Expectations Matter: When (not) to Use Machine Learning Earnings Forecasts*

John L. Campbell[†], Harrison Ham[‡], Zhongjin (Gene) Lu[§], and Katherine Wood[¶]

Abstract

We comprehensively examine whether machine learning technology can meaningfully improve earnings forecasts, and if so, whether market expectations appear to reflect those superior forecasts. First, we use a consistent methodology to evaluate a comprehensive list of machine-learning forecasts from 1990 to 2020. We find evidence that the best machine learning forecast outperforms analysts' forecasts, but the improvement declines over time and is small when analysts face stronger incentives to be accurate. Second, using earnings response coefficient (ERC) tests, we infer that investors' expectations put more weight on analysts' forecasts than prescribed by the best machine forecast. Investors' overweighting becomes statistically insignificant among large-cap firms with more sophisticated investors. Third, our time-series analyses suggest that analyst and machine forecasts are converging over time and that analysts' information production remains critical, blurring the line between human and machine forecasts. Overall, our study provides an updated and comprehensive take on the most accurate earnings forecast and the best proxy for investors' earnings expectations.

*We thank an anonymous reviewer, Karthik Balakrishnan, Supdita Basu, Matt Bloomfield, Mark Bradshaw, Sean Cao, Mike Chen, Braiden Coleman, Joe Comprix, Sandip Dhole, Paul Fischer, Mukesh Garg, Rebecca Hann, Jie He, Frank Heflin, Steven Malliaris, Michael Kimbrough, K. Ramesh, Willie Reddic, Sandra Schafhaütle, Karen Ton, Ben Whipple, Mark Zakota, and seminar participants at the Chinese University of Hong Kong, the University of Georgia, the University of Maryland, Monash University, Nanyang Technological University, Rice University, Singapore Management University, Syracuse University, and the Villanova University Fall Accounting Conference for helpful comments and suggestions.

†Professor of Accounting and Finance, Terry College of Business, University of Georgia, johnc@uga.edu

‡PhD candidate, Terry College of Business, University of Georgia, harrison.ham@uga.edu

§Assistant Professor of Finance, Terry College of Business, University of Georgia; corresponding author, author email: zlu15@uga.edu

¶Assistant Professor of Finance, Bentley University, kwood@bentley.edu

1. Introduction

Earnings forecasts play an important role in numerous lines of accounting and finance research. Given the exponential growth in data production and computing power over the last two decades, a natural question to examine is how advancements in machine learning technology have impacted the ability to forecast a firm's earnings. In this paper, we examine two closely related questions. First, to what extent do machine learning (ML) earnings forecasts outperform analyst earnings forecasts in terms of statistical accuracy (i.e., minimizing ex-post forecast errors)? Second, do market earnings expectations align more with the most statistically optimal machine forecast or with analysts' forecasts?

One key obstacle to answering these questions is the proliferation of specification choices in machine learning forecasts, which leads to thousands of potential ML forecasts. As a result, the existing literature has no consensus on whether ML forecasts are superior to analysts nor the extent of ML improvement, if any. Given the current state of the literature, skeptics not only raise specification mining concerns but also point to recent studies documenting the statistical superiority of analysts' forecasts over ML forecasts.[1] Furthermore, there remains the practical implementation issue of determining the appropriate benchmark from thousands of potential machine learning forecasts.

We overcome these challenges using a unifying empirical approach and systematically evaluate the statistical accuracy of earnings forecasts proposed in the recent ML literature, as well as their correlations with investors' earnings expectations. We offer the following new findings. First, specification choices have a large impact on model performance, but only a few choices matter and can be chosen judiciously. Second, the statistically optimal ML forecast consistently outperforms analysts' forecasts, but the magnitude of the outperformance exhibits substantial variation across size, forecast horizon, and over time, thus cautioning against any simplistic broad-brush characterization of machine or analysts' superiority. Finally, using machine learning, analysts', and crowdsourced forecasts to examine market expectations, we find that investors place excessive weight on analysts'

[1] As our Table 1 shows, most studies report that ML models outperform analysts' forecasts (Ball and Ghysels, 2018; Cao and You, 2020; van Binsbergen et al., 2022; Uddin et al., 2022). However, (de Silva and Thesmar, 2022) report that analysts' forecasts are superior for 1 quarter to 1 year ahead earnings, whereas ML forecasts are superior for 2-4 years ahead earnings.

forecasts compared to a statistically optimal forecast, and this overweighting exhibits systematic variation.

We start our analysis by comprehensively reviewing common specification choices used in the recent machine learning earnings forecasting literature. To assess the impact of these specification choices, we first compile a comprehensive list of 3,024 machine learning models that represent the range of these specification choices (see Table 1). We then use a consistent measure of statistical forecast accuracy to evaluate these ML forecasts relative to analysts' forecasts. Motivated by Bradshaw et al. (2012), we use ML Superiority, which is defined as the absolute value of the error in the analysts' forecast minus the absolute value of the error in a given machine learning forecast.

We find significant variation in the accuracy of machine forecasts due to specification choice. Nearly 90 percent of the 3,024 machine learning models evaluated underperform analysts' forecasts. Thus, understanding the impact of specification choices is key to making sense of prior papers' results on machine or analysts' superiority. Our results show that utilizing a non-linear algorithm, combined with an outlier-resistant loss function and a temporal train-validation split, significantly improves forecast accuracy. This approach, particularly when focused on refining analysts' forecasts by correcting their predictable biases, consistently yields the most statistically accurate earnings forecast.[2]

Having determined the most accurate machine learning specification, we then proceed to examine our first research question regarding the statistical accuracy of machine forecasts relative to analysts' forecasts (as measured by ML Superiority). Over our sample of forecasts from 1990 to 2020, we find consistent evidence that the statistically optimal ML forecast outperforms analysts' forecasts. However, in situations where analysts are expected to exert more effort and face stronger incentives to be accurate (i.e., short-horizon forecasts and large-cap stocks), the ML improvement is an order of magnitude smaller compared to that for long-horizon forecasts and among small-cap firms. For example, for quarterly EPS forecasts made in the month prior to the announcement, the

[2]Furthermore, when we divide the sample into decade-long subsamples, we find the top performing models consistently use these four key specification choices. Sections 2 through 4 provide more details behind this methodological exercise. Furthermore, our website contains programming code and statistical estimates to the extent other researchers wish to use these statistically optimal machine learning forecasts. (ReadMe File and Machine Learning Forecasts)

statistically optimal ML forecast reduces the average absolute error by a modest 0.22 cents per $10 stock price (or 0.86 cents on an annualized basis), versus 5.18 cents per $10 stock price for annual EPS forecasts made 12 months prior to the announcement. The accuracy gain among all quarterly EPS forecasts is further reduced to 0.16 (or 0.62 cents on an annualized basis) cents per $10 stock price among large-cap firms.

The statistically optimal ML forecast also outperforms an alternative human forecast crowd-sourced from the social media website Estimize. While Estimize forecasts demonstrate a lower level of pessimistic bias than I/B/E/S analysts' forecasts, Estimize forecasts are actually associated with higher mean absolute errors than analysts. In contrast, the statistically optimal ML forecast presents both lower biases and smaller mean absolute errors when compared to both Estimize and analysts' forecasts.

Finally, our time-series analysis indicates a diminishing outperformance of the statistically optimal ML forecast relative to analysts over time, suggesting that analysts incorporate machine learning techniques to enhance their forecasts. This recognition is conceptually important because if analysts are integrating machine learning forecasts into their analyses, and vice versa, it blurs the distinction between human and machine forecasts as both forecasts result from collaborations between humans and machines. Therefore, despite being limited to commonly used algorithms in the existing machine learning earnings forecast literature, to the extent that advancements in forecasting algorithms benefit both human and machine forecasts without significantly widening their performance gap, our findings on the relative accuracy of the best ML forecasts versus analysts' forecasts may retain enduring relevance.

We then turn to examine our research question regarding the extent to which investors' earnings expectations align with the statistically optimal ML forecast or with analysts' forecasts. Under the rational expectation hypothesis, investors' expectations should align perfectly with the statistically optimal forecast. However, if investors share the same biases as analysts, as suggested by (Bordalo et al., 2019), investors' earnings expectations may correlate more closely with analysts' forecasts. We test these competing hypotheses in two approaches. First, assuming that earnings announcement returns are driven by revisions in market earnings expectations, we use earnings response coefficients (ERC) to infer which earnings forecast is a more precise proxy for market-expected earnings. Second, if investors share the same biases as analysts, then ex-ante measures of analysts' bias should predict abnormal stock returns.

