross-sectional information, the beta estimates obtained from rolling regressions lead to substantial average forecast errors, ranging from 19.17% for the OLS_5Y_M model to 9.44% for the EWMA_L model. Winsorizing or shrinking the rolling-beta estimates toward a well-defined prior, assigning portfolio beta estimates to individual stocks, or exploiting the long-memory properties of beta time series substantially reduce the average MSE. Consistent with Cosemans et al. (2016), Becker et al. (2021), and Welch (2022), the best performing estimators among our benchmark approaches are slope-winsorized betas (8.77%), hybrid betas (8.53%), and long-memory betas (8.29%).

Turning to the machine learning methods, we observe that tree-based models and neural networks further reduce the average prediction error relative to the best benchmark beta estimators (average MSEs between 7.77% and 8.04%). In contrast, linear regressions (both simple, 9.15%, and penalized, 8.89%) have notably higher average MSEs.¹³ Therefore, using information from a large set of predictors in isolation is not sufficient to produce superior beta estimates, and much of the outperformance of the RF, GBRT, and NN_1 models is due to their ability to

¹¹ Additional robustness tests in Section C of the Supplementary Material document that the predictive performance for the neural network models deteriorates slightly with the number of hidden layers. Therefore, in the main part of this article, we only present and discuss the results for the simplest NN_1 architecture.

¹² In Section C of the Supplementary Material, we additionally analyze the time series and cross-sectional properties of the benchmark and machine learning-based beta estimators. We find that the machine learning-based estimators have the lowest standard deviations and produce the least extreme beta estimates (see Table C1 in the Supplementary Material).

¹³ Our finding that tree-based models perform particularly well in estimating market beta is consistent with Jorovski et al. (2020), although they do not analyze any of the top three performing benchmark models or neural networks.

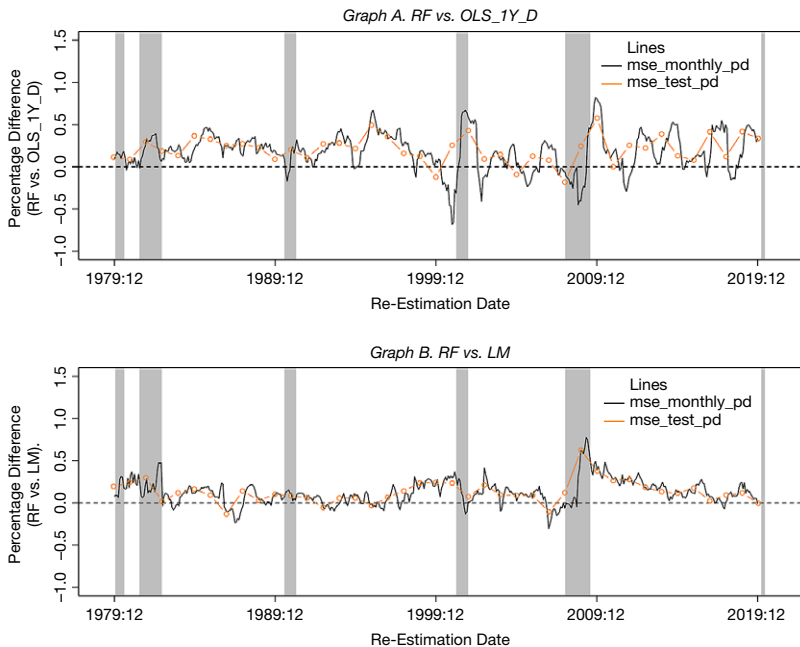
TABLE 2
Forecast Errors (Value-Weighted MSEs)

Table 2 shows the differences in forecast errors obtained from the forecasting models presented in Section IV. Panel A reports the time series averages of the monthly value-weighted MSEs: $MSE_{t+k|t}^{(j)} = \sum_{i=1}^{N_t} w_{i,t} \left(\beta_{i,t+k}^R - \beta_{i,t+k|t}^{F,(j)} \right)^2$, with $k = 12$, where N_t is the number of stocks in the sample at the end of month t , and $w_{i,t}$ is the market capitalization-based weight of stock i . Panel B reports the fraction of months during the out-of-sample period for which the column model is i) in the Hansen et al. (2011) model confidence set (MCS) and ii) significantly better than the row model in a pairwise comparison (according to the Diebold and Mariano (1995) test statistics). The DM tests of equal predictive ability examine the differences in stock-level squared forecast errors (SEs): $SE_{i,t+k|t}^{(j)} = \left(\beta_{i,t+k}^R - \beta_{i,t+k|t}^{F,(j)} \right)^2$, with $k = 12$. The DM test statistic in month t for comparing model j with a competing model i is $DM_t^{(j,i)} = \frac{\bar{d}_t^{(j,i)}}{\hat{\sigma}_{\bar{d}_t^{(j,i)}}}$, where $d_{i,t}^{(j,i)} = SE_{i,t+k|t}^{(i)} - SE_{i,t+k|t}^{(j)}$ is the difference in SEs, $\bar{d}_t^{(j,i)} = \sum_{i=1}^{N_t} w_{i,t} d_{i,t}^{(j,i)}$ the value-weighted cross-sectional average of these differences, and $\hat{\sigma}_{\bar{d}_t^{(j,i)}}$ the Newey and West (1987) standard error of $\bar{d}_t^{(j,i)}$ (with 4 lags to account for possible heteroscedasticity and autocorrelation). Statistical tests are based on the 10% significance level. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020. The first beta estimates are obtained in Dec. 1979.

| | Benchmark Estimators | | | | | | | | | | ML Estimators | | | | |
|---|----------------------|----------|--------|--------|-------|---------|---------|--------|-------------|-----------|---------------|--------|-------|-------|-------|
| | OLS_5Y_M | OLS_1Y_D | EWMA_S | EWMA_L | BSW | VASICEK | KAROLYI | HYBRID | FAMA-FRENCH | LONG-MEMO | LM | ELANET | RF | GBRT | NN_1 |
| <i>Panel A. Average Forecast Errors</i> | | | | | | | | | | | | | | | |
| MSE, value-weighted (%) | 19.17 | 9.70 | 9.55 | 9.44 | 8.77 | 8.91 | 8.97 | 8.53 | 9.11 | 8.29 | 9.15 | 8.89 | 7.77 | 8.04 | 7.79 |
| <i>Panel B. Forecast Errors over Time</i> | | | | | | | | | | | | | | | |
| In MCS | 3.74 | 32.64 | 40.54 | 41.79 | 49.69 | 46.99 | 50.10 | 52.81 | 43.04 | 67.15 | 51.77 | 57.80 | 82.54 | 68.19 | 78.59 |
| <i>Benchmark Estimators</i> | | | | | | | | | | | | | | | |
| vs. OLS_5Y_M | | 87.73 | 87.73 | 88.36 | 92.52 | 91.68 | 90.64 | 94.39 | 90.85 | 97.51 | 90.44 | 89.81 | 96.47 | 96.05 | 96.67 |
| vs. OLS_1Y_D | 1.87 | | 32.22 | 48.44 | 74.22 | 81.50 | 83.58 | 72.35 | 50.31 | 60.29 | 46.78 | 50.73 | 68.61 | 61.75 | 62.99 |
| vs. EWMA_S | 1.25 | 28.27 | | 39.29 | 52.60 | 50.52 | 48.02 | 54.05 | 40.75 | 55.72 | 43.04 | 45.74 | 62.99 | 59.88 | 59.88 |
| vs. EWMA_L | 1.25 | 16.84 | 20.58 | | 55.51 | 52.60 | 50.52 | 59.46 | 39.50 | 54.26 | 43.24 | 46.15 | 62.79 | 58.21 | 59.46 |
| vs. BSW | 0.62 | 5.20 | 13.31 | 12.47 | | 22.45 | 22.25 | 40.96 | 16.01 | 42.00 | 34.93 | 38.88 | 59.88 | 53.43 | 54.47 |
| vs. VASICEK | 1.04 | 6.86 | 14.55 | 13.72 | 29.73 | | 19.75 | 50.52 | 12.68 | 43.04 | 35.97 | 39.50 | 60.50 | 56.55 | 56.13 |
| vs. KAROLYI | 1.04 | 5.82 | 15.59 | 12.89 | 29.94 | 30.15 | | 49.69 | 18.50 | 43.04 | 34.93 | 40.54 | 57.80 | 52.39 | 54.05 |
| vs. HYBRID | 0.83 | 3.53 | 12.89 | 13.51 | 21.21 | 23.08 | 21.83 | | 14.97 | 31.60 | 31.39 | 33.89 | 51.77 | 48.23 | 50.94 |
| vs. FAMA-FRENCH | 1.46 | 14.14 | 18.09 | 19.33 | 31.60 | 30.15 | 30.15 | 44.28 | | 46.36 | 33.89 | 39.09 | 63.83 | 60.71 | 60.29 |
| vs. LONG-MEMO | 0.00 | 14.97 | 16.22 | 17.46 | 19.33 | 18.71 | 20.17 | 22.04 | 16.22 | | 21.62 | 27.23 | 43.04 | 40.33 | 43.04 |
| <i>ML Estimators</i> | | | | | | | | | | | | | | | |
| vs. LM | 3.12 | 27.65 | 28.27 | 30.98 | 34.72 | 33.89 | 34.93 | 37.21 | 28.27 | 40.75 | | 37.63 | 65.28 | 54.47 | 63.83 |
| vs. ELANET | 3.33 | 25.78 | 26.61 | 29.73 | 34.30 | 34.93 | 35.55 | 33.47 | 27.23 | 34.10 | 18.92 | | 56.13 | 49.69 | 54.26 |
| vs. RF | 0.00 | 9.77 | 10.60 | 12.68 | 13.31 | 13.93 | 14.35 | 14.55 | 6.86 | 10.81 | 5.82 | 8.94 | | 18.30 | 20.17 |
| vs. GBRT | 0.42 | 16.22 | 18.71 | 18.71 | 22.04 | 22.87 | 22.66 | 23.70 | 13.31 | 19.54 | 9.77 | 18.71 | 39.92 | | 34.93 |
| vs. NN_1 | 0.00 | 12.89 | 13.10 | 13.72 | 17.67 | 17.88 | 18.09 | 17.88 | 11.64 | 11.02 | 6.65 | 13.93 | 23.70 | 16.22 | |
| t | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 |

FIGURE 1
Relative Forecast Errors over Time

Figure 1 plots the forecast errors for random forests (RF, introduced in Section IV.B) over the sample period relative to those obtained by 1-year rolling betas (OLS_1Y_D, introduced in Section IV.A) in Graph A and betas obtained from simple linear regressions (LM, introduced in Section IV.B) in Graph B. We compute the relative forecast error as the percentage difference between the MSEs of the two models and show both the monthly relative forecast error and the relative average forecast error within each calendar year of our test sample (MSE_MONTHLY_PD and MSE_TEST_PD, respectively). These percentage differences are calculated as 1 minus the MSE of the random forests divided by the MSE of the respective benchmark model. The orange unfilled circles are assigned to the re-estimation dates, that is, the dates when the forecasts of the stock-level betas (over the next year) are obtained. The graphs also visualize the NBER recession periods (gray-shaded areas). The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020, while the first beta estimates are obtained in Dec. 1979.



capture nonlinearity and interactive effects. Random forests perform best, with an average MSE of 7.77%. They reduce the average forecast error relative to the most commonly used estimation techniques, the OLS_5Y_M and OLS_1Y_D models, by 59% and 20%, respectively. Even relative to the best performing benchmark approach, the LONG-MEMO model, random forests reduce the forecast error by more than 6%.

Since, by construction, these numbers only reflect the average predictive performance of a forecasting model, we also examine the prediction errors over time to assess *when* machine learning estimators perform particularly well. First, we visually examine the differences in MSEs between the forecasting models over the sample period. For the sake of brevity, we focus on comparing random forests with 1-year rolling betas and simple linear regressions. Figure 1 shows the forecast errors for random forests over the sample period relative to those achieved by OLS_1Y_D (Graph A) and LM (Graph B). It also contains the recession periods, as defined by the National Bureau of Economics Research (NBER; gray-shaded areas). We

compute the relative forecast error as the percentage difference between the MSEs of the two models and show both the monthly relative forecast error and the relative average forecast error within each calendar year of our test sample (MSE_MONTHLY_PD and MSE_TEST_PD, respectively).¹⁴

The visualizations indicate that random forests reduce the forecast errors relative to the OLS_1Y_D and LM models most of the time over the sample period, suggesting that random forests are generally able to provide more precise stock market beta forecasts. In addition, larger-than-average positive differences during or after most recessions (recognizing that MSEs are computed based on a 1-year forecast horizon) imply that the RF model outperforms the two benchmarks even more strongly in distressed economic environments, when it is particularly difficult to accurately predict market betas.

Second, to statistically assess the differences in prediction errors, Panel B of Table 2 reports the fraction of months in the out-of-sample period for which the column model is i) in the Hansen et al. (2011) MCS and ii) significantly better than the row model in a pairwise comparison (according to the Diebold and Mariano (DM) (1995) test statistics).¹⁵ The MCS approach incorporates an adjustment for multiple testing and includes the best forecast model(s) based on a certain confidence level. The DM tests of equal predictive ability examine the differences in stock-level squared forecast errors (SEs):

$$(4) \quad SE_{i,t+k|t}^{(j)} = \left(\beta_{i,t+k}^R - \beta_{i,t+k|t}^{F,(j)} \right)^2 \text{ with } k = 12.$$

The DM test statistic in month t for comparing model j with a competing model l is $DM_t^{(j,l)} = \frac{\bar{d}_t^{(j,l)}}{\hat{\sigma}_{\bar{d}_t^{(j,l)}}}$, where $d_{i,t}^{(j,l)} = SE_{i,t+k|t}^{(j)} - SE_{i,t+k|t}^{(l)}$ is the difference in SEs, $\bar{d}_t^{(j,l)} = \sum_{i=1}^{N_t} w_{i,t} d_{i,t}^{(j,l)}$ is the value-weighted cross-sectional average of these differences, and $\hat{\sigma}_{\bar{d}_t^{(j,l)}}$ is the heteroscedasticity and autocorrelation consistent standard error of $\bar{d}_t^{(j,l)}$. We use the Newey and West (1987) estimator with 4 lags for calculating these standard errors.

We find that regression trees and neural networks are in the MCS for most of the 481 sample months, with fractions ranging from 68.19% for GBRT to 82.54% for RF. Therefore, we can reject the null hypothesis that random forests are the best model in only about 17% of the sample months, which is only slightly higher than the expected proportion of false positives at the 10% significance level. This

¹⁴These percentage differences are calculated as 1 minus the MSE of the random forests divided by the MSE of the respective benchmark model. We follow the convention that positive differences indicate superior predictive performance, that is, reduced prediction errors, of the random forest relative to the OLS_1Y_D and LM models, respectively.

¹⁵We follow Becker et al. (2021) in testing statistical significance at the 10% level, which corresponds to 90% model confidence sets. As a robustness test, we also use the Giacomini and White (2006) test to assess the relative conditional predictive performance of each model in pairwise comparisons in Section C of the Supplementary Material (see Table C2 in the Supplementary Material). Consistent with the results shown here, we find that the machine learning-based beta estimators outperform all other methods not only in terms of their unconditional predictive ability but also in terms of their conditional predictive ability.

overwhelmingly suggests that the RF model in particular, but also other machine learning techniques, provides very accurate predictions for stock market betas. The machine learning methods are in the MCS more than twice as often as the 1-year rolling betas (32.64%). Therefore, we can reject the null hypothesis that 1-year rolling betas, which are the most commonly used estimators, provide the best beta forecasts in more than 67% of the months. The fractions with which the machine learning estimators are in the MCS are also considerably larger than those of the best performing benchmark approaches (e.g., slope-winsorized betas, 49.69%; hybrid betas, 52.81%; and long-memory betas, 67.15%). Overall, the machine learning-based models are predominantly among the top performers.

The results of the monthly DM tests confirm this observation. Both tree-based models and neural networks dominate most established approaches and linear regressions. In up to 96% of the months, they produce a significantly lower MSE than the respective benchmark models. For almost all benchmark models, this fraction is well above or close to 50%, at least for the RF, GBRT, and NN_1 models. The machine learning estimators even significantly outperform the best performing benchmark model, long-memory betas, for at least 40.33% of the sample months. Conversely, the benchmark models rarely produce significantly lower MSEs than the machine learning-based approaches. Even the best benchmark, the LONG-MEMORY model, significantly outperforms the RF, GBRT, and NN_1 models in less than 20% of the sample months.

Taken together, these results indicate a clear outperformance of regression trees and neural networks over established beta estimators.¹⁶ When comparing the machine learning techniques with each other, the RF model appears to be slightly superior to the GBRT and NN_1 models. Random forests have the largest MCS fraction, and as shown in Panel B of Table 2, they outperform GBRT and neural networks more often than they are dominated by them.

One could argue that machine learning methods have an inherent advantage when evaluated on the basis of MSE. This is because they are explicitly trained to predict the MSE as accurately as possible, while most benchmark models are not. To account for this aspect, we also analyze the predictive performance of the models using an alternative evaluation metric, the mean squared hedging error (MSHE). For each stock, we compute the squared hedging error as the squared difference between the realized return and the return implied by the market model along with the beta forecast: $(R_{i,t+k} - \beta_i^F R_{M,t+k})^2$. $R_{i,t+k}$ is the excess return of stock i , and $R_{M,t+k}$ is the market excess return. Consistent with the previous analysis, we set k to 12 months.

Table 3 shows the results, which are qualitatively similar to those based on the MSE metric. The RF, GBRT, and NN_1 models produce the lowest average MSHEs

¹⁶In Section C of the Supplementary Material, we use Mincer and Zarnowitz (1969) regressions to test the unbiasedness of the different forecasting models. The results are shown in Table C6 in the Supplementary Material. As expected, we find that the best performing machine learning models are the least biased. In an additional robustness test, we extend our beta forecasting approach to the size and value factors of Fama and French (1993). The results in Table C11 in the Supplementary Material provide first and preliminary evidence that machine learning methods are also useful for the prediction of factor betas. In particular, the machine learning-based models produce smaller forecast errors than the benchmarks for both size (SMB) and value (HML) betas.

TABLE 3
Forecast Errors (Value-Weighted MSHEs)

Table 3 shows the differences in forecast errors obtained from the forecasting models presented in Section IV. Panel A reports the time series averages of the monthly value-weighted MSHEs: $MSHE_{i,t+k|t}^{(j)} = \sum_{i=1}^{N_t} w_{i,t} \left(R_{i,t+k} - \beta_{i,t+k|t}^{F,(j)} R_{M,t+k} \right)^2$, with $k = 12$, where N_t is the number of stocks in the sample at the end of month t , and $w_{i,t}$ is the market capitalization-based weight of stock i . Panel B reports the fraction of months during the out-of-sample period for which the column model is i) in the Hansen et al. (2011) model confidence set (MCS) and ii) significantly better than the row model in a pairwise comparison (according to the Diebold and Mariano (1995) test statistics). The DM tests of equal predictive ability examine the differences in stock-level squared forecast errors (SEs): $SHE_{i,t+k|t}^{(j)} = \left(R_{i,t+k} - \beta_{i,t+k|t}^{F,(j)} R_{M,t+k} \right)^2$, with $k = 12$. The DM test statistic in month t for comparing the model under investigation j with a competing model i is $DM_t^{(j,i)} = \frac{\bar{d}_t^{(j,i)}}{\hat{\sigma}_{\bar{d}_t^{(j,i)}}}$, where $\bar{d}_t^{(j,i)} = SHE_{i,t+k|t}^{(i)} - SHE_{i,t+k|t}^{(j)}$ is the difference in SEs, $\hat{\sigma}_{\bar{d}_t^{(j,i)}} = \sum_{i=1}^{N_t} w_{i,t} \hat{\sigma}_{i,t}^{(j,i)}$ the value-weighted cross-sectional average of these differences, and $\hat{\sigma}_{\bar{d}_t^{(j,i)}}$ the Newey and West (1987) standard error of $\bar{d}_t^{(j,i)}$ (with 4 lags to account for possible heteroscedasticity and autocorrelation). Statistical tests are based on the 10% significance level. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020. The first beta estimates are obtained in Dec. 1979.

| | Benchmark Estimators | | | | | | | | | | ML Estimators | | | | |
|------------------------------------|----------------------|----------|--------|--------|-------|---------|---------|--------|-------------|-----------|---------------|--------|-------|-------|-------|
| | OLS_5Y_M | OLS_1Y_D | EWMA_S | EWMA_L | BSW | VASICEK | KAROLYI | HYBRID | FAMA-FRENCH | LONG-MEMO | LM | ELANET | RF | GBRT | NN_1 |
| Panel A. Average Forecast Errors | | | | | | | | | | | | | | | |
| MSHE, value-weighted (%) | 7.74 | 7.62 | 7.61 | 7.61 | 7.55 | 7.56 | 7.57 | 7.53 | 7.55 | 7.52 | 7.54 | 7.53 | 7.47 | 7.49 | 7.48 |
| Panel B. Forecast Errors over Time | | | | | | | | | | | | | | | |
| In MCS | 57.17 | 56.13 | 60.29 | 58.63 | 65.49 | 67.15 | 66.53 | 62.16 | 69.44 | 60.71 | 73.60 | 74.84 | 72.77 | 72.97 | 71.31 |
| Benchmark Estimators | | | | | | | | | | | | | | | |
| vs. OLS_5Y_M | | 42.00 | 40.33 | 42.41 | 46.15 | 44.70 | 44.49 | 46.36 | 45.32 | 47.19 | 50.31 | 49.69 | 50.31 | 50.31 | 51.77 |
| vs. OLS_1Y_D | 19.75 | | 20.58 | 22.66 | 49.06 | 56.55 | 55.72 | 43.04 | 41.58 | 35.14 | 34.10 | 37.84 | 46.78 | 38.46 | 39.09 |
| vs. EWMA_S | 18.92 | 22.04 | | 25.57 | 39.09 | 42.00 | 40.12 | 37.84 | 37.01 | 33.68 | 33.47 | 38.88 | 47.19 | 36.59 | 38.25 |
| vs. EWMA_L | 19.96 | 19.96 | 16.84 | | 44.91 | 46.15 | 43.87 | 37.21 | 37.84 | 34.30 | 32.64 | 38.25 | 47.61 | 37.63 | 37.01 |
| vs. BSW | 18.30 | 15.18 | 13.72 | 12.68 | | 20.17 | 21.83 | 30.98 | 20.58 | 26.82 | 25.16 | 30.77 | 40.12 | 31.81 | 29.94 |
| vs. VASICEK | 18.50 | 16.63 | 14.35 | 13.72 | 18.09 | | 17.05 | 33.47 | 21.83 | 28.07 | 23.91 | 30.98 | 39.92 | 31.60 | 29.11 |
| vs. KAROLYI | 20.17 | 12.27 | 14.97 | 12.47 | 23.08 | 28.07 | | 33.89 | 26.20 | 30.35 | 24.53 | 33.06 | 40.12 | 32.64 | 30.56 |
| vs. HYBRID | 16.01 | 12.27 | 13.93 | 12.68 | 22.66 | 24.12 | 23.49 | | 22.66 | 22.25 | 25.57 | 28.69 | 38.25 | 28.48 | 27.65 |
| vs. FAMA-FRENCH | 17.67 | 14.97 | 13.93 | 14.76 | 14.55 | 17.88 | 18.71 | 25.57 | | 28.27 | 21.00 | 24.95 | 33.06 | 27.03 | 27.23 |
| vs. LONG-MEMO | 16.22 | 20.37 | 20.79 | 21.21 | 25.57 | 25.36 | 24.95 | 29.73 | 32.02 | | 27.03 | 35.76 | 40.96 | 35.55 | 33.68 |
| ML Estimators | | | | | | | | | | | | | | | |
| vs. LM | 19.96 | 18.92 | 17.26 | 18.71 | 20.17 | 20.37 | 21.21 | 23.49 | 20.79 | 18.09 | | 29.73 | 30.77 | 25.16 | 24.53 |
| vs. ELANET | 18.92 | 17.26 | 16.22 | 17.26 | 17.46 | 17.67 | 17.88 | 20.37 | 16.63 | 20.37 | 18.30 | | 28.27 | 22.87 | 25.16 |
| vs. RF | 15.59 | 16.01 | 15.80 | 17.46 | 15.80 | 17.46 | 18.92 | 20.79 | 14.97 | 20.37 | 11.85 | 17.05 | | 14.35 | 17.67 |
| vs. GBRT | 16.63 | 16.01 | 18.09 | 17.67 | 18.30 | 18.09 | 18.50 | 21.62 | 18.71 | 20.37 | 17.46 | 22.66 | 28.27 | | 20.17 |
| vs. NN_1 | 15.59 | 13.31 | 14.97 | 14.35 | 17.46 | 18.09 | 17.67 | 19.33 | 21.21 | 15.18 | 16.84 | 28.48 | 30.98 | 19.96 | |
| T | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 | 481 |

and are in the MCS more often than the benchmark models. In addition, these models produce significantly lower average MSHEs than the benchmarks considerably more often than vice versa. Again, the RF model performs best overall.