Our ERC regression analysis indicates that both hypotheses have merit under different circumstances. Using forecasts from our 1,512 ML models for FQ earnings in univariate ERC regressions, we find that those forecasts which are statistically more accurate tend to have larger ERCs. This indicates that market earnings expectations are rational, aligning more closely with these accurate forecasts. However, our bivariate ERC regressions also reveal that compared to the optimal weighting of analysts' forecasts as prescribed by the ML forecast, investors disproportionately favor analysts' forecasts when forming earnings expectations. Furthermore, this overweighting is inversely related to firm size and institutional ownership, becoming statistically insignificant among large-cap stocks with above-median institutional ownership. This result suggests that in areas where prices are expected to be more efficient, investors weight analysts' forecasts more in line with the statistically optimal ML forecast. Finally, while crowdsourced estimates offer no additional value beyond what is already captured by statistically optimal forecasts in reflecting market expectations, the initiation of Estimize coverage is associated with a reduction in investors' overweighting of analysts' forecasts.

Given that our ERC regression results indicate that investors' overweighting of analysts' forecasts varies systematically with size and institutional ownership, we further examine the return predictive power of the ex-ante conditional bias in analysts' forecasts in small- cap/large-cap and above-/below-median institutional ownership groups. We use the analysts' bias identified by the statistically optimal ML forecast (i.e., ML – AF) to predict monthly stock returns. We find that analysts' bias has statistically significant predictive power for stock returns among small firms or firms with low institutional ownership. However, this is not the case among large-cap stocks with high institutional ownership, supporting the notion that market expectations align more closely with the statistically optimal ML forecast among these stocks.

The main contribution of this study is to provide an updated and comprehensive take on the (statistically) most accurate earnings forecast and the best proxy for investors' earnings expectations. We contribute to the academic community by providing the code and estimates for statistically optimal ML forecasts. These optimal ML forecasts are crucial for small-cap firms and longer forecasting horizons for applications requiring precise earnings predictions to ascertain a firm's fundamental value. Moreover, for large-cap stocks with substantial institutional ownership, the optimal ML forecast serves as the preferred proxy for market earnings expectations, whereas in other segments of the market investors' expectations appear to be more of a blend between the optimal ML forecast and analysts' forecasts.

Our findings also contribute to the machine learning earnings forecast literature by using a consistent methodology to evaluate a comprehensive list of ML forecasts.[3] Our study helps characterize the impact of specification choices in ML model performance. As a result of our exhaustive approach examining over 3,000 model specifications, we provide much more confidence in the results reported in contemporaneous studies such as the importance of combining human private information with machine learning van Binsbergen et al. (2022); Cao et al. (2021), as well as the fact that machines produce better forecasts for smaller firms and longer forecast horizons (e.g., de Silva and Thesmar (2022)). In contrast to the impression made by prior studies, we show that the accuracy improvement of machine forecasts over analysts' forecasts has declined quite substantially to a modest level towards the end of our sample period. Furthermore, the accuracy improvement of machine forecasts is small for near-term EPS forecasts and for large-cap stocks which account for 87% of the market capitalization.

Our findings further the understanding of investors' earnings expectations. Prior studies in this area largely assume that analysts' earnings forecasts are the best proxy for market expectations Ball and Brown (1968); Bernard and Thomas (1989, 1990); Kothari (2001); Bradshaw et al. (2012); with some arguing that analysts and investors might have different objective functions (e.g., Gu and Wu (2003); Basu and Markov (2004); Weiss et al. (2008)) and others arguing that investors might exhibit the same cognitive biases as analysts (e.g., Bertomeu et al. (2021)). We demonstrate that neither analysts' forecasts nor the statistically optimal forecast are a perfect proxy for market earnings expectations, but combining these two forecasts is extremely useful in characterizing the

[3]Table 1 presents a summary of papers in this area, showing differences in methodology and ultimately conclusions as to whether machine or analysts' forecasts appear to dominate.

market earnings expectations. We present new evidence showing that investors' earnings expectations deviate from statistically optimal forecasts by disproportionately overweighting analysts' forecasts. This overweighting decreases in institutional ownership and firm size and market expectations appear to converge to the optimal earnings forecasts for large-cap stocks with above-median institutional ownership.

Finally, our results are related to the emerging social media research that shows crowdsourced earnings forecasts are less biased and help investors unravel analysts' biases. We demonstrate that there are limits to the ability of crowdsourcing to unravel analysts' bias. Specifically, the optimal ML forecast exhibits less bias compared to crowdsourced earnings forecast, surpasses them in accuracy, and covers more firms. Nevertheless, corroborating the finding in Jame et al. (2016); Schafhautle and Veenman (2024), our findings show that following the initiation of Estimize coverage, investors' expectations shift closer to the statistically optimal forecast and further away from analysts' forecasts.

2. Experiment Design

2.1. ML Model Specification Choices

When forecasting earnings, researchers must first determine the dataset (y, X), where y is the target variable and X is the predictor set. Given the dataset (y, X), researchers fit a linear or non-linear regression model $(f(X))$ in order to minimize the sum of a loss function (L) and a potential regularization term (G) with hyperparameter(s) γ. The model takes the following form:[4]

$$f(X) = \arg\min_{\{\beta\}} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(y_i, \beta | X_i) + G(y_i, \beta | X_i, \gamma) \right\}, \text{for } i \text{ in the estimation window.} \quad (1)$$

There are 6 specification choices directly related to model training: (1) the loss function, (2) the ML algorithm (which determines the form of the regularization term), (3) the cross-validation scheme of parameter tuning, (4) the frequency of hyperparameter re-tuning, (5) the frequency of model re-fitting, and (6) the estimation window.

In Table 1, we comprehensively review the variations in specification choices made by exist-

[4]Because this paper requires readers to understand machine learning and involves substantial nuance in specification choices within ML, we take the unusual step of including an Internet Appendix with a first round journal submission. Internet Appendix Section 1 provides the OLS and LASSO models as two examples of the loss function and regularization term.

ing ML earnings forecast studies.[5] We identify six commonly used ML algorithms: OLS, Lasso, Ridge, Elastic Net (EN), Random Forest (RF), and Gradient Boosted Regression Trees (GBRT), each implemented with a diverse set of specification choices. From these studies, we also identify variation in four potentially important choices not directly related to model training. We detail these specification choices in Table 2 and offer an in-depth discussion regarding these choices in Section 2 of the Internet Appendix.

From the full combination of the choices listed in Table 2, we derive 3,024 potential ML models. We evaluate the forecasting performance of this exhaustive list of ML models to assess the impact of ML specification choices.

2.2. Model Performance Evaluation Metric

To enable a direct comparison to prior literature, we use the superiority measure from Bradshaw et al. (2012) to evaluate out-of-sample forecasting accuracy. Specifically, we define ML Superiority for the forecast made in month t, for firm i and earnings with fiscal period end T, as follows:[6]

$$\begin{aligned}\text{ML superiority} \equiv &\left|\frac{\text{EPS}_{i,T}}{\text{Price}_{i,t}} - \frac{\text{Analysts Forecast}_t\left(\text{EPS}_{i,T}\right)}{\text{Price}_{i,t}}\right| \\ &- \left|\frac{\text{EPS}_{i,T}}{\text{Price}_{i,t}} - \frac{\text{ML Model Forecast}_t\left(\text{EPS}_{i,T}\right)}{\text{Price}_{i,t}}\right|\end{aligned} \tag{2}$$

We use "ML Superiority" to refer to the average ML Superiority of an ML model (over firm-month observations unless otherwise stated). A more positive ML Superiority means ML forecasts are more accurate than analysts' forecasts. Because the measure is based on the absolute value of EPS forecast errors per dollar stock price, it is economically intuitive and allows comparison across firms. We follow Bradshaw et al. (2012) and winsorize analysts' error and ML error (scaled by price) to the range [-1, 1] throughout our analysis. We drop observations with stock price less than or equal to $1 to minimize the impact of extreme values, but our results are robust to removing this filter. In all following analyses, the ML Superiority measure is annualized, meaning that all variables in Eq. (2) are multiplied by four for next quarter (i.e. FQ) earnings.

[5] Our review focuses on studies that compare the forecast accuracy of ML forecasts versus analysts' forecasts. A related but separate strand of literature studies the statistical earnings forecasts for firms not covered by analysts, such as Hou et al. (2012) and Chattopadhyay et al. (2023).