B. Forecast Errors of Cross-Sectional Portfolio Sorts

In the next step, we provide more insight into when machine learning methods outperform traditional estimators. To do so, we identify types of stocks, for example, high or low beta stocks, large or small stocks, etc., for which differences in prediction errors between beta estimators are particularly large. Following the study by Cosemans et al. (2016), we first examine the extent to which the different abilities to predict future market betas can be attributed to underestimating the betas of low-beta stocks and overestimating those of high-beta stocks. We sort the stocks into decile portfolios based on their predicted betas at the end of month t . In each month, we compute the value-weighted MSE between the predicted betas and the realized betas over the next year within each portfolio. To gain insight into the direction of measurement errors, we also compute the fraction of stocks within each decile portfolio for which the difference between the predicted and realized betas is positive. Ratios below 0.5 indicate that an estimator, on average, underestimates realized betas, while figures above 0.5 indicate an average overestimation.

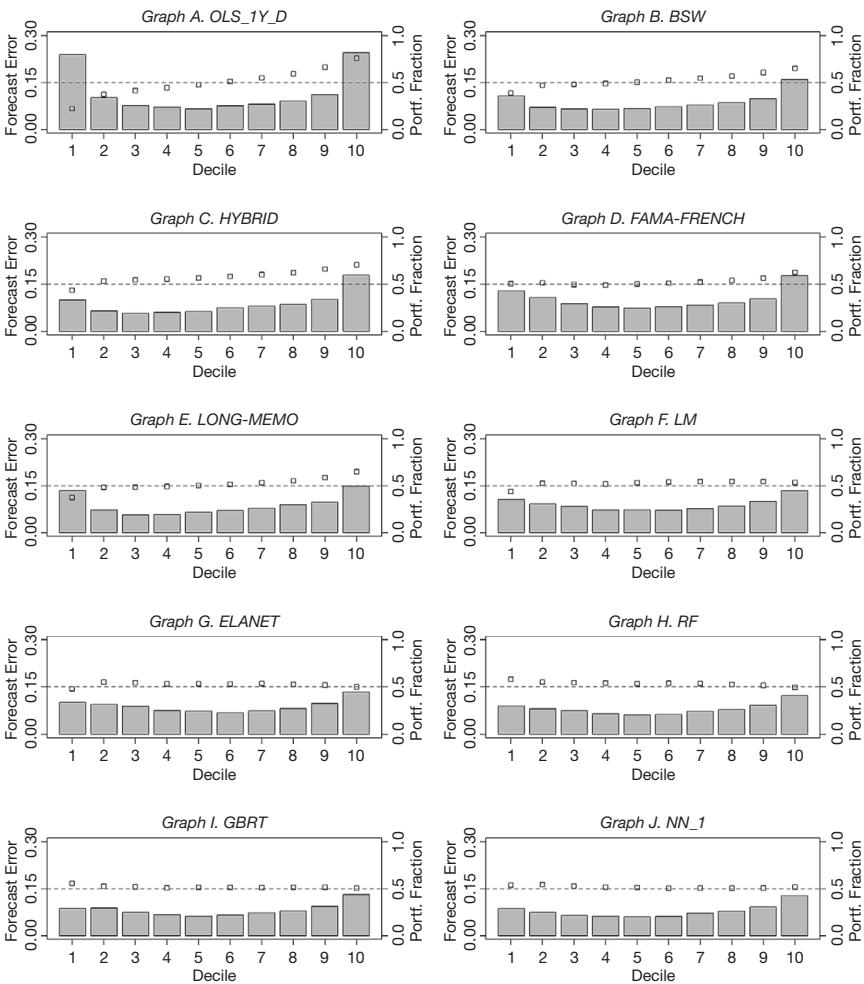
Figure 2 plots the time series averages of the monthly forecast errors within each decile portfolio (gray bars). To keep the presentation focused, we henceforth omit the OLS_5Y_M, EWMA_S, EWMA_L, VASICEK, and KAROLYI models. The results of these models are generally worse than those of the competing forecasting model(s) within the same family shown in Figure 2, that is, the OLS_1Y_D, BSW, and HYBRID models. For all approaches, the extreme portfolios generate the largest average forecast errors. Rolling-beta estimates perform the worst; winsorizing or shrinking them to a well-defined prior, assigning portfolio beta estimates to individual stocks, or exploiting the long-memory properties of beta time series reduce the average forecast error in the extreme beta deciles. The machine learning approaches reduce the forecast errors for (almost) all decile portfolios. In addition, the forecast error distributions are more uniform compared to the classical estimators.

Figure 2 also illustrates the average overestimation fractions (black unfilled squares). The problem of underestimating the betas of stocks in low-beta deciles and overestimating those in high-beta deciles is evident for all benchmark approaches (albeit to varying degrees). This problem is less pronounced for some of the better performing models (e.g., the FAMA-FRENCH and LONG-MEMO models), although they cannot avoid it entirely. In contrast, the machine learning techniques show no evidence of systematic underestimation in the low-beta deciles or systematic overestimation in the high-beta deciles. This pattern can be explained by less extreme beta estimates, as indicated by the lower cross-sectional forecast dispersions (as shown in Table C1 of the Supplementary Material).

In the next step, we analyze how differences in forecast errors across beta estimators are related to other firm characteristics or industry classifications. For the sake of brevity, we focus on the comparison of RF with OLS_1Y_D and LM. We repeat the procedure outlined in Figure 2, but now sort the stocks into decile portfolios based on firm size (ME), book-to-market (BM), momentum (MOM),

FIGURE 2
Average Forecast Errors of Portfolio Sorts Based on Beta Estimates

Figure 2 shows the time series averages of the monthly forecast errors of the portfolios sorted on the basis of beta estimates (gray bars). Stocks are sorted into decile portfolios based on their predicted betas at the end of each month t , separately for each of the selected forecasting models introduced in Section IV. The forecast error in this test is defined as the value-weighted MSE between beta forecasts and realized betas over the next year within each portfolio. Added to these visualizations are the time series averages of the overestimation fractions (black unfilled squares). They are computed at the end of each month t as the fraction of stocks within each decile portfolio for which the difference between beta forecasts and realized betas is positive. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020, while the first beta estimates are obtained in Dec. 1979.



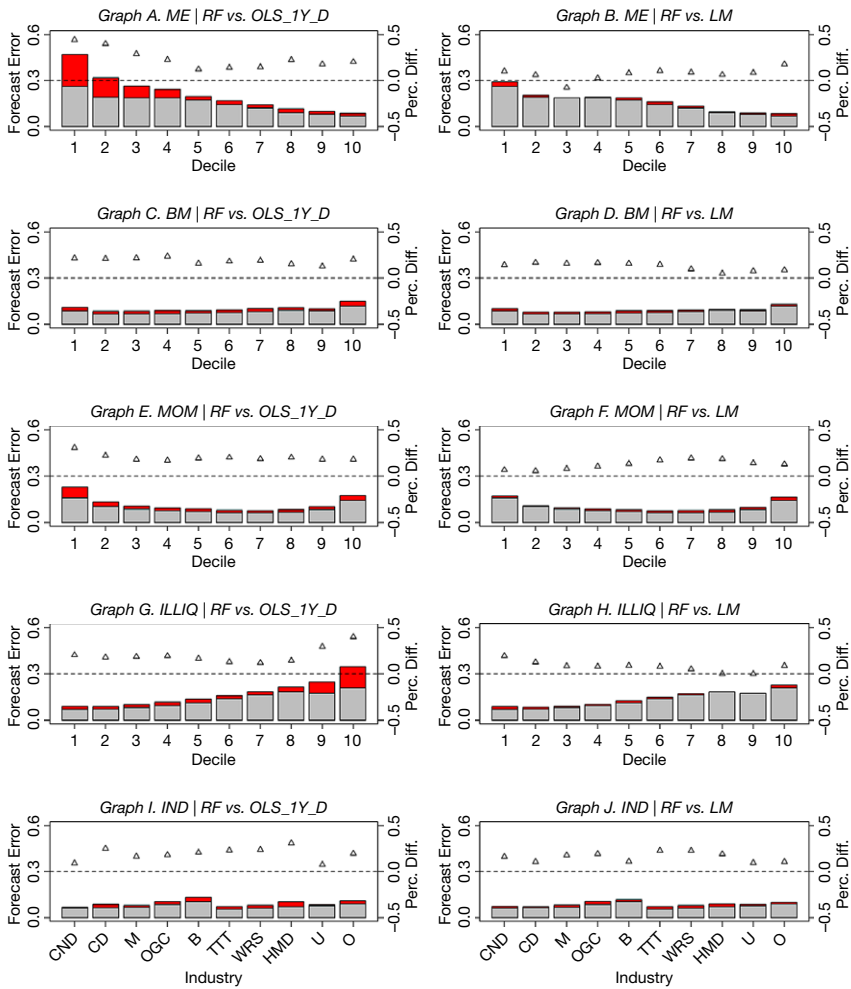
illiquidity (ILLIQ), and industry classification (IND).¹⁷ Figure 3 plots the time series averages of the forecast errors within each decile portfolio for the RF model (gray bars) and the corresponding benchmark model (red bars). To these visualizations, we add the percentage differences in the average forecast errors relative to

¹⁷For these visualizations, we consider only 10 rather than 47 dummies, corresponding to the industry classification of Fama and French (1997).

FIGURE 3

Average Forecast Errors of Portfolio Sorts Based on Firm Characteristics and Industry Classification

Figure 3 shows the time series averages of the monthly forecast errors of portfolio sorts based on firm characteristics and industry classifications. To do so, the procedure outlined in Figure 2 is repeated, but the stocks are sorted into decile portfolios based on firm size (ME), book-to-market (BM), momentum (MOM), illiquidity (ILLIQ), and Fama and French's (1997) industry classification, that is, consumer nondurables (CND), consumer durables (CD), manufacturing (M), oil, gas, and coal extraction and products (OGC), business equipment (B), telephone and television transmission (TTT), wholesale, retail, and some services (WRS), healthcare, medical equipment, and drugs (HMD), utilities (U), and other (O). The graphs plot the time series averages of the forecast errors within each decile portfolio for the RF model (gray bars) and the respective benchmark model (red bars), that is, the OLS_1Y_D (Graphs A, C, E, G, I) and LM (Graphs B, D, F, H, J) models. The monthly forecast error in this test is defined as the value-weighted mean squared error (MSE) between beta forecasts and realized betas over the next year within each portfolio. These visualizations are complemented by the percentage differences in average forecast errors relative to the respective benchmark model (black unfilled triangles), calculated as 1 minus the average MSE of the random forests divided by the average MSE of the respective benchmark model. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020, while the first beta estimates are obtained in Dec. 1979.



the respective benchmark model (black unfilled triangles), calculated as 1 minus the average MSE of the random forests divided by the average MSE of the respective benchmark model.

Taken together, the graphs of Figure 3 suggest that random forests reduce the forecast errors relative to the OLS_1Y_D (Graphs A, C, E, G, and I) and LM (Graphs B, D, F, H, and J) models for almost all decile portfolios, as indicated by percentage differences greater than 0 (triangles above the dashed line). Consistent with Figure 2, random forests generally provide more accurate beta estimates. The higher average MSEs for 1-year rolling betas and simple linear regressions are not predominantly driven by high forecast errors for just a few stocks with specific firm characteristics. However, compared to the OLS_1Y_D model, random forests reduce the forecast errors even more within the extreme decile portfolios, both in absolute and relative terms.

Especially for small and illiquid stocks, the RF model provides more accurate beta estimates than those obtained from the OLS_1Y_D model. In addition, although less pronounced, we observe improvements in random forests for value and loser stocks. Because the outperformance of random forests over simple linear regressions is marginal for these stocks, we attribute this observation to the inclusion of slow-moving firm fundamentals as predictors rather than to the RF model's ability to capture nonlinearity and interactions. Finally, we observe that the random forests outperform the two benchmarks in every single industry. Compared to the OLS_1Y_D model, their value-added is greatest for "consumer durables" (CD), "wholesale, retail, and some services" (WRS), and "healthcare, medical equipment, and drugs" (HMD), while the RF model is most beneficial relative to simple linear regressions for "telephone and television transmission" (TTT) and "wholesale, retail, and some services" (WRS).

In summary, for traditional approaches, very small and very large beta forecasts as well as beta forecasts for stocks with extreme firm characteristics or within specific industries should be used with caution. In contrast, machine learning-based beta forecasts appear to be uniformly useful for all types of stocks.

VI. Economic Value of Market Beta Forecasts

Since accurate beta estimates are critical for several applications in academia and industry, we now examine whether statistically more accurate forecasts lead to economic gains in portfolio construction. In this section, we focus our analysis only on applications that derive economic value directly from the beta estimates. Therefore, we assess hedging approaches, market-neutral anomaly portfolios, and MVPs. For many other portfolio optimization applications, however, beta estimates are not sufficient and must be supplemented with forecasts for expected returns. We do not consider these approaches in this article.

A. Anomaly Portfolio Hedging

In a first step, we analyze the hedging performance of different beta estimators for anomaly portfolios. Specifically, we consider a typical hedge fund strategy, where an investor attempts to exploit an anomaly without taking any market risk. We examine the average ex post realized betas of ex ante market-neutral long–short

anomaly portfolios. We consider commonly used anomaly variables, such as size (ME), book-to-market (BM), and illiquidity (ILLIQ). Note that we separate this pure hedging objective from strategies to enhance the performance of anomaly portfolios in Section VI.B.

The portfolio optimization we use extends the procedure described in the study by Hollstein et al. (2019). At the end of each month t , we first sort the stocks into decile portfolios based on the respective anomaly variables. We then use the out-of-sample beta forecasts of the stocks in the top and bottom deciles to construct the long and short portfolios, respectively. We require that the ex ante predicted portfolio beta is 1 for both decile portfolios. To achieve this, we choose portfolio weights in the top and bottom decile portfolios that solve the following optimization problem, separately for each beta estimator:

$$(5) \quad \min_{w_t} \sum_i \left(w_{i,t} - w_{i,t}^* \right)^2 \text{ s.t.} \\ w_{i,t} \geq 0 \\ \sum_{i=1}^{N_t} w_{i,t} \beta_{i,t+1|t}^F = 1.$$

The optimization algorithm aims to minimize the sum of the squared deviations from the original market capitalization-based weights $w_{i,t}^*$ of the stocks in the respective portfolios. This approach helps ensure that the resulting long and short portfolios are indeed investable for a hedge fund. The first constraint implies that each decile portfolio must be long-only, while the second implies that the ex ante-predicted portfolio betas must be 1. Combining the long and short portfolios (by multiplying the computed weights for the short portfolio by -1) yields a long–short (HML) anomaly portfolio that is ex ante market neutral. Alternatively, investors can use the long-only portfolios, which can then be made market neutral by taking a short position in exchange-traded index funds or futures written on a market proxy as the underlying. Consistent with the ex ante portfolio betas, the ex post portfolio betas are the weighted averages of the realized betas over the next year.

Table 4 reports the time series averages of the ex post portfolio betas (β) of the long (H), short (L), and long–short (HML) anomaly portfolios. The t -statistics in parentheses are based on Newey and West (1987) robust standard errors with 11 lags. The null hypotheses for the H and L portfolios are that the ex post betas are 1 and for the HML portfolio that the ex post beta is 0.

We find that only the RF, GBRT, and NN_1 approaches produce long–short portfolios that are truly market neutral ex post for all stock market anomalies. For example, the RF model yields realized portfolio betas ranging from -0.03 for the illiquidity anomaly to 0.08 for the size anomaly. According to t -statistics ranging from -0.49 to 0.66 , these betas are insignificantly different from 0. In contrast, the ex post betas of market-neutral strategies derived from the benchmark estimators are economically large and statistically significant in most cases. For example, the OLS_1Y_D estimator fails to generate market-neutral size and illiquidity

TABLE 4
Anomaly Portfolio Hedging

Table 4 reports the properties of market-neutral anomaly portfolios constructed using beta estimates from the forecasting models introduced in Section IV. The anomaly variables are size (ME), book-to-market (BM), and illiquidity (ILLIQ). The portfolio optimization is described in Section VI.A. The table presents the time series averages of the ex post portfolio betas (β) for the long (H), short (L), and long-short (HML) anomaly portfolios constructed from decile sorts. The t -statistics in parentheses are based on Newey and West (1987) robust standard errors with 11 lags. The null hypotheses for the H and L portfolios are that the ex post betas are 1 and for the HML portfolio that the ex post beta is 0. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020. The first beta estimates are obtained in Dec. 1979.

| Model | ME | | | BM | | | ILLIQ | | |
|-----------------------------|----------------|-----------------|-----------------|------------------|-----------------|-----------------|------------------|------------------|-----------------|
| | β_{HML} | β_H | β_L | β_{HML} | β_H | β_L | β_{HML} | β_H | β_L |
| <i>Benchmark Estimators</i> | | | | | | | | | |
| OLS_1Y_D | 0.34 (3.54) | 0.99 (-0.14) | 0.66 (-4.29) | -0.04 (-0.81) | 0.96 (-0.94) | 1.00 (0.13) | -0.30 (-4.80) | 0.68 (-6.26) | 0.99 (-0.80) |
| BSW | 0.33 (3.27) | 1.01 (0.17) | 0.68 (-3.94) | -0.04 (-0.74) | 0.98 (-0.46) | 1.02 (0.63) | -0.30 (-4.60) | 0.70 (-5.95) | 1.00 (-0.15) |
| HYBRID | 0.45 (5.33) | 1.01 (0.14) | 0.56 (-6.98) | -0.05 (-1.10) | 0.96 (-1.07) | 1.02 (0.48) | -0.40 (-7.71) | 0.59 (-10.57) | 0.99 (-0.38) |
| FAMA-FRENCH | 0.19 (2.29) | 0.99 (-0.33) | 0.80 (-2.78) | -0.02 (-0.33) | 1.00 (0.05) | 1.02 (0.78) | -0.18 (-3.36) | 0.81 (-4.54) | 0.98 (-1.39) |
| LONG-MEMO | 0.25 (3.88) | 0.98 (-0.49) | 0.73 (-4.71) | -0.01 (-0.10) | 0.99 (-0.35) | 0.99 (-0.24) | -0.22 (-5.33) | 0.74 (-6.90) | 0.97 (-2.38) |
| <i>ML Estimators</i> | | | | | | | | | |
| LM | 0.18 (2.58) | 0.95 (-1.74) | 0.77 (-3.75) | -0.04 (-0.93) | 0.94 (-1.62) | 0.98 (-0.51) | -0.14 (-3.28) | 0.79 (-5.65) | 0.94 (-3.44) |
| ELANET | 0.20 (2.45) | 0.96 (-1.24) | 0.76 (-3.64) | -0.01 (-0.23) | 0.97 (-0.70) | 0.98 (-0.66) | -0.17 (-3.41) | 0.78 (-5.80) | 0.95 (-2.83) |
| RF | 0.08 (0.66) | 1.00 (-0.07) | 0.92 (-0.77) | -0.02 (-0.29) | 0.99 (-0.22) | 1.01 (0.28) | -0.03 (-0.49) | 0.95 (-0.83) | 0.98 (-1.17) |
| GBRT | 0.09 (0.93) | 1.01 (0.44) | 0.92 (-0.90) | 0.00 (-0.02) | 1.02 (0.37) | 1.02 (0.71) | -0.06 (-0.94) | 0.93 (-1.09) | 0.99 (-0.48) |
| NN_1 | 0.04 (0.29) | 1.01 (0.24) | 0.97 (-0.28) | -0.02 (-0.31) | 1.00 (-0.04) | 1.01 (0.54) | -0.04 (-0.62) | 0.95 (-0.89) | 0.99 (-0.87) |

portfolios. Moreover, the machine learning-based models perform well for the long-only portfolios, with ex post betas close to 1.

In summary, the results of this simple portfolio hedging exercise illustrate the practical consequences of inaccurate beta estimates. An investment strategy that is supposed to be market neutral ex ante may still have significant market risk ex post. We find that traditional estimation techniques fail to produce truly market-neutral portfolios ex post, while the machine learning-based approaches perform significantly better.

B. Anomaly Portfolio Performance

Market betas can be used not only for pure hedging but also to improve the performance of anomaly investments. Grundy and Martin (2001) and Daniel and Moskowitz (2016) consider market-neutral momentum strategies that attempt to improve performance by hedging the dynamic market exposure. Frazzini and Pedersen (2014) and Novy-Marx and Velikov (2022) consider different specifications of betting-against-beta anomalies. Finally, idiosyncratic volatility is also related to beta estimates. Therefore, the use of better beta estimates can potentially help to improve the investment performance of these anomalies.

In this section, we consider the abnormal returns of market-neutral momentum, idiosyncratic volatility, and betting-against-beta strategies based on the

different beta estimators. Each month, we construct decile portfolios by sorting the stocks based on their momentum (MOM), idiosyncratic volatility (IVOL), and beta estimates (BAB). For the latter, we use the predicted beta of each forecasting model. The anomaly portfolios go long and short in the extreme deciles. For momentum, the resulting portfolio goes long in decile 10 and short in decile 1, while those for the other 2 anomalies go long in decile 1 and short in decile 10. Finally, the portfolios are hedged each month with a position in the market portfolio equal to the negative of the portfolio beta predicted by the forecasting models.

For all analyses of portfolio performance in this section and the subsequent [Section VI.C](#), we follow Chan, Karceski, and Lakonishok (1999) and Cosemans et al. (2016) and focus on liquid and investable stocks with market capitalizations above the 20th percentile of NYSE stocks.¹⁸ We report the annualized alphas of the returns over the next month of these strategies with respect to the CAPM and the Fama and French (2015) 5-factor model (FF5). Finally, we also report the ex post betas of the strategies (β). The *t*-statistics (in parentheses) are based on Newey and West (1987) robust standard errors, with 4 lags for the alpha tests and 11 lags for the beta tests.