[6] When EPS is scaled by price in the forecasting step, ML Superiority is defined as the difference between $\left|\frac{\text{EPS}_{i,T}}{\text{Price}_{i,t}} - \frac{\text{Analysts Forecast}_t\left(\text{EPS}_{i,T}\right)}{\text{Price}_{i,t}}\right|$ and $\left|\frac{\text{EPS}_{i,T}}{\text{Price}_{i,t}} - \text{ML Model Forecast}_t\left(\frac{\text{EPS}_{i,T}}{\text{Price}_{i,t}}\right)\right|$.

3. Data and Sample Construction

Our dataset consists of the intersection of firms in CRSP, Compustat, and I/B/E/S. Analysts' EPS forecasts (AF) in this paper refer to I/B/E/S median consensus analysts' forecasts, which are publicized on the third Thursday of each month. The two-year ahead annual earnings (FY2) refers to the forecast horizon of FY2 in I/B/E/S. The next quarterly earnings (FQ) refers to the forecast horizon of quarter one (Q1) in I/B/E/S when analysts' forecasts are available, and if not, it corresponds to quarter two (Q2).[7] Our dataset for the FY2 forecasts begins in January 1983, in line with Bradshaw et al. (2012), and ends in December 2020. Due to limited observations before 1985, the dataset FQ forecast starts in January 1985 and ends in December 2020.

Our predictor set consists of 77 features that include the WRDS Financial Suite Ratios (van Binsbergen et al. (2022); de Silva and Thesmar (2022)), the current month's stock price, return, six-month momentum, industry momentum, and market capitalization from CRSP, as well as analyst related variables: the current analysts' forecast, the three-month revision of the analysts' forecast, the most recently realized earnings (annual EPS for FY2 and quarterly EPS for FQ), the realized analysts' forecast error, and the distance between the current month and the end of the forecast period.[8] In robustness tests, we also include macroeconomic variables.

We require the current analysts' forecast, the most recently realized earnings, the stock price, and price-to-sales to be non-missing, following Bradshaw et al. (2012). Additionally, we require returns, market capitalization, and both momentum variables to be non-missing. After cleaning our data, we are left with an average number of 2,421 and 2,338 firms each month in our test dataset for the FY2 and FQ forecasts, respectively. Table A3 describes the filters we apply to arrive at our final dataset. We then winsorize the variables in our predictor set on a monthly basis at the 1% and 99% levels.[9]

[7] I/B/E/S consensus forecasts are calculated on the third Thursday. So, if Q1 earnings as of the third Thursday are announced before month-end, Q2 EPS becomes the next quarterly earnings at month-end. We find similar results if we use the detail file to compute the consensus forecasts at month-end. We prefer the I/B/E/S consensus file because it is more stable over different historical versions than the detail file (Call et al. (2021)). Additionally, our results are also similar if we use the consensus mean analysts' forecast rather than the consensus median.

[8] The realized analysts' forecast error is calculated using past FY2 (FQ) forecasts that have the same (or the most similar) distance between the forecast date and the end of the forecast period as the current FY (FQ) forecasts. See detailed definitions in Tables A1 and A2.

[9] We fill the missing values for variables with the FF38 industry median value and if unavailable, the cross-sectional median value in each month following the winsorization.

Finally, in each estimation window, we standardize the predictors in the training dataset and use the mean and standard deviation of the training dataset to standardize the predictors in the test dataset. Specifically, at the end of each month t, we train the model using the dataset $(y_{i\tau}, X_{i\tau})$ with the target variable $y_{i\tau}$ known between months $t-119$ (or the beginning of our sample if using the expanding training window) and t. We then apply the fitted model to the predictor values as of month t (i.e., predictors in the test set) and generate the predicted value for the target variable, which would be the ML forecast for month t. Our first ML forecasts are generated in June 1990 to allow sufficient amount of training data and the last ML forecasts are generated in May 2020 to allow observations of realized earnings. The forecast accuracy as measured by ML Superiority in Eq. (2) is evaluated on an out-of-sample basis. See the Internet Appendix Section 3 for a more detailed discussion of the timeline.

4. Impact of Machine Learning Model Specifications

4.1. What Specification Choices Matter

We start by presenting the distribution of ML Superiority for our 3,024 estimated machine learning models in Figure 1. We find significant variability in the accuracy of ML forecasts, as nearly 90 percent of machine learning models have a negative ML Superiority (i.e., underperforming analysts' forecasts). Thus, understanding the impact of specification choices is key to making sense of prior papers' results on machine or analysts' superiority.[10]

To assess the impact of each specification choice listed in Table 2, we calculate the average ML Superiority and runtime over all possible combinations of choices across each of the choice sets. For instance, when evaluating estimation window choices, we compute the average ML Superiority and runtime for all ML models with an expanding window versus those with a rolling window. For conciseness, Table 3 presents results for choices associated with the highest and lowest ML Superiority within a specification choice set.

Table 3 shows that the loss function is the most critical factor affecting ML model performance

[10]Many studies report that ML models outperform analysts' forecasts: for short-horizon earnings from 1 to 3 quarters ahead (Ball and Ghysels (2018), Cao and You (2020), van Binsbergen et al. (2022), and Uddin et al. (2022)) and for longer horizons from 1 to 3 years ahead (So (2013), Ball and Ghysels (2018), Cao and You (2020), and van Binsbergen et al. (2022)). A notable exception is de Silva and Thesmar (2022), who report that analysts' forecasts are superior for 1 quarter to 1 year ahead earnings, whereas ML forecasts are superior for 2-4 years ahead earnings.

when forecasting earnings. Opting for the MAE loss function over the MSE loss function boosts average ML Superiority by 6.32%. This improvement is large because given that the median absolute value of EPS over price is 6.07% for FQ (annualized by multiplying by 4) and 6.43% for FY2.[11] This result highlights the importance of handling the outliers in the target variable when implementing ML models for earnings forecasting.

The second most crucial specification choice is the CV scheme for hyperparameter tuning. We observe that using the time-series CV instead of the panel CV scheme (i.e., standard 5-fold CV) increases the average ML Superiority by 1.47%. Our comparative analysis addresses the gap noted in Bertomeu (2020) that there is no guidance from theoretical or simulation studies on the appropriate CV approach when using accounting data. Our results indicate the assumption of independent observations inherent in standard k-fold CV is violated in the earnings forecasting setting, and support Bertomeu (2020)'s advocacy for the use of time-series CV that preserves the data's temporal order.

The other three specification choices related to ML model training have a substantially smaller impact on model performance. Different choices of training windows, refitting the model, and tuning the hyperparameters result in a difference in average ML Superiority of 0.27%, 0.10%, and 0.09%, respectively.

For the four specification choice sets not directly related to ML model training, we find that: (1) adopting Frankel and Lee (1998)'s indirect approach to forecast earnings (as opposed to forecasting EPS directly) yields an increase in ML Superiority of 1.21%. This result suggests that ML forecasts are more accurate when the machine focuses on correcting predictable analysts' forecast biases; (2) including analysts' forecasts in the predictor set improves ML Superiority by 0.68%, which corroborates the findings of van Binsbergen et al. (2022); de Silva and Thesmar (2022) in a more exhaustive and definitive manner; (3) ML Superiority is higher for long-distance forecasts (FY2) than for short-distance forecasts (FQ) by 0.65%, aligning with existing research that shows analysts' forecasts are more accurate for near-term earnings; and (4) price scaling all EPS-related variables

[11] We show in Internet Appendix Section 4 that the best-performing choice within each specification choice set remains the same across the board when utilizing Mean Squared Errors rather than Mean Absolute Errors to compute the superiority measure.

in the forecasting step enhances ML Superiority by 0.48%.

Our results in Table 3 thus pinpoint three key specification choices: MAE as the loss function, a time-series CV scheme, and the indirect forecasting approach. To highlight the importance of these choices, Figure 1 overlays the distribution of the ML Superiority for ML models that employ these three choices over the distribution of the ML Superiority for all models. We observe that under the constrained specification choices, ML models exhibit considerably less variation in performance and consistently outperform analysts' forecasts.

Table 3 also reports the computational runtime for each specification choice.[12] Specification choices significantly affect runtime, with the most substantial impact totaling a difference of 212 computational hours attributed to variation in the frequency of hyperparameter tuning. Overall, the process of training and fitting all 3,024 machine learning models requires approximately five years of computational time.