[Table 5](#) shows the results. We start with the market-neutral momentum strategies. The strategies based on all the different beta estimators produce positive alphas that are statistically significant, but it is the machine learning-based approaches that produce the largest alphas. For example, for the main benchmark model, OLS_1Y_D, the Fama and French (2015) 5-factor alpha for the market-neutral momentum strategy is 8.62%. Those for the other benchmarks are of a similar magnitude below 9%, and only the LONG-MEMO model achieves a higher Fama and French (2015) 5-factor alpha of 9.44%. The corresponding alphas for the RF, GBRT, and NN_1 models are 9.79%, 9.74%, and 9.49%, respectively. In addition, only the machine learning-based estimators generate truly market-neutral portfolios. For all benchmarks, the realized betas of the supposedly market-neutral portfolios are significantly negative, while they are insignificantly different from 0 for the machine learning models. Therefore, market-neutral momentum portfolios based on machine learning clearly outperform those based on the benchmarks.

Next, we consider market-neutral idiosyncratic volatility portfolios. For these, we also find that the machine learning-based estimators perform well. They generate the largest alphas. For example, the Fama and French (2015) 5-factor alphas are 9.46% for RF, 9.65% for GBRT, and 9.69% for NN_1, compared to only 8.93% for the OLS_1Y_D estimator. While the portfolios are ex post market neutral for the GBRT and NN_1 models, this is not the case for the RF model.

Finally, we turn to the betting-against-beta strategies. Again, the estimated alphas are largest for the machine learning-based estimators.¹⁹ The Fama and French (2015) 5-factor alpha for the OLS_1Y_D estimator is 6.95%. The alphas

¹⁸Therefore, we focus on the economically most important stocks (Hou, Xue, and Zhang (2020)). As shown in [Section V.B](#), the machine learning-based methods outperform the benchmarks even more for microcap stocks, which we exclude from this analysis.

¹⁹Consistent with the higher betting-against-beta portfolio returns, we find that the CAPM cannot be saved even with machine learning-based betas. A Fama and MacBeth (1973) regression test of the model (untabulated) shows that the intercepts are similarly large and significantly positive for all beta estimators, and the slope coefficients are generally insignificant and negative. Therefore, the rejection of the CAPM does not seem to be due to inaccurate beta forecasts.

TABLE 5
Anomaly Portfolio Performance

Table 5 shows the investment performance of market-neutral anomaly portfolios. Each month, we construct decile portfolios by sorting the stocks by their momentum (MOM), idiosyncratic volatility (IVOL), and beta estimates (BAB). For the latter, we use the predicted beta of each forecasting model. The anomaly portfolios go long and short in the extreme deciles. For momentum, the resulting portfolio goes long in decile 10 and short in decile 1, while those for the other two anomalies go long in decile 1 and short in decile 10. Finally, the portfolios are hedged each month with a position in the market portfolio equal to the negative of the portfolio beta predicted by the forecasting models. The table reports the annualized alphas of the returns over the next month of these strategies with respect to the CAPM and the Fama and French (2015) 5-factor model (FF5). Finally, the ex post betas of the strategies (β) are shown. The t -statistics in parentheses are based on Newey and West (1987) robust standard errors, with 4 lags for the alpha tests and 11 lags for the beta tests. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020 and have a market capitalization above the 20th percentile of NYSE stocks. The first beta estimates are obtained in Dec. 1979.

| Model | MOM | | | IVOL | | | BAB | | |
|-----------------------------|------------------------|-----------------------|------------------|------------------------|-----------------------|------------------|------------------------|-----------------------|------------------|
| | α_{CAPM} [%] | α_{FF5} [%] | β | α_{CAPM} [%] | α_{FF5} [%] | β | α_{CAPM} [%] | α_{FF5} [%] | β |
| <i>Benchmark Estimators</i> | | | | | | | | | |
| OLS_1Y_D | 5.95 (1.67) | 8.62 (2.46) | −0.09 (−2.40) | 10.94 (4.25) | 8.93 (4.19) | 0.05 (1.12) | 8.13 (2.68) | 6.95 (2.32) | 0.41 (3.37) |
| BSW | 6.20 (1.74) | 8.89 (2.57) | −0.10 (−2.07) | 10.93 (4.31) | 8.87 (4.24) | −0.05 (−1.20) | 6.91 (2.31) | 5.64 (1.93) | 0.23 (2.33) |
| HYBRID | 6.06 (1.70) | 8.67 (2.48) | −0.11 (−2.77) | 10.84 (4.28) | 8.81 (4.20) | 0.01 (0.29) | 9.54 (3.15) | 7.89 (2.71) | 0.23 (1.81) |
| FAMA–FRENCH | 6.01 (1.68) | 8.76 (2.55) | −0.12 (−2.48) | 10.91 (4.18) | 8.86 (4.11) | −0.06 (−1.50) | 8.05 (2.83) | 6.74 (2.46) | 0.20 (1.86) |
| LONG-MEMO | 6.62 (1.86) | 9.44 (2.82) | −0.11 (−2.17) | 11.46 (4.52) | 9.45 (4.45) | −0.04 (−1.27) | 8.95 (3.11) | 7.64 (2.67) | 0.19 (2.47) |
| <i>ML Estimators</i> | | | | | | | | | |
| LM | 6.84 (1.88) | 9.32 (2.58) | −0.09 (−1.34) | 11.68 (4.59) | 9.54 (4.47) | −0.08 (−1.91) | 9.00 (2.68) | 7.01 (2.30) | 0.03 (0.39) |
| ELANET | 6.26 (1.70) | 8.66 (2.44) | −0.09 (−1.30) | 11.61 (4.48) | 9.35 (4.34) | −0.09 (−2.10) | 9.03 (2.83) | 7.19 (2.51) | −0.03 (−0.33) |
| RF | 7.05 (1.97) | 9.79 (2.91) | −0.07 (−1.13) | 11.54 (4.54) | 9.46 (4.49) | −0.09 (−2.23) | 9.31 (2.91) | 7.81 (2.57) | −0.05 (−0.55) |
| GBRT | 6.98 (1.95) | 9.74 (2.90) | −0.04 (−0.70) | 11.74 (4.60) | 9.65 (4.53) | −0.07 (−1.57) | 10.13 (3.13) | 8.69 (2.82) | −0.02 (−0.26) |
| NN_1 | 6.79 (1.89) | 9.49 (2.79) | −0.04 (−0.66) | 11.80 (4.63) | 9.69 (4.58) | −0.01 (−0.31) | 10.01 (3.13) | 8.45 (2.71) | 0.03 (0.31) |

for the RF, GBRT, and NN_1 models are clearly larger at 7.81%, 8.69%, and 8.45%, respectively. Finally, all machine learning approaches produce portfolios that are ex post market neutral, while the betting-against-beta portfolios generated by the benchmark models are generally not.^{20,21}

²⁰Novy-Marx and Velikov (2022) argue that, in addition to unconditional market neutrality, conditional market neutrality is important for betting-against-beta portfolios. In Section C of the Supplementary Material, we apply their regression approach to confirm that all machine learning-based betting-against-beta strategies are also conditionally market neutral (see Table C10 in the Supplementary Material).

²¹In Table C12 in the Supplementary Material, we report the t -statistics of pairwise tests on the differences in alphas between the different estimators. The statistically strongest differences are observable for the momentum portfolios. For these portfolios, the RF and GBRT forecast models generate significantly larger FF5 alphas than the OLS_1Y_D, BSW, HYBRID, and FAMA–FRENCH models. For the IVOL and BAB portfolios, most of the t -statistics for the alpha comparisons between the machine learning and benchmark models are not statistically significant. However, to put this result in perspective, note that all of these anomaly strategies have two inputs: (i) the anomaly signal and (ii) the beta estimates. Because (i) is the same for all strategies, it seems natural that not all differences in alphas are statistically significant.

C. Minimum Variance Portfolios

The previous subsections demonstrate that machine learning-based estimators produce better market-neutral anomaly portfolios. However, market betas can also be used to construct portfolios of particular interest to investors that are not based on a single anomaly. MVPs are a prominent example. Since expected returns do not enter the MVP optimization, differences in stock weights in the optimized portfolio result solely from differences in the estimated covariance matrices. As in Cosemans et al. (2016), we assume a single-factor structure for the high-dimensional covariance matrix of the stocks. Therefore, the stock market beta forecasts can be used to obtain covariance matrix forecasts. Ultimately, differences in beta estimates are the only source of differences in stock weights in the MVP.

At the end of each month t , we predict the out-of-sample covariance matrix as $\Omega_{t+1|t} = s_{M,t+1|t}^2 B_{t+1|t} B'_{t+1|t} + D_{t+1|t}$, where $B_{t+1|t}$ is the $N_t \times 1$ vector of out-of-sample beta forecasts, $s_{M,t+1|t}^2$ is the out-of-sample forecast of the market variance (variance of market excess returns), and $D_{t+1|t}$ is the diagonal matrix containing the out-of-sample forecasts of the idiosyncratic variances $d_{i,t+1|t}^2$. The idiosyncratic returns are computed as the differences between realized and estimated stock returns: $r_{i,t} - \beta_{i,t}^F r_{M,t}$. Both market and idiosyncratic variances are obtained from daily returns over the past year ending in month t , so that these historical values are used as predictions for month $t + 1$. We use these out-of-sample covariance forecasts to construct the MVP by selecting portfolio weights that solve the following problem, separately for each beta estimator:

$$(6) \quad \min_{w_t} w_t' \Omega_{t+1|t} w_t \text{ s.t.} \\ 0 \leq w_{i,t} \leq 0.05 \\ \sum_{i=1}^{N_t} w_{i,t} = 1.$$

The first constraint implies that the portfolio weights must be within a reasonable range (due to short selling restrictions and industry maximum weight rules), while the second implies that the portfolio must be fully invested. Since a hedge fund typically has a short investment horizon, we rebalance the portfolio at the end of each month t and record the realized return in the next month $t + 1$. To evaluate the performance of the resulting MVP, we obtain return and risk measures based on the monthly portfolio returns.²²

Panel A of Table 6 reports the results for the different MVPs. We observe that the machine learning-based approaches produce better MVPs than all the benchmark models. Most importantly, the ex post standard deviations of the MVPs are substantially lower for the machine learning-based estimators. For example, the ex post MVP standard deviation of the RF estimator is 11.42%, while that of the

²²We use monthly returns to be consistent with industry practice and the previous academic literature (e.g., Ghysels and Jacquier (2006), Cosemans et al. (2016)). The results are qualitatively similar when using daily returns. Therefore, the differences in beta estimates based on daily and monthly returns documented by Gilbert, Hrdlicka, Kalodimos, and Siegel (2014) do not seem to play a major role in this application.

TABLE 6
Minimum Variance Portfolios

Table 6 reports the properties of the MVPs constructed based on beta estimates obtained from the forecasting models introduced in Section IV. For the portfolio optimization, we impose a single-factor structure on the covariance matrix of stock returns. Therefore, the market betas are the primary determinants of the stock weights in the MVP. The approach is described in detail in Section VI.C. Each month, we compute the weights that minimize the expected portfolio variance, subject to the constraints that the weights are positive, that each individual weight is less than 5%, and that the weights sum to 1. The forecasts for the market and idiosyncratic variances are based on daily returns over the previous year. Panel A presents the annualized risk and return measures of the resulting MVPs. Std. Dev. reports the ex post time series standard deviation and DWND the ex post downside standard deviation (of negative returns). Min is the lowest monthly excess return and MaxDD is the maximum drawdown of the MVP from peak to trough over multimonth periods. TV is the terminal value in Dec. 2019 of a \$1 investment in the MVP in Dec. 1979. Mean is the average portfolio return, and SR is the Sharpe ratio. Panel B reports the ex post market betas of the MVPs (β_{pv}) as well as the beta of a market-neutral MVP that hedges the expected market risk (depending on the portfolio beta forecast) each month with an additional investment in the market portfolio (β_{mn}). The *t*-statistics in parentheses are based on Newey and West (1987) robust standard errors with 11 lags. The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020 and have a market capitalization above the 20th percentile of NYSE stocks. The first beta estimates are obtained in Dec. 1979.

| Model | Panel A. Minimum Variance | | | | | | | Panel B. Market Neutrality | |
|-----------------------------|---------------------------|-------------|------------|--------------|-----------|-------------|------|----------------------------|------------------|
| | Std. Dev. (%) | DWND (%) | Min (%) | MaxDD (%) | TV (%) | Mean (%) | SR | β_{pv} | β_{mn} |
| <i>Benchmark Estimators</i> | | | | | | | | | |
| OLS_1Y_D | 12.40 | 9.94 | −24.24 | 43.55 | 27.98 | 9.12 | 0.74 | 0.39 (5.19) | −0.22 (−6.41) |
| BSW | 12.06 | 9.29 | −21.91 | 43.07 | 27.29 | 9.01 | 0.75 | 0.37 (4.41) | −0.16 (−4.01) |
| HYBRID | 12.20 | 9.36 | −21.88 | 41.12 | 26.24 | 8.93 | 0.73 | 0.36 (4.30) | −0.15 (−2.96) |
| FAMA–FRENCH | 12.02 | 9.52 | −23.04 | 37.08 | 25.62 | 8.85 | 0.74 | 0.40 (7.95) | −0.07 (−1.99) |
| LONG-MEMO | 11.93 | 9.04 | −20.35 | 34.73 | 28.96 | 9.15 | 0.77 | 0.36 (4.64) | −0.12 (−3.23) |
| <i>ML Estimators</i> | | | | | | | | | |
| LM | 11.71 | 8.75 | −12.98 | 50.15 | 35.35 | 9.62 | 0.82 | 0.41 (4.36) | −0.13 (−2.63) |
| ELANET | 11.91 | 9.98 | −22.34 | 49.34 | 30.94 | 9.32 | 0.78 | 0.41 (4.28) | −0.10 (−1.96) |
| RF | 11.42 | 8.31 | −19.32 | 39.12 | 33.07 | 9.42 | 0.82 | 0.35 (4.23) | 0.01 (0.15) |
| GBRT | 11.10 | 8.42 | −18.82 | 38.85 | 42.41 | 10.01 | 0.90 | 0.36 (4.24) | 0.00 (−0.01) |
| NN_1 | 11.16 | 8.11 | −16.28 | 35.08 | 37.51 | 9.70 | 0.87 | 0.35 (4.42) | −0.03 (−0.61) |

OLS_1Y_D estimator is 12.40%. Even for the best benchmark estimator, the ex post standard deviation is higher than for all machine learning-based approaches. While minimum variance is the sole objective of the optimization, investors may also care about other portfolio performance metrics. The machine learning-based estimators perform well on each of them. They produce the smallest downward variations, the least negative minimum returns, small maximum drawdowns, as well as the highest average returns, terminal values (based on an initial investment of \$1 in Dec. 1979), and Sharpe ratios. Taken together, machine learning methods can be used to generate better MVPs.²³

Hedge funds may also be interested in making these MVPs market neutral. Therefore, we analyze the realized betas of the pure MVPs as well as their

²³In Section C of the Supplementary Material, we show results separately for the first and second halves of the sample period. These results are qualitatively similar, implying that the performance of the MVPs is stable over time (see Tables C8 and C9 in the Supplementary Material).

market-neutral versions. First, one might suspect that the machine learning-based estimators outperform the benchmarks simply because their MVPs have different ex post betas. As shown in Panel B of Table 6, this is not the case. For all estimators, the ex post MVP betas (β_{pv}) are of similar magnitude. Second, we analyze the performance of a strategy that uses long or short positions in the market portfolio to hedge the predicted beta of the MVP. For example, if the expected beta is 0.2, the hedging strategy adds an additional short position of 0.2 times the portfolio value in the market portfolio. We report the ex post beta (β_{mv}) of this strategy, examining the ability of the different estimators to hedge the market risk. The machine learning-based estimators, in particular RF, GBRT, and NN_1, also excel at this task and generate truly market-neutral MVPs. In contrast, the ex post realized betas of the hedging strategy are significantly different from 0 for all benchmarks.

We conclude that machine learning-based beta estimates are not only statistically superior to their more traditional benchmarks, but they also contain superior economic information that can be exploited to arrive at better portfolio decisions.

VII. Properties and Operating Scheme of Machine Learning Estimators

The previous sections show that machine learning-based estimators outperform the established beta estimators both statistically and economically. In a final step, we focus on determining *how* these techniques, often referred to as “black boxes,” achieve this outperformance. We address the black box issue in market beta estimation by examining the properties and operating scheme of random forests because, on balance, random forests outperform both GBRT and neural networks.²⁴ In particular, we examine changes in the inherent model complexity over time and decompose predictions into the contributions of individual variables using relative variable importance metrics. Moreover, in Section C of the Supplementary Material, we discuss examples that illustrate the patterns of nonlinear and interactive effects in the relationship between predictor variables and beta estimates.

A. Model Complexity

Since we reestimate the random forests on an annual basis, it is interesting to measure whether the model complexity changes over time or rather remains stable. Since the RF model is nonparametric and tree-based, we use the number of trees added to the ensemble prediction (MC) to measure model complexity. For example, a large number of trees indicates high model complexity, that is, the respective random forest needs information from several different bootstrap-replicated trees to optimally explain the cross-sectional variation in realized betas within the validation sample. A smaller number of trees, on the other hand, indicates that a less complex model is sufficient to meet the goal of minimizing the validation error. To contextualize the model complexity measure, we compute the time series mean of the monthly MSEs within the validation sample (MSE_VALI) and the test sample

²⁴In results not reported, we observe that the patterns identified for both the GBRT and the NN_1 model, as well as their implications, are qualitatively similar.

(MSE_TEST) for each annual reestimation cycle. We also relate the MSE_TEST metrics of the RF model to those obtained for the standard benchmark model (OLS_1Y_D) and compute the monthly percentage difference in test sample MSEs (MSE_TEST_PD).²⁵

Figure 4 illustrates the model complexity of random forests over the sample period and its association with MSE_VALI (Graph A), MSE_TEST (Graph B), and MSE_TEST_PD (Graph C). Again, it also contains the NBER recession periods (gray-shaded areas). Consistent with Gu et al. (2020) and Drobetz and Otto (2021), the graphs suggest that the complexity of the RF model varies significantly over time. For example, many trees are required at multiple reestimation dates during the period between 2000 and 2009 (with a global peak in Dec. 2001). In contrast, the complexity of the RF model is much lower in the following years, that is, in the period 2010–2019.

In Graph A of Figure 4, we find co-movement between MC and MSE_VALI, as indicated by a time series correlation of 0.90. A *t*-test significantly rejects the null hypothesis that the correlation coefficient is 0 (untabulated). Consequently, the model complexity varies substantially over time depending on the stock market conditions. In particular, we find that the complexity of the RF model is high during periods that are difficult to predict (large validation errors). When future betas are easier to predict, a smaller number of trees is sufficient for the ensemble forecast to minimize the validation error.

It is interesting to note that we do not find any synchronicity between the MC and MSE_TEST metrics (time series correlation of -0.06) in Graph B of Figure 4. According to an insignificant *t*-statistic, we cannot reject the null hypothesis that this correlation is 0 and conclude that neither high- nor low-complexity forecasts systematically coincide with high or low forecast errors within the test samples. This result highlights the need to adaptively determine the hyperparameters governing model complexity from the sample data rather than forcing them to remain constant. Finally, there is no significant co-movement between MC and MSE_TEST_PD in Graph C, with a time series correlation of only 0.17. Random forests reduce the test sample forecast error relative to the OLS_1Y_D model for most of the sample period (as shown in Figure 1). However, the relative out-performance appears to be only weakly related to the complexity of the RF model.

B. Variable Importance

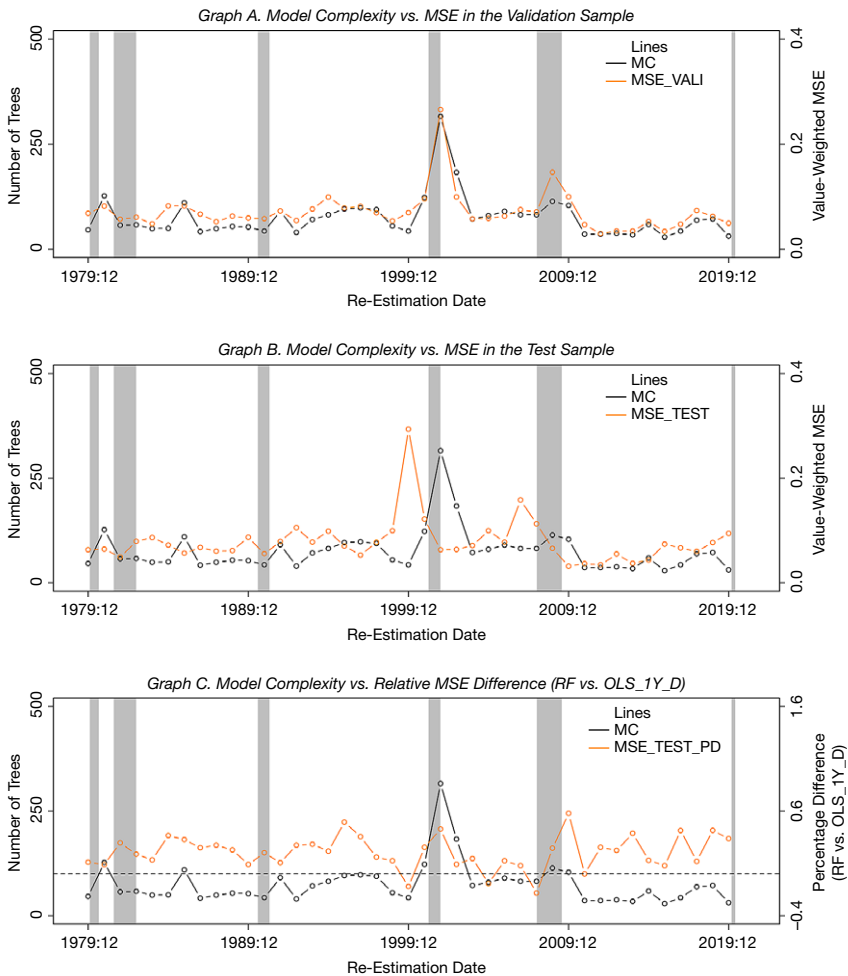
Since the degree of model complexity is time-varying, it is instructive to examine whether the contribution of each predictor to the overall predictive ability of random forests also changes over time.²⁶ We compute the variable importance using a 2-step approach, separately for each reestimation date: First, we compute the

²⁵It is defined as 1 minus the random forest MSE divided by that of the benchmark model. Again, we follow the convention that positive differences indicate superior predictive performance of random forests relative to the OLS_1Y_D model.

²⁶For the sake of brevity, we exclude the industry classifiers throughout the variable importance tests because they are among the least informative predictors. Note that the patterns identified and their implications are qualitatively similar when they are included.

FIGURE 4
Model Complexity over Time

Figure 4 illustrates the model complexity of random forests (RF, introduced in Section IV.B) over the sample period and its association with forecast errors. Since the RF model is nonparametric and tree-based, the number of trees added to the ensemble prediction (MC) is used to measure model complexity. Forecast errors are computed as the time series mean of the monthly mean squared errors (MSEs) within each validation sample (MSE_VALI) and test sample (MSE_TEST), respectively. The relative MSE difference (between random forests and 1-year rolling betas (OLS_1Y_D, introduced in Section IV.A)) is the percentage difference in test-sample average MSEs during each calendar year (MSE_TEST_PD), calculated as 1 minus the MSE of the random forests model divided by the MSE of the benchmark model. In particular, this figure plots MC over time, along with MSE_VALI (Graph A), MSE_TEST (Graph B), and MSE_TEST_PD (Graph C), respectively. The black and orange unfilled circles are assigned to the re-estimation dates, that is, the dates at which the forecasts of the stock-level betas (over the next year) are obtained. The graphs also visualize the NBER recession periods (gray-shaded areas). The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the sample period from Mar. 1970 to Dec. 2020, while the first beta estimates are obtained in Dec. 1979.



absolute variable importance as the increase in the value-weighted MSE from setting all values of a given predictor to its uninformative median within the training sample. Second, we normalize the absolute variable importance measures to sum to 1, indicating the relative contribution of each variable to the RF model.