Finally, we present the best-performing specification for each ML algorithm.[13] Table 4 shows that ML models with the best-performing specifications consistently outperform analysts' forecasts for both FQ and FY2. The ML Superiority varies from 0.113% (OLS) to 0.201% (GBRT) for FQ, and from 0.267% (OLS) to 0.601% (GBRT) for FY2. The GBRT algorithm delivers the highest ML Superiority for both FQ and FY2, indicating the value of allowing the model to capture non-linear predictable relationships in the data.[14] In situations where faster computing is preferable, we recommend researchers use the three aforementioned specification choices (MAE loss function, time-series CV, and indirect approach) and then select configurations from the remaining choice sets to minimize computing time without substantially impacting model performance. We offer these time-saving specification recommendations in the Internet Appendix Section 4 and show the superiority and computational hours in comparison to our best specifications.

[12]The computational runtime for a ML model includes the time spent on tuning hyperparameters, training the model, and generating the ML model forecast at the end of each month. To accelerate the process, we utilize large-scale computing servers at our university, allowing us to run computing jobs concurrently.

[13]Given that most extant studies do not scale EPS in the forecasting step, we restrict our results to models with this scaling choice through the rest of the paper.

[14]To examine whether the top-performing specifications are stable over time, we divide the thirty-year sample into three decade-long subsamples and evaluate the top-performing specifications within each decade. Section 5 of the Internet Appendix presents these results and shows that the top-performing models in each decade consistently use the GBRT algorithm with the three key specification choices mentioned above.

Overall, our findings identify significant variability in ML model performance and demonstrate the importance of four key specification choices — (1) the GBRT algorithm that exploits non-linear relationships, (2) the MAE loss function that is robust to outliers, (3) the indirect approach that focuses on minimizing analysts' forecast errors, and (4) the time-series CV scheme that preserves time-series order in training and validation samples. The remaining other specification choices have relatively minor effect on the forecasting performance. For brevity, we focus on the GBRT model with the best-performing specification in Table 4, and hereafter refer to it as the statistically optimal ML forecast.[15]

4.2. The Role of Analysts' Forecasts

If the statistically optimal ML model consistently outperforms analysts' forecasts, is there enduring relevance of analysts' forecasts in the machine learning era? To answer this question, we examine three scenarios that differ in how the ML models (with the best-performing specifications for a given algorithm) encode information from analysts' forecasts. First, we use ML models to forecast EPS directly without analysts' forecasts included in the predictor set (direct approach w/o analysts); second, with analysts' forecasts included in the predictor set (direct approach w/ analysts); and third, with using the ML models to predict analysts' forecast errors (indirect approach).

The first two rows of Table 5 show the results for the direct approach w/o analysts. Under this approach, none of the six ML models outperform analysts' predictions for forecasting FQ EPS, and only one model, GBRT, significantly outperforms analysts' for forecasting FY2 EPS. Even so, the outperformance of GBRT is only 0.205% (t-stat=2.39), which represents a modest improvement because it is equivalent to a reduction in absolute EPS forecast errors of 2.1 cents for a $10 stock.

Rows 3 and 4 of Table 5 shows that the direct approach w/ analysts leads to a substantial increase in ML Superiority across all models. Despite this improvement, for FQ forecasts, only the GBRT model can outperform analysts' forecasts, with a small ML Superiority of 0.0921% (t-stat=3.51). For FY2 forecasts, all models except the RF outperform analysts, with the GBRT

[15]To further alleviate the specification mining concern, in addition to the GBRT model with the best-performing specification in Table 4, our website will also provide an alternative forecast based on the same four key specifications: the GBRT algorithm, the MAE loss function, the indirect approach, and the time-series CV, but it averages across all possible combinations of choices over the frequency of hyperparameter re-tuning, the frequency of model re-fitting, and the estimation window.

standing out as the top model with a large ML Superiority of 0.536% (t-stat=7.20).[16]

When using the indirect approach, *all* ML algorithms with the best-performing specifications outperform analysts in a statistically significant way for both FQ and FY2. It is noteworthy that switching from the direct forecasting approach (w/ analysts) to the indirect forecasting approach significantly narrows the performance gap between the worst and the best algorithm, from 2.5% to 0.1% for FQ and from 1.0% to 0.3% for FY2.

In summary, these findings demonstrate the indispensable role of analysts' forecasts in accurate earnings predictions: even sophisticated ML models struggle to match analysts' forecasts without including analysts' information. Once ML models are allowed to learn from analysts, all six ML models—including the elementary OLS model—–can exceed analysts in terms of accuracy.

4.3. Feature Importance

Since the top ML model outperforms analysts by correcting the predictable errors in analysts' forecasts, we further analyze what variables does the statistically optimal ML model rely on to detect errors in analysts' forecasts? Conceptually, the predictable errors in analysts' forecasts may be due to analysts' inability to use all available information or their behavioral biases (Bertomeu (2020)). If the predictable errors are related to the analysts' inability to use all information, then the most important features of the top ML Model should capture information that is not contained in analysts' forecasts. However, if the predictable errors are related to behavioral biases, then analyst-related variables would be the most important features.

We use the drop-column feature importance to identify the most important features of the top ML model. The basic idea is that if a feature is not as important, excluding it from the predictor set should not noticeably decrease the model's out-of-sample performance. The drop-column feature importance not only accounts for the inter-correlations between features but also accounts for the substitution effect between the dropped feature and the remaining features.[17] Specifically, we calculate the decrease in the ML Superiority measure as defined in Eq. (2) when feature k is

[16]The enhanced ML superiority resulting from the addition of analyst variables to the predictor set can serve as a measure of the additional information that ML models are capable of extracting from these analyst variables. We provide a detailed analysis of this additional value added in Internet Appendix Section 6, demonstrating that analysts contribute more value for short-horizon forecasts and for smaller and more complex firms. The value added by analysts is very persistent over time.

[17]The Internet Appendix Section 7 provides more analysis of substantial non-linear and interactive relations in predicting analysts' forecast errors.

excluded from the predictor set. We then scale the decrease by the ML Superiority measure when all predictors are used.

$$\%\Delta\text{superiority}_k = \frac{\text{superiority}_{\text{All}} - \text{superiority}_{\text{All}\setminus\{k\}}}{\text{superiority}_{\text{All}}} \tag{3}$$

A $\%\Delta\text{superiority}_k$ of 50% indicates that removing variable k from the predictor set reduces the model's superiority (relative to analysts' forecasts) by half.

Figure 2 presents the top 10 features for the top ML model for FQ and FY2 forecasts. The current analysts' forecasts (AF) show up as the most influential feature for both horizons. Removing AF from the predictor set results in a 44.5% reduction in ML Superiority for FY2 and a 22.9% reduction in ML Superiority for FQ. The stock price (PRC) and the realized forecast errors of prior analysts' forecasts (ErrAF) rank as the second most influential feature for FY2 and FQ, respectively. The removal of these features result in decreases of 17.8% and 17.7% in ML Superiority. All the other features do not play as critical of a role because removing them reduces ML Superiority by less than 10%. Our finding that AF is the most important feature for predicting analysts' forecast errors leans towards behavioral bias interpretations.

4.4. The Magnitude of ML Superiority

4.4.1. Cross Sectional Variation in ML Superiority

The accuracy of analysts' earnings forecasts are known to covary with firm characteristics.[18] We thus examine whether these characteristics also affect the extent to which the statistically optimal ML forecast outperforms analysts' forecast. Table 6 reports the Fama and MacBeth (1973) multivariate regression of the ML Superiority on these firm attributes. For ease of interpretation, we use the normalized rank (i.e., the rank scaled by the number of stocks in a cross-section) of these firm characteristics as independent variables. Therefore, if one characteristic changes from the 25th percentile to the 75th percentile while the other characteristics remain the same, the corresponding change in ML Superiority is 0.5 times the respective multi-variate regression coefficient.

Table 6 shows that the ML Superiority is larger among stocks with smaller size, more business

[18] Table A4 in Appendix A provides detailed definitions of these firm attributes, which are related to information uncertainty, firm complexity, price informativeness, analysts' incentives, and earnings management (citations of where attributes are from). We account for time-series and cross-sectional correlations by computing the t−statistics based on Newey-West standard errors with 24 lags.

segments, higher idiosyncratic volatility, lower institutional ownership and high bid-ask spreads. These results are consistent with the prior study that analysts' bias tend to be larger among these stocks. A standout finding is the large coefficient on firm size, which is several times larger than the coefficients on any other characteristic. We therefore further report the ML Superiority by size quintiles in Figure 3. We find that for stocks in the top (largest) size quintile, the ML Superiority is just 0.062% and 0.064% for FQ and FY2, respectively. These improvements are small as they are equivalent to a reduction in absolute EPS forecast errors of 0.62 and 0.64 cents for a $10 stock. In contrast, the improvement is an order of magnitude larger among firms in the bottom size quintile, with ML Superiority of 0.569% for FQ and 1.712% for FY2. Given that stocks in the top size quintile account for 87% of total equity market capitalization, on average during our sample period, we conclude that the ML forecasts proposed in the literature do not substantially improve the earnings forecasting accuracy for the majority of the market capitalization.