Graph A of Figure 5 plots time series averages of the relative variable importance measures for the predictor categories introduced in Section III.²⁷ Historical betas are the most informative variables, together accounting for more than 60% of the variable importance. This is as expected, since realized betas are highly persistent and have long-memory properties. Technical indicators and accounting-based predictors are also important, while macroeconomic indicators seem to have only little impact on the betas. Therefore, time variation in market betas is driven more by changes in the firm fundamentals than by changes in the underlying economic conditions.²⁸

Graph B of Figure 5 presents the time series averages of the relative variable importance measures for the 10 most influential predictors. Random forests place most of their weights on only five variables, leaving 29 variables of significantly lower importance (ignoring industry classifications). Consistent with the above, the most influential predictors are the 3 sample estimates of beta, with the largest weight placed on the rolling betas using a 1-year window of daily returns (OLS_1Y_D), followed by those using a 3-month window of daily returns (OLS_3M_D) and a 5-year window of monthly returns (OLS_5Y_M), respectively. In addition, a firm's turnover (TO) and size (ME) are also important. Overall, the average relative contribution of the top five variables to the RF model is 82.99%.²⁹

Interestingly, the variables that are most important in predicting future market betas are different from those that are most helpful in predicting returns, as documented by Gu et al. (2020). For example, both market betas and turnover are not among the top 10 return predictors in their analysis. Only firm size appears in the list of the top predictors for both market betas and returns. In contrast, several of the major return predictors, such as short-term reversal, value, and momentum, have little relevance in predicting future market beta.

Because the overall relative variable importance measures reflect only the average contribution of a predictor to the predictive performance of the random forests, we also examine the relative variable importance metrics over time. Volatile metrics indicate that all covariates in the predictor set are essential; stable numbers would imply that we can permanently remove uninformative predictors, as they may reduce the signal-to-noise ratio of the RF model. Therefore, we next focus on the 29 least important predictors for which removal can be considered.

To identify the time variability in the relative importance measures only within this subset of predictors, we drop the remaining variables before normalizing the absolute variable importance measures to sum to 1 at each reestimation date. Graph C of Figure 5 presents the resulting relative variable importance over the sample period. Although we still observe substantial differences, the graph shows that the

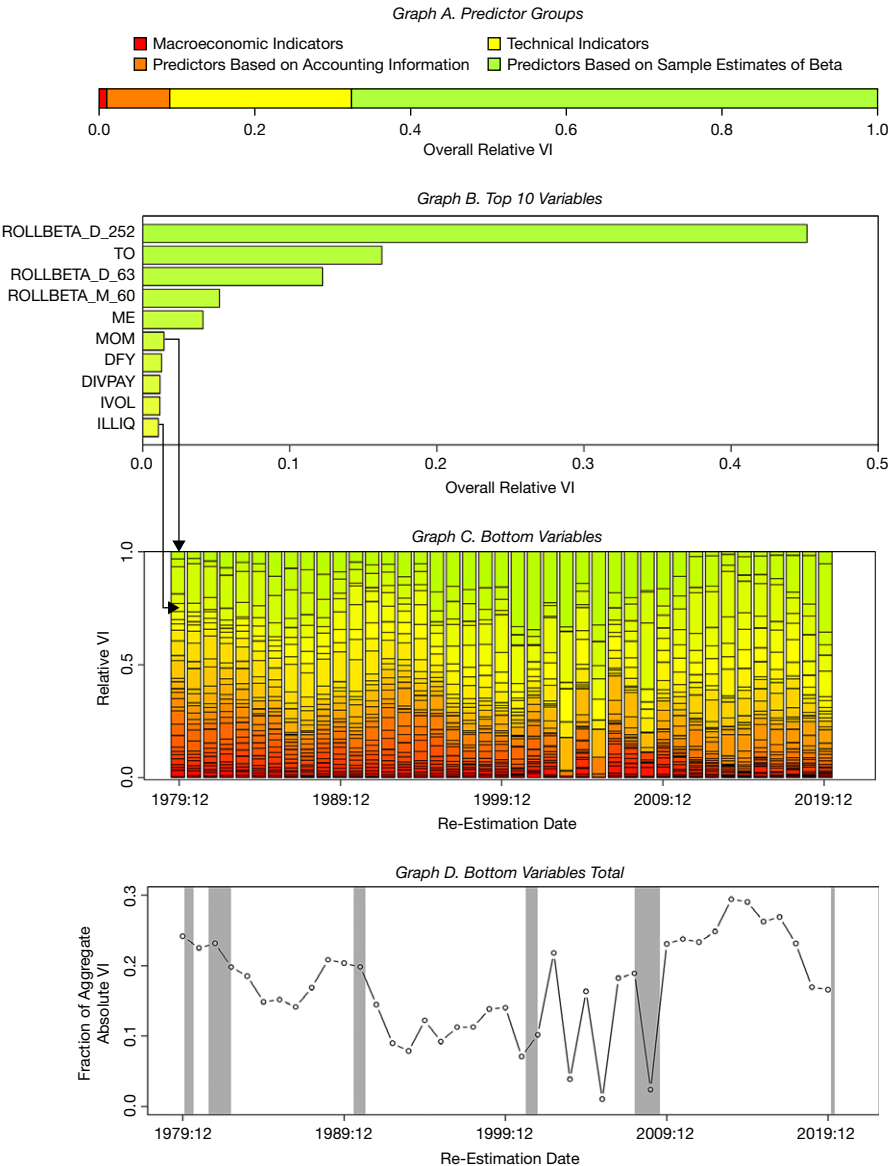
²⁷We simultaneously set all values of all predictors within each category to their uninformative median values within the training sample before computing the absolute and relative variable importance metrics as described in the text.

²⁸As shown in Section C of the Supplementary Material (see Table C3 in the Supplementary Material), our results are robust when we include a set of other macroeconomic variables in addition to the default spread.

²⁹Using a smaller and different set of predictor variables, Jorovski et al. (2020) also find that a lagged beta is the most influential variable. However, they do not include historical betas with different horizons, which turns out to be important in our empirical analysis.

FIGURE 5
Relative Variable Importance in Aggregate and over Time

Figure 5 shows the relative importance of the variables included as predictors in the random forests (RF, introduced in Section IV.B). For this purpose, the variable importance matrix is computed based on a 2-step approach, separately for each re-estimation date: First, the absolute variable importance is computed as the increase in value-weighted mean squared error (MSE) from setting all values of a given predictor to its uninformative median within the training sample. Second, the absolute variable importance measures are normalized to sum to 1, indicating the relative contribution of each variable to the RF model. Graphs A and B show the time series averages of the relative variable importance measures for the predictor categories introduced in Section III and the 10 most influential predictors, respectively. Focusing on the 29 least important predictors, Graph C presents the resulting relative variable importance metrics over the sample period. For this purpose, the remaining variables are omitted before the absolute variable importance measures are normalized to sum to 1 at each re-estimation date. Graph D visualizes the fraction of the aggregate absolute variable importance (i.e., the sum of the increases in value-weighted MSE across all variables) attributable to the 29 least important predictors over the sample period, along with the NBER recession periods (gray-shaded areas). The sample includes all firms that were or are listed on the NYSE, AMEX, or NASDAQ in any month during the period Mar. 1970–Dec. 2020, while the first beta estimates are obtained in Dec. 1979.



relative variable importance metrics change significantly over time. Therefore, we conclude that each predictor variable is an important contributor to the random forests (albeit to time-varying degrees).

In addition, we evaluate the overall contribution of the 29 least important covariates to the predictive performance of the RF model. To do so, at each re-estimation date, we compute the fraction of the total absolute variable importance attributable to this subset of predictors. Graph D of Figure 5 plots this fraction over the sample period, along with the NBER recession periods (gray-shaded areas). At times, the aggregate contribution is exceptionally low, such as during the global financial crisis (Dec. 2008). In contrast, it is high in the subsequent period, reaching almost 30% at its peak. As a result, even the unconditionally unimportant covariates can conditionally play an important role in the performance of the RF model. These results justify the use of our comprehensive set of predictors. While using information from multiple different sources seems to be essential for the predictive performance of machine learning-based prediction models, it also makes them prone to potential misspecification. Therefore, we only include those covariates that have been shown theoretically or empirically to explain and predict time-varying market betas. This partially explains our finding that we should *not* remove certain predictors: We refrain from including too many predictors that are likely irrelevant in the first place.

VIII. Conclusion

Using a large universe of U.S. stocks and a long sample period, we compare the predictive performance of machine learning-based beta estimators (linear regression, tree-based models, and neural networks) with that of established benchmarks. Machine learning techniques outperform established approaches both statistically and economically. Random forests perform best, but GBRT and neural networks also work well. In particular, machine learning methods produce the lowest forecast and hedging errors. In addition, they outperform the benchmark models in hedging anomaly portfolios and constructing MVPs.

An important economic reason for the outperformance of machine learning methods is their ability to capture the joint information content of a large set of firm characteristics that appear to affect betas. However, random forests, GBRT, and neural networks outperform linear regression, while incorporating the same covariates. We conclude that much of this outperformance is further attributable to the machine learning methods' ability to exploit nonlinear and interactive patterns.

Supplementary material

To view supplementary material for this article, please visit <http://doi.org/10.1017/S0022109024000036>.

References

- Amihud, Y., and H. Mendelson. "The Liquidity Route to a Lower Cost of Capital." *Journal of Applied Corporate Finance*, 12 (2000), 8–25.