4.4.2. Forecast Horizon Effect in ML Superiority

Bradshaw et al. (2012) have documented that analysts' forecasts tend to become more accurate as the announcement date approaches. We thus follow the methodology in Bradshaw et al. (2012) and analyze how ML Superiority varies as the distance between the forecast date and the earnings announcement changes in Figure 4.

We find the ML Superiority of the statistically optimal forecast becomes progressively smaller as the distance to earnings announcement decreases. Panel A of Figure 4 demonstrates a strong horizon effect. For FQ forecasts made one month prior to the earnings announcement, the ML Superiority of the statistically optimal forecast is merely 0.086%. Though statistically significant, this ML Superiority is small: an accuracy improvement of less than 0.86 cents for forecasting FQ EPS for a $10 stock (annualized by multiplying by 4). As the distance between the forecast date and earnings announcement increases, we see a notable increase in ML Superiority. For FQ forecasts generated 3 months prior to the earnings announcement, GBRT's ML Superiority quadruples to 0.273%. Panel B shows that GBRT's ML Superiority for FY2 forecasts made 12 months before the earnings announcement is 0.518% (5.18 cents for forecasting FY2 EPS for a $10 stock), over six times larger than the ML Superiority for FQ forecasts made one month prior to the earnings announcement. This superiority further increases to 0.674% for forecasts made 19 months ahead of the earnings announcement, then tapers slightly to 0.571% for forecasts made 23 months in advance.[19]

These results thus indicate that the statistically optimal ML forecast cannot substantially improve upon analysts' forecasts for near-term earnings, but it can bring material enhancement for FY2 forecasts.

4.4.3. Time Series Variation in ML Superiority

We next examine whether there is a time trend in the ML superiority over analysts. While advancements in statistical models increase model forecasting accuracy over time, analysts can also incorporate these technological advancements in their forecasting processes to refine their forecasts. So it is ex-ante unclear which force dominates in our sample period.

Figure 5 shows the 10-year rolling average of the ML superiority for the statistically optimal forecast. We observe a pronounced downward trend in its advantage over analysts' forecasts, with this edge more than halving over our sample period. The decline is particularly steep prior to 2012, but it then stabilizes, remaining positive and statistically significant for both the FQ and FY2 forecasts at the end of our sample.[20]

This time-series trend supports the idea that analysts increasingly use machine learning tools to minimize their errors causing their forecasts to converge with the statistically optimal forecasts over time. Combined with earlier findings that the statistically optimal forecast relies on analysts' information, our analysis suggests that the distinction between machine and human forecasts is blurring as they increasingly inform and converge with each other.

4.5. Comparison to Crowdsourced Earnings Estimates

Prior research such as Jame et al. (2016); Schafhautle and Veenman (2024) demonstrate that earnings forecasts from the crowdsourcing platform Estimize offer additional value beyond that provided by I/B/E/S analysts' forecasts. In this section, we compare the performance of our

[19] Further supporting the horizon effect, we find in untabulated results that the average GBRT's ML Superiority for FY1 forecasts is 0.25%, which is between 0.20% for FQ forecasts and 0.60% for FY2 forecasts. Section 8 of the Internet Appendix presents the same analysis for the best ML forecast based on the other 6 algorithms.

[20] In unreported results, we find that the statistically optimal forecast underperform analysts' forecasts when the macroeconomy is coming out of a recession.

statistically optimal machine learning (ML) forecasts with Estimize forecasts in terms of bias and statistical accuracy.[21]

In Panel A of Figure 6, we examine the pessimism bias of analysts' short-term forecasts documented in prior studies. We first replicate the findings in Jame et al. (2016); Schafhautle and Veenman (2024) that Estimize forecasts are characterized by a less pronounced walkdown in forecasts as the earnings announcement approaches. For example, at the month end immediately preceding the earnings announcement, the I/B/E/S consensus forecast exhibits strong pessimism, falling short of the actual earnings for 73% of the observed firm-quarters. In contrast, only 61% of Estimize forecasts underestimate actual earnings. We then apply the same analysis on the statistically optimal ML forecasts and find that the statistically optimal forecast shows an even lower level of pessimism bias than Estimize forecasts. Only 53% of the ML forecasts are lower than the actual earnings prior to announcements, with noticeably less walkdown in the forecasts compared to both the I/B/E/S and Estimize forecasts. [22]

In Panel B of Figure 6, we compare the forecast accuracy of these forecasts by computing Estimize Superiority (constructed analogously to ML Superiority). Consistent with Jame et al. (2016), we find that the Estimize consensus forecast is less accurate than the I/B/E/S analysts' consensus forecast, but the combined forecast of these two outperforms the I/B/E/S analysts' consensus forecast.[23] Importantly, we find that the statistically optimal ML forecast outperforms both the Estimize consensus and the combined forecast in terms of accuracy.

Overall, the statistically optimal ML forecast not only exhibits less pessimism bias compared to crowdsourced earnings forecast, but it also surpasses them in accuracy. These results support the idea that there are limits to the extent that crowdsourcing unravels bias in analysts' forecasts,

[21]We follow the prior literature (e.g Jame et al. (2016); Brown and Khavis (2018); Da and Huang (2020)) in constructing the Estimize consensus forecast. If a contributor makes multiple forecasts on one day for a given firm/forecast quarter, the average of the forecasts is retained. We require that the realized earnings given by Estimize matches that of the reported realized earnings in I/B/E/S. Flagged forecasts, quarterly forecasts that are made over 120 days prior to as well as those made after the earnings announcement are removed. We retain the most recent forecast made by a contributor for a given firm-quarter and use the median of these forecasts to create the consensus forecast at the end of each month. The results are robust to constructing the consensus forecast as of I/B/E/S forecast date (statpers). Please see our Internet Appendix Section 9 for details.

[22]Note that a perfectly unbiased estimate is equivalent to 50% of the forecasts being lower than the actual earnings.

[23]Tables 8 and 9 of Jame et al. (2016) indicate that the optimal combined forecast is roughly an equal-weighted average of the Estimize consensus forecast and the I/B/E/S analysts' consensus forecast.

and that machine learning does a better job of this.

5. Machine Learning Earnings Forecasts and Market Expectations

After determining the statistically optimal ML earnings forecast, we now examine the extent to which investors' expectations align with the statistically optimal machine learning forecast or analysts' forecasts. If investors are fully rational, market-expected earnings should align more closely with the most statistically accurate forecast. However, if investors share the same biases as analysts, as suggested by Bordalo et al. (2019), market-expected earnings may exhibit excess correlation with analysts' forecasts, diverging from the statistically optimal forecast.

We test these competing hypotheses in two approaches. First, we use earnings response coefficients (ERC) to infer which earnings forecast is a more precise proxy for market-expected earnings. Second, we investigate whether investors share the same biases as analysts by examining the return predictability of ex-ante measures of analysts' bias.

5.1. ERC Tests

To the extent that earnings announcement returns are predominantly driven by revisions in market earnings expectations in reaction to the announcements, earnings surprises computed using a more precise proxy for market-expected earnings should be associated with larger ERCs. Therefore, we regress abnormal returns around earnings announcements on earnings surprise measures based on analysts' forecasts or the statistically optimal earnings forecasts as follows:

$$R_{d-1,d+1} = c + \beta \times \mathrm{SUE}_{t-1,d} + \epsilon_{d-1,d+1} \tag{4}$$

, where $R_{d-1,d+1}$ is the size-adjusted 3-day return around the earnings announcement day d.[24] $\mathrm{SUE}_{t-1,d}$ is the earnings surprise measured as the difference between quarterly I/B/E/S reported earnings announced in month t and the expected earnings (proxied by analysts' or machine FQ forecasts made at month end prior to the announcement) scaled by the price at the prior month end, and β is the earnings response coefficient (ERC). Following Basu (1997), we trim $\mathrm{SUE}_{t-1,d}$

[24]To create the size adjusted returns around the earnings announcements, we first assign firms in each month to five size buckets based on the 20th, 40th, 60th, and 80th percentile of the size distribution for NYSE listed firms. Then, we calculate the value-weighted average daily return within each size group. Finally, for the three days around a given earnings announcement, we take a firm's three-day return and subtract off the three-day return of the corresponding size matched portfolio.

and $R_{d-1,d+1}$ at the 1% and 99% levels by quarter.[25]

5.1.1. How Market Earnings Expectations Align with ML Forecasts

We first examine how market earnings expectations align with forecasts from our 1,512 ML models for FQ earnings. We run the univariate regression in Eq. (4) using each of these ML forecasts as well as analysts' forecast. We find that the resulting ERC have a strong positive association with ML Superiority, having a rank correlation of 0.971. To visualize this strong positive association, we sort ML forecasts into deciles based on their ML Superiority and show the distribution of the ERCs within each decile in Figure 7. SUEs based on the machine forecasts in the top ML Superiority decile are predominantly associated with higher ERCs than the analysts' forecasts-based SUE, while the opposite is true for the machine forecasts in the other deciles. Overall, only 8% of the machine forecasts are associated with higher ERCs than analysts' forecasts, corroborating our earlier findings that the specification choices of the ML model matter and that 90% of ML forecasts are less accurate than analysts' forecasts.