- Andersen, T. G.; T. Bollerslev; F. X. Diebold; and G. Wu. "Realized Beta: Persistence and Predictability." In *Advances in Econometrics: Econometric Analysis of Economic and Financial Time Series*, T. Fomby, and D. Terrel, eds. Amsterdam, The Netherlands: Elsevier (2006), 1–39.
- Ang, A., and J. Chen. "CAPM Over the Long Run: 1926–2001." *Journal of Empirical Finance*, 14 (2007), 1–40.
- Bailey, D. H.; J. M. Borwein; M. L. de Prado; and Q. J. Zhu. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the American Mathematical Society*, 61 (2014), 458–471.
- Bailey, D. H.; J. M. Borwein; M. L. de Prado; and Q. J. Zhu. "The Probability of Backtest Overfitting." *Journal of Computational Finance*, 20 (2017), 1460–1559.
- Bali, T. G.; A. Goyal; D. Huang; F. Jiang; and Q. Wen. "The Cross-Sectional Pricing of Corporate Bonds Using Big Data and Machine Learning." Working Paper, Georgetown University (2022).
- Barber, B. M.; X. Huang; and T. Odean. "Which Factors Matter to Investors? Evidence from Mutual Fund Flows." *Review of Financial Studies*, 29 (2016), 2600–2642.
- Beaver, W.; P. Kettler; and M. Scholes. "The Association Between Market Determined and Accounting Determined Risk Measures." *Accounting Review*, 45 (1970), 654–682.
- Becker, J.; F. Hollstein; M. Prokopczuk; and P. Sibbertsen. "The Memory of Beta." *Journal of Banking and Finance*, 124 (2021), 106026.
- Berk, J. B., and J. H. van Binsbergen. "Assessing Asset Pricing Models Using Revealed Preference." *Journal of Financial Economics*, 119 (2016), 1–23.
- Bianchi, D.; M. Büchner; and A. Tamoni. "Bond Risk Premiums with Machine Learning." *Review of Financial Studies*, 34 (2021), 1046–1089.
- Black, F.; M. C. Jensen; and M. S. Scholes. "The Capital Asset Pricing Model: Some Tests." In *Studies in the Theory of Capital Markets*, M. C. Jensen, ed. New York, NY: Praeger (1972), 79–121.
- Bollerslev, T.; R. F. Engle; and J. M. Wooldridge. "A Capital Asset Pricing Model with Time-Varying Covariances." *Journal of Political Economy*, 96 (1988), 116–131.
- Bucci, A. "Realized Volatility Forecasting with Neural Networks." *Journal of Financial Econometrics*, 18 (2020), 502–531.
- Campbell, J. Y.; M. Lettau; B. G. Malkiel; and Y. Xu. "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk." *Journal of Finance*, 56 (2001), 1–43.
- Chan, L. K.; J. Karceski; and J. Lakonishok. "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model." *Review of Financial Studies*, 12 (1999), 937–974.
- Chincarini, L. B.; D. Kim; and F. Moneta. "Beta and Firm Age." *Journal of Empirical Finance*, 58 (2020), 50–74.
- Christensen, K.; M. Siggaard; and B. Veliyev. "A Machine Learning Approach to Volatility Forecasting." *Journal of Financial Econometrics*, 21 (2023), 1680–1727.
- Cochrane, J. H. "Presidential Address: Discount Rates." *Journal of Finance*, 66 (2011), 1047–1108.
- Connor, G.; M. Hagmann; and O. Linton. "Efficient Semiparametric Estimation of the Fama–French Model and Extensions." *Econometrica*, 80 (2012), 713–754.
- Connor, G., and O. Linton. "Semiparametric Estimation of a Characteristic-Based Factor Model of Common Stock Returns." *Journal of Empirical Finance*, 14 (2007), 694–717.
- Cosemans, M.; R. Frehen; P. C. Schotman; and R. Bauer. "Estimating Market Betas Using Prior Information Based on Firm Fundamentals." *Review of Financial Studies*, 29 (2016), 1072–1112.
- Daniel, K., and T. J. Moskowitz. "Momentum Crashes." *Journal of Financial Economics*, 122 (2016), 221–247.
- Daniel, K.; L. Mota; S. Rottke; and T. Santos. "The Cross-Section of Risk and Returns." *Review of Financial Studies*, 33 (2020), 1927–1979.
- Diebold, F., and R. Mariano. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics*, 13 (1995), 253–263.
- Donaldson, R. G., and M. Kamstra. "An Artificial Neural Network-GARCH Model for International Stock Return Volatility." *Journal of Empirical Finance*, 4 (1997), 17–46.
- Drobetz, W., and T. Otto. "Empirical Asset Pricing via Machine Learning: Evidence from the European Stock Market." *Journal of Asset Management*, 22 (2021), 507–538.
- Fama, E. F., and K. R. French. "The Cross-Section of Expected Stock Returns." *Journal of Finance*, 47 (1992), 427–465.
- Fama, E. F., and K. R. French. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33 (1993), 3–56.
- Fama, E. F., and K. R. French. "Industry Costs of Equity." *Journal of Financial Economics*, 43 (1997), 153–193.
- Fama, E. F., and K. R. French. "Dissecting Anomalies." *Journal of Finance*, 63 (2008), 1653–1678.
- Fama, E. F., and K. R. French. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics*, 116 (2015), 1–22.

- Fama, E. F., and J. D. MacBeth. "Risk, Return, and Equilibrium: Tests." *Journal of Political Economy*, 81 (1973), 607–636.
- Fan, J.; Y. Liao; and W. Wang. "Projected Principal Component Analysis in Factor Models." *Annals of Statistics*, 44 (2016), 219–254.
- Fernandes, M.; M. C. Medeiros; and M. Scharh. "Modeling and Predicting the CBOE Market Volatility Index." *Journal of Banking and Finance*, 40 (2014), 1–10.
- Ferson, W. E., and C. R. Harvey. "Conditioning Variables and the Cross Section of Stock Returns." *Journal of Finance*, 54 (1999), 1325–1360.
- Frazzini, A., and L. H. Pedersen. "Betting Against Beta." *Journal of Financial Economics*, 111 (2014), 1–25.
- Freyberger, J.; B. Höppner; A. Neuhierl; and M. Weber. "Missing Data in Asset Pricing Panels." *Review of Financial Studies*, forthcoming (2024).
- Freyberger, J.; A. Neuhierl; and M. Weber. "Dissecting Characteristics Nonparametrically." *Review of Financial Studies*, 33 (2020), 2326–2377.
- Ghysels, E., and E. Jacquier. "Market Beta Dynamics and Portfolio Efficiency." Working Paper, Boston University (2006).
- Giacomini, R., and H. White. "Tests of Conditional Predictive Ability." *Econometrica*, 74 (2006), 1545–1578.
- Gilbert, T.; C. Hrdlicka; J. Kalodimos; and S. Siegel. "Daily Data is Bad for Beta: Opacity and Frequency-Dependent Betas." *Review of Asset Pricing Studies*, 4 (2014), 78–117.
- Graham, J. R. "Presidential Address: Corporate Finance and Reality." *Journal of Finance*, 77 (2022), 1975–2049.
- Graham, J. R., and C. R. Harvey. "The Theory and Practice of Corporate Finance: Evidence from the Field." *Journal of Financial Economics*, 60 (2001), 187–243.
- Grundy, B. D., and J. S. M. Martin. "Understanding the Nature of the Risks and the Source of the Rewards to Momentum Investing." *Review of Financial Studies*, 14 (2001), 29–78.
- Gu, S.; B. Kelly; and D. Xiu. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies*, 33 (2020), 2223–2273.
- Gu, S.; B. Kelly; and D. Xiu. "Autoencoder Asset Pricing Models." *Journal of Econometrics*, 222 (2021), 429–450.
- Gulen, H.; Y. Xing; and L. Zhang. "Value Versus Growth: Time-Varying Expected Stock Returns." *Financial Management*, 40 (2011), 381–407.
- Hansen, P. R., and A. Lunde. "Consistent Ranking of Volatility Models." *Journal of Econometrics*, 131 (2006), 97–121.
- Hansen, P. R.; A. Lunde; and J. M. Nason. "The Model Confidence Set." *Econometrica*, 79 (2011), 453–497.
- Harvey, C. R., and Y. Liu. "Evaluating Trading Strategies." *Journal of Portfolio Management*, 40 (2014), 108–118.
- Harvey, C. R., and Y. Liu. "Backtesting." *Journal of Portfolio Management*, 42 (2015), 13–28.
- Harvey, C. R.; Y. Liu; and H. Zhu. "... and the Cross-Section of Expected Returns." *Review of Financial Studies*, 29 (2016), 5–68.
- Hillebrand, E., and M. C. Medeiros. "The Benefits of Bagging for Forecast Models of Realized Volatility." *Econometric Reviews*, 29 (2010), 571–593.
- Hollstein, F., and M. Prokopczuk. "Estimating Beta." *Journal of Financial and Quantitative Analysis*, 51 (2016), 1437–1466.
- Hollstein, F.; M. Prokopczuk; and C. Wese Simen. "Estimating Beta: Forecast Adjustments and the Impact of Stock Characteristics for a Broad Cross-Section." *Journal of Financial Markets*, 44 (2019), 91–118.
- Hou, K.; C. Xue; and L. Zhang. "Replicating Anomalies." *Review of Financial Studies*, 33 (2020), 2019–2133.
- Jacobs, M. T., and A. Shivdasani. "Do You Know Your Cost of Capital?" *Harvard Business Review*, 90 (2012), 118–124.
- Jacoby, G.; D. J. Fowler; and A. A. Gottesman. "The Capital Asset Pricing Model and the Liquidity Effect: A Theoretical Approach." *Journal of Financial Markets*, 3 (2000), 69–81.
- Jagannathan, R., and Z. Wang. "The Conditional CAPM and the Cross-Section of Expected Returns." *Journal of Finance*, 51 (1996), 3–53.
- Jourovski, A.; V. Dubikovskyy; P. Adell; R. Ramakrishnan; and R. Kosowski. "Forecasting Beta Using Machine Learning and Equity Sentiment Variables." In *Machine Learning for Asset Management: New Developments and Financial Applications*, E. Jurczenko, ed. London, UK: Wiley-ISTE (2020), 231–260.
- Karolyi, G. A. "Predicting Risk: Some New Generalizations." *Management Science*, 38 (1992), 57–74.

- Kelly, B.; T. J. Moskowitz; and S. Pruitt. "Understanding Momentum and Reversal." *Journal of Financial Economics*, 140 (2021), 726–743.
- Kelly, B. T.; S. Pruitt; and Y. Su. "Characteristics are Covariances: A Unified Model of Risk and Return." *Journal of Financial Economics*, 134 (2019), 501–524.
- Kim, S.; R. A. Korajczyk; and A. Neuhierl. "Arbitrage Portfolios." *Review of Financial Studies*, 34 (2020), 2813–2856.
- Kozak, S.; S. Nagel; and S. Santosh. "Shrinking the Cross-Section." *Journal of Financial Economics*, 135 (2020), 271–292.
- Leippold, M.; Q. Wang; and W. Zhou. "Machine Learning in the Chinese Stock Market." *Journal of Financial Economics*, 145 (2021), 64–82.
- Lewellen, J. "The Cross-Section of Expected Stock Returns." *Critical Finance Review*, 4 (2015), 1–44.
- Lintner, J. "Security Prices, Risk, and Maximal Gains from Diversification." *Journal of Finance*, 20 (1965), 587–615.
- Lo, A. W., and A. C. MacKinlay. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *Review of Financial Studies*, 3 (1990), 431–467.
- Luong, C., and N. Dokuchaev. "Forecasting of Realised Volatility with the Random Forests Algorithm." *Journal of Risk and Financial Management*, 11 (2018), 61.
- Mincer, J. A., and V. Zarnowitz. "The Evaluation of Economic Forecasts." In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, J. A. Mincer, ed. Cambridge, MA: NBER (1969), 3–46.
- Mittnik, S.; N. Robinsonov; and M. Spindler. "Stock Market Volatility: Identifying Major Drivers and the Nature of Their Impact." *Journal of Banking and Finance*, 58 (2015), 1–14.
- Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica*, 34 (1966), 768–783.
- Newey, W. K., and K. D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55 (1987), 703–708.
- Novy-Marx, R. "Operating Leverage." *Review of Finance*, 15 (2011), 103–134.
- Novy-Marx, R., and M. Velikov. "Betting Against Betting Against Beta." *Journal of Financial Economics*, 143 (2022), 80–106.
- Patton, A. J. "Volatility Forecast Comparison Using Imperfect Volatility Proxies." *Journal of Econometrics*, 160 (2011), 246–256.
- Petkova, R., and L. Zhang. "Is Value Riskier Than Growth?" *Journal of Financial Economics*, 78 (2005), 187–202.
- Rahimikia, E., and S. H. Poon. "Machine Learning for Realised Volatility Forecasting." Working Paper, University of Manchester (2020).
- Schorfheide, F., and K. I. Wolpin. "On the Use of Holdout Samples for Model Selection." *American Economic Review*, 102 (2012), 477–481.
- Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *Journal of Finance*, 19 (1964), 425–442.
- Vasicek, O. "A Note on Using Cross-Sectional Information in Bayesian Estimation of Market Betas." *Journal of Finance*, 28 (1973), 1233–1239.
- Van Binsbergen, J. H.; X. Han; and A. Lopez-Lira. "Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases." *Review of Financial Studies*, 36 (2023), 2361–2396.
- Welch, I. "Simply Better Market Betas." *Critical Finance Review*, 11 (2022), 37–64.
- West, K. D. "Forecast Evaluation." In *Handbook of Economic Forecasting*, G. Elliott, C. Granger, and A. Timmermann, eds. Amsterdam, The Netherlands: Elsevier (2006), 99–134.