Columns 1 and 2 of Table 7 tabulate the univariate ERC regression results using SUE^{ML} (based on the statistically optimal forecast) and SUE^{AF}, respectively. We find that both the ERC and the R^2 are larger when using the statistically optimal forecast rather than analysts' forecast to compute the earnings surprises. Overall, our results indicate that investors' earnings expectations are rational in the sense that they align more closely with the more accurate earnings forecasts.

5.1.2. Investors' Weighting of the Statistically Optimal Forecasts and Analysts' Forecasts

We then test the extent to which investors' earnings expectations align with the statistically optimal machine learning forecast and analysts' forecasts through the following bivariate regressions:

$$R_{d-1,d+1} = c + \beta_1 \times \text{SUE}^{ML}_{t-1,d} + \beta_2 \times \left(\frac{\text{ML}_{t-1} - \text{AF}_{t-1}}{P_{t-1}} \right) + \epsilon^{ML}_{d-1,d+1} \tag{5}$$

where ML_{t-1} and AF_{t-1} are the best statistical FQ forecast and the analysts' FQ forecast (made at the month end prior to the announcement), respectively; the second term in the regression captures the deviation of analysts' forecast from the statistically optimal forecast (hereafter referred to as

[25] We construct the earnings announcement days following the methodology in Dellavigna and Pollet (2009); Johnson and So (2018). Our results are robust to the inclusion of time fixed effects.

Bias^{AF}).[26]

To see why these regression coefficients are informative, suppose that the true SUE based on market expectations (SUE^M) is a weighted average of SUE^{AF} and SUE^{ML}:

$$\text{SUE}^M_{t-1,d} = w_{AF}\text{SUE}^{AF}_{t-1,d} + w_{ML}\text{SUE}^{ML}_{t-1,d} \tag{6}$$

and the ERC based on the true SUE is as follows:

$$R_{d-1,d+1} = c + \beta^M \times \text{SUE}^M_{t-1,d} + \epsilon^M_{d-1,d+1}. \tag{7}$$

Substitute $\text{SUE}^M_{t-1,d}$ of Eq. (7) out using Eq. (6), we have:

$$R_{d-1,d+1} = c + \beta^M \times (w_{AF} + w_{ML})\,\text{SUE}^{ML}_{t-1,d} + \beta^M \times w_{AF}\left(\frac{\text{ML}_{t-1} - \text{AF}_{t-1}}{P_{t-1}}\right) + \epsilon^M_{d-1,d+1}$$

Therefore, we can infer the relative weight the market earnings expectations assign to analysts' forecasts and the statistically optimal forecast from the regression coefficients in Eq. (5):

$$\frac{\beta_2}{\beta_1} = \frac{\beta^M \times w_{AF}}{\beta^M \times (w_{AF} + w_{ML})} = \frac{w_{AF}}{w_{AF} + w_{ML}} \tag{8}$$

Under the rational expectation hypothesis, investors' earnings expectations should perfectly align with the statistically optimal forecast, leading to a β_2 value of zero. Note even in this case, investors' earnings expectations still depend on analysts' forecasts. Here investors assign the same weight on analysts' forecasts as the statistically optimal forecast does. On the other hand, if investors overweight analysts' forecasts relative to the statistically optimal forecast, then β_2 should be positive. The ratio of β_2 relative to β_1 measures the extent of investors' excessive reliance on analysts' forecasts relative to that prescribed by the statistically optimal forecast.

Column 3 of Table 7 shows that $\beta_1 = 0.25$ (t-stat = 6.5) and $\beta_2 = 0.096$ (t-stat = 6.0), implying that $\frac{w_{AF}}{w_{AF}+w_{ML}} = 38.2\%$. Contrary to the rational expectations hypothesis, where investors would assign 100% weight to the statistically optimal forecast, our empirical analysis shows that market

[26]Note that $\frac{\text{ML}_{t-1}-\text{AF}_{t-1}}{P_{t-1}}$ is equal to the difference between $\text{SUE}^{AF}_{t-1,d}$ and $\text{SUE}^{ML}_{t-1,d}$. So, an alternative specification is to regress $R_{d-1,d+1}$ on $\text{SUE}^{AF}_{t-1,d}$ and $\text{SUE}^{ML}_{t-1,d}$, but this specification suffers from multicolinearity as $\text{SUE}^{AF}_{t-1,d}$ and $\text{SUE}^{ML}_{t-1,d}$ are highly correlated.

expectations place approximately a 6:4 ratio on the statistically optimal forecast and analysts' forecasts, respectively. This 40% weight is in excess of the optimal weight on analysts' forecasts as prescribed by the statistically optimal forecast.

5.1.3. Systematic Variation in Investors' Overweighting of Analysts' Forecasts

We further investigate how investors' overweighting of analysts' forecasts differs across firm size and across varying levels of investor sophistication. To do so, we interact SUE^{ML} and Bias^{AF} with firm size and institutional ownership (IO). For ease of interpretation, we again use the normalized rank (i.e., the rank scaled by the number of stocks in a cross-section) of these firm characteristics. Column 4 of Table 7 shows a positive coefficient on the interaction term between SUE^{ML} and firm size and a negative coefficient on the interaction term between Bias^{AF} and firm size, both of which are highly significant. The signs of these coefficients mean that moving from small-cap to large-cap stocks, market expectations assign a progressively larger weight on the statistically optimal and smaller weight on the analysts' forecasts. The regression coefficients on the interaction terms indicate that if the size rank changes from the 25th percentile to the 75th percentile, investors' over-weighting of analysts' forecasts reduce by 22.8%, which is economically large relative to the average 38.2% excess weight.

Column 5 of Table 7 shows that the regression coefficient on the interaction term between SUE^{ML} and IO is positive and significant, whereas the coefficient on the interaction term between Bias^{AF} and IO is negative and insignificant. This means when going from stocks with low IO to stocks with high IO, market expectations assign progressively more weight on the statistically optimal forecast without substantially changing the weight on analysts' forecasts. The magnitude of these coefficients indicate that when the rank of institutional ownership goes from the 25th percentile to the 75th percentile, the market weight on the analyst reduces by 22.3%, again an economically large reduction when compared to the average 38.2% excess weight. Further, in untabulated results, when we restrict the sample of firms to only include those with above median IO and above median firm size, the coefficient on Bias^{AF} becomes statistically insignificant, with an economically small excess weight on the analyst forecast of 7.3%. This implies that among large-cap firms with more sophisticated investors, the market's over-reliance on analysts' forecasts when forming earnings expectations becomes statistically insignificant.

5.1.4. The Effect of Crowdsourced Estimates on Investors' Overweighting of Analysts' Forecasts

We then examine whether the Estimize consensus forecast provides additional information for measuring market earnings expectations beyond what is captured by the statistically optimal ML forecast. We conduct the same bivariate ERC regressions of earnings announcement returns on SUE^{ML} and a measure that captures the deviation of a forecast from the statistically optimal forecast. This analysis is limited to firm/months with Estimize consensus forecasts. Similar to Bias^{AF}, we construct $\text{Bias}^{Estimize}$ and $\text{Bias}^{Combine}$ for which we replace I/B/E/S analysts' consensus forecasts with Estimize consensus forecasts and the combined forecast, respectively.[27]

Columns (1) to (3) of Panel C in Table 8 show that the regression coefficients on SUE^{ML} are positive and highly significant, while the coefficients on Bias^{AF}, $\text{Bias}^{Estimize}$, and $\text{Bias}^{Combine}$ are all statistically insignificant. These results indicate that market expectations align almost exclusively with the statistically optimally forecast in this sample, attributing statistically insignificant additional weight to the I/B/E/S analysts' forecast, the Estimize consensus, or the combined forecast. This result is consistent with our earlier findings that market expectations tend to be more efficient for large-cap stocks and for those with substantial institutional ownership, which have greater coverage on the Estimize platform. Our results complement the finding in Jame et al. (2016), suggesting that while Estimize forecasts provide incremental value to I/B/E/S analysts' forecasts in measuring market earnings expectations, this incremental value is effectively incorporated in the statistically optimal ML forecast.

Although the crowdsourced forecasts do not appear to provide extra value beyond what is already captured by statistically optimal forecasts in measuring market earnings expectations, the availability of alternative forecasts from social media can nevertheless affect how investors process analysts' forecasts (Jame et al. (2016); Drake et al. (2023); Schafhautle and Veenman (2024)). Following the empirical strategy in these papers, we examine how the staggered addition of firms to the Estimize platform affects investors' weighting of the statistically optimal forecast and the analysts' forecasts.

[27] See the definition of Estimize consensus forecasts and the combined forecast in Subsection 4.5.

We interact SUE^{ML} and Bias^{AF} in Eq. (5) with an indicator variable Post, which is equal to one for periods following the initialization of Estimize coverage for a given firm. We control for the linear time trend effect to ensure the results are indeed driven by the initiation of Estimize coverage. To allow for a control group and pre-coverage period, following Schafhautle and Veenman (2024), we use all firm-months in our dataset from January 2010 to May 2020, so our analysis also includes firms that are never covered by Estimize.

Column 4 of Table 8 shows that the regression coefficient on the interaction term between SUE^{ML} and Post is 0.199 and highly significant (t-stat = 4.77), while the coefficient on the interaction term between Bias^{AF} and Post is -0.055 and insignificant (t-stat = -0.81). These results suggest that following the initialization of coverage on Estimize, investors' expectations assign more weight on the statistically optimal ML forecast. As a result, the initialization of Estimize coverage leads to a decrease in the excess weight on the analysts' forecast of 20.7%, which is economically significant relative to the average excessive weight of 40% as documented in the previous subsection.[28]

Our results indicate that the introduction of alternative forecasts from social media leads to a decreased reliance on analysts' forecasts, corroborating earlier findings by Jame et al. (2016); Drake et al. (2023); Schafhautle and Veenman (2024). More importantly, we provide new evidence that, as a result, investors' expectations converge towards objectively more accurate forecasts.

5.2. Return Predictability of Analysts' Bias

In our second market expectation test, we examine the return predictability of analysts' biases identified by the statistically optimal ML forecast (i.e., Bias$^{AF} \equiv \frac{\text{ML}_{t-1} - \text{AF}_{t-1}}{P_{t-1}}$). Given that our ERC results indicate that investors' overweighting of analysts' forecasts varies systematically with size and institutional ownership, we conduct the analysis by dividing our sample into small-cap/large-cap and above-/below-median institutional ownership groups. We form monthly portfolios by first dividing stocks in the cross section into four groups based on the cross sectional medians of firm size and institutional ownership. Then, within each group we divide the firms into quintiles based on Bias^{AF} and compute the value weighted quintile portfolio returns. Finally, we report the average excess return, FF5 alpha, and q^5 alpha of the long short Q5-Q1 portfolio for each group in

[28]The 20.7% decrease in the excess weight on the analysts' is computed after controlling for the effect due to the time trend (i.e., setting Time=0): $\frac{.210-.055}{.485+.199} - \frac{.210}{.485} \approx .207$.

Table 9.

Panel A presents the results for the Q5-Q1 portfolio sorted on Bias^{AF} for analysts' FQ forecasts. We find that Bias^{AF} is positively associated with future returns across all four groups. When analysts' forecasts exceed the statistically optimal forecasts (i.e., a negative Bias^{AF} reflecting analysts' optimism), future stock returns are lower. This confirms our previous finding that investors' expectations place excessive weights on analysts' forecasts. Moreover, the Q5-Q1 portfolio returns are the largest among the small firms with low IO with a statistically significant excess return of 1.825% per month (t-stat=8.04). In contrast, the Q5-Q1 portfolio returns are almost an order of magnitude smaller and statistically insignificant among large firms with high IO. This contrasting pattern is robust to using either the FF5 or the q^5 model to compute risk-adjusted alphas. These results thus corroborate our ERC test results that investors' overweighting of analysts' forecasts is concentrated among small firms with low IO.

Panel B presents the results for the Q5-Q1 portfolio sorted on Bias^{AF} for analysts' FY2 forecasts. We find a similar pattern that the return predictability of analysts' bias is strongest among small firms with low IO and weakest among large firms with high IO. Interestingly, we find that the return predictability of analysts' FY2 bias is not substantially different that of analysts' FQ bias, despite the fact that ML superiority is much larger for FY2 forecasts. This highlights the distinction between a forecast's statistical accuracy and suitability as a proxy for market earnings expectations.

Taken together, our two market expectations tests demonstrate that while neither the statistically optimal ML forecast nor analysts' forecasts are the prefect proxy for investors' expectations, employing both of them helps further our understanding of investors' expectations. On the one hand, market expectations appear to align closely with the statistically optimal forecast. On the other hand, market expectations still tend to emphasize analysts' forecasts more than what the statistically optimal forecast would recommend. This tendency to overweight analysts' forecasts is more pronounced for firms that are small or have low institutional ownership.

6. Conclusion

We comprehensively examine the superiority of various machine learning methods and analysts' forecasts in predicting earnings to provide an updated view on which earnings forecast minimizes ex-post forecast errors and which best aligns with investors' earnings expectations. To determine the most appropriate machine learning specification, we evaluate the impact of specification choices used in recent machine learning earnings forecast studies by exhaustively comparing 3,024 models derived from the full combination of nine sets of specification choices and six machine learning algorithms. We find that only a handful of specification choices significantly impact machine learning model forecast accuracy, with most having a minimal effect. Our analysis, complete with codes and estimates, provides a much needed bridge between the earlier literature and the recent (and future) machine learning earnings forecasting studies.

Contrary to the impression created by many recent machine learning studies, we find that even the best machine expectation is only marginally more accurate than analysts' forecasts in most cases, especially when the analyst faces stronger incentives to be accurate. Despite this fact however, we show that in cases where analyst and machine expectations differ, investors' expectations put more weight on analysts' forecasts than prescribed by the best machine forecast. Investors' overweighting becomes statistically insignificant among large-cap firms with more sophisticated investors. Finally, our time-series analyses suggest that analyst and machine forecasts are converging over time and that analysts' information production remains a critical input to the best machine forecast, blurring the line between human and machine forecasts. Overall, our study provides an updated and comprehensive take on the most accurate earnings forecast and the best proxy for investors' earnings expectations.

References

Amir, E., Lev, B., and Sougiannis, T. 2003. Do financial analysts get intangibles? European Accounting Review, 12(4), 635–659. ISSN 0963-8180, 1468-4497. doi:10.1080/0963818032000141879.

Ball, R. and Brown, P. 1968. An Empirical Evaluation of Accounting Income Numbers. Journal of Accounting Research, 6(2), 159–178. ISSN 0021-8456. doi:10.2307/2490232. Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].

Ball, R. T. and Ghysels, E. 2018. Automated Earnings Forecasts: Beat Analysts or Combine and Conquer? Management Science, 64(10), 4936–4952. ISSN 0025-1909, 1526-5501. doi:10.1287/mnsc.2017.2864.

Basu, S. 1997. The conservatism principle and the asymmetric timeliness of earnings1. Journal of Accounting and Economics, 24(1), 3–37. ISSN 01654101. doi:10.1016/S0165-4101(97)00014-1.

Basu, S. and Markov, S. 2004. Loss function assumptions in rational expectations tests on financial analysts earnings forecasts. Journal of Accounting and Economics, 38, 171–203. ISSN 0165-4101. doi:10.1016/j.jacceco.2004.09.001.

Bernard, V. L. and Thomas, J. K. 1989. Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium? Journal of Accounting Research, 27, 1–36. ISSN 0021-8456. doi:10.2307/2491062.

Bernard, V. L. and Thomas, J. K. 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics, 13(4), 305–340. ISSN 0165-4101. doi:10.1016/0165-4101(90)90008-R.

Bertomeu, J. 2020. Machine learning improves accounting: discussion, implementation and research opportunities. Review of Accounting Studies, 25(3), 1135–1155. ISSN 1573-7136. doi:10.1007/s11142-020-09554-9.

Bertomeu, J., Cheynel, E., Liao, Y., and Milone, M. 2021. Using Machine Learning to Measure Conservatism. SSRN Electronic Journal. ISSN 1556-5068. doi:10.2139/ssrn.3924961.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. 2019. Diagnostic Expectations and Stock Returns. The Journal of Finance. ISSN 0022-1082, 1540-6261. doi:10.1111/jofi.12833.

Bradshaw, M. T., Drake, M. S., Myers, J. N., and Myers, L. A. 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. Review of Accounting Studies, 17(4), 944–968. ISSN 1380-6653, 1573-7136. doi:10.1007/s11142-012-9185-8.

Bradshaw, M. T., Lee, L. F., and Peterson, K. 2016. The Interactive Role of Difficulty and Incentives in Explaining the Annual Earnings Forecast Walkdown. The Accounting Review, 91(4), 995–1021. ISSN 0001-4826. Publisher: American Accounting Association.

Brown, L. D. and Khavis, J. The Reliability of Crowdsourced Earnings Forecasts. 2018. doi: 10.2139/ssrn.3034323.

Call, A. C., Hewitt, M., Watkins, J., and Yohn, T. L. 2021. Analysts' annual earnings forecasts and changes to the I/B/E/S database. Review of Accounting Studies, 26(1), 1–36. ISSN 1573-7136. doi:10.1007/s11142-020-09560-x.

Cao, K. and You, H. Fundamental Analysis Via Machine Learning. 2020. doi:10.2139/ssrn.3706532.

Cao, S. S., Jiang, W., Wang, J. L., and Yang, B. 2021. From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. SSRN Electronic Journal. ISSN 1556-5068. doi: 10.2139/ssrn.3840538.

Chattopadhyay, A., Fang, B., and Mohanram, P. 2023. Machine Learning, Earnings Forecasting, and Implied Cost of Capital - US and International Evidence. Working Paper.

Chen, A. Y. and Zimmermann, T. 2022. Open Source Cross-Sectional Asset Pricing. Critical Finance Review, page 69.

Da, Z. and Huang, X. 2020. Harnessing the Wisdom of Crowds. Management Science, 66(5), 1847–1867. ISSN 0025-1909. doi:10.1287/mnsc.2019.3294. Publisher: INFORMS.

Das, S., Levine, C. B., and Sivaramakrishnan, K. 1998. Earnings Predictability and Bias in Analysts' Earnings Forecasts. The Accounting Review, 73(2), 277–294. ISSN 0001-4826. Publisher: American Accounting Association.

de Silva, T. and Thesmar, D. Noise in Expectations: Evidence from Analyst Forecasts. 2022. doi:10.2139/ssrn.3762348.

Dechow, P. M. and Dichev, I. D. 2002. The Quality of Accruals and Earnings: The Role of Accrual Estimation Errors. The Accounting Review, 77, 35–59. ISSN 0001-4826. Publisher: American Accounting Association.

Dellavigna, S. and Pollet, J. M. 2009. Investor Inattention and Friday Earnings Announcements. The Journal of Finance, 64(2), 709–749. ISSN 1540-6261. doi:10.1111/j.1540-6261.2009.01447.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2009.01447.x.

Drake, M. S., Moon, J. R., Twedt, B. J., and Warren, J. D. 2023. Social media analysts and sell-side analyst research. Review of Accounting Studies, 28(2), 385–420. ISSN 1573-7136. doi:10.1007/s11142-021-09645-1.

Fama, E. F. and MacBeth, J. D. 1973. Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy, 81(3), 607–636.

Francis, J., LaFond, R., Olsson, P. M., and Schipper, K. 2004. Costs of Equity and Earnings Attributes. The Accounting Review, 79(4), 967–1010. ISSN 0001-4826. Publisher: American Accounting Association.

Frankel, R., Kothari, S. P., and Weber, J. 2006. Determinants of the informativeness of analyst research. Journal of Accounting and Economics, 41(1), 29–54. ISSN 0165-4101. doi:10.1016/j.jacceco.2005.10.004.

Frankel, R. and Lee, C. M. 1998. Accounting valuation, market expectation, and cross-sectional stock returns. Journal of Accounting and Economics, 25(3), 283–319. ISSN 0165-4101. doi:10.1016/S0165-4101(98)00026-3.

Gu, F. and Wang, W. 2005. Intangible Assets, Information Complexity, and Analysts Earnings Forecasts. Journal of Business Finance & Accounting, 32(9-10), 1673–1702. ISSN 1468-5957. doi:10.1111/j.0306-686X.2005.00644.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0306-686X.2005.00644.x.

Gu, Z. and Wu, J. S. 2003. Earnings skewness and analyst forecast bias. Journal of Accounting and Economics, 35(1), 5–29. ISSN 0165-4101. doi:10.1016/S0165-4101(02)00095-2.

Hou, K., van Dijk, M. A., and Zhang, Y. 2012. The implied cost of capital: A new approach. Journal of Accounting and Economics, 53(3), 504–526. ISSN 0165-4101. doi:10.1016/j.jacceco.2011.12.001.

Jame, R., Johnston, R., Markov, S., and Wolfe, M. C. 2016. The Value of Crowdsourced Earnings Forecasts. Journal of Accounting Research, 54(4), 1077–1110. ISSN 0021-8456, 1475-679X. doi:10.1111/1475-679X.12121.

Johnson, T. L. and So, E. C. 2018. Asymmetric Trading Costs Prior to Earnings Announcements: Implications for Price Discovery and Returns. Journal of Accounting Research, 56(1), 217–263. ISSN 00218456. doi:10.1111/1475-679X.12189.

Kerr, J., Sadka, G., and Sadka, R. 2020. Illiquidity and Price Informativeness. Management Science, 66(1), 334–351. ISSN 0025-1909. doi:10.1287/mnsc.2018.3154. Publisher: INFORMS.

Kothari, S. P. 2001. Capital markets research in accounting. Journal of Accounting and Economics, page 127.

Kross, W., Ro, B., and Schroeder, D. 1990. Earnings Expectations: The Analysts' Information Advantage. The Accounting Review, 65(2), 461–476. ISSN 0001-4826. Publisher: American Accounting Association.

Lehavy, R., Li, F., and Merkley, K. 2011. The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. The Accounting Review, 86(3), 1087–1115. ISSN 0001-4826, 1558-7967. doi:10.2308/accr.00000043.

Ljungqvist, A., Marston, F., Starks, L. T., Wei, K. D., and Yan, H. 2007. Conflicts of interest in sell-side research and the moderating role of institutional investors. Journal of Financial Economics, 85(2), 420–456. ISSN 0304-405X. doi:10.1016/j.jfineco.2005.12.004.

Lobo, G. J., Song, M., and Stanford, M. 2012. Accruals quality and analyst coverage. Journal of Banking & Finance, 36(2), 497–508. ISSN 0378-4266. doi:10.1016/j.jbankfin.2011.08.006.

Lys, T. and Soo, L. G. 1995. Analysts' Forecast Precision as a Response to Competition. Journal of Accounting, Auditing & Finance, 10(4), 751–765. ISSN 0148-558X. doi: 10.1177/0148558X9501000404. Publisher: SAGE Publications Inc.

Newey, W. K. and West, K. D. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703. ISSN 00129682.

Schafhautle, S. G. and Veenman, D. 2024. Crowdsourced Forecasts and the Market Reaction to Earnings Announcement News. The Accounting Review, 99(2), 421–456. ISSN 0001-4826, 1558-7967. doi:10.2308/TAR-2021-0055.

So, E. C. 2013. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? Journal of Financial Economics, 108(3), 615–640. ISSN 0304-405X. doi: 10.1016/j.jfineco.2013.02.002.

Uddin, A., Tao, X., Chou, C.-C., and Yu, D. 2022. Are missing values important for earnings forecasts? A machine learning perspective. Quantitative Finance, 22(6), 1113–1132. ISSN 1469-7688, 1469-7696. doi:10.1080/14697688.2021.1963825.

van Binsbergen, J. H., Han, X., and Lopez-Lira, A. 2022. Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. The Review of Financial Studies, page hhac085. ISSN 0893-9454. doi:10.1093/rfs/hhac085.

Weiss, D., Naik, P. A., and Tsai, C.-L. 2008. Extracting Forward-Looking Information from Security Prices: A New Approach. The Accounting Review, 83(4), 1101–1124. ISSN 0001-4826. Publisher: American Accounting Association.

Figure 1: Distributions of ML Superiority

This figure presents the distribution of the ML Superiority of our 3,024 ML models. The gray bars show the distribution of the ML Superiority measure for all models. The green bars represent the subset of ML models that use the MAE objective function, the indirect forecasting method, and time-series cross validation. The sample contains forecasts made between June 1990 and May 2020.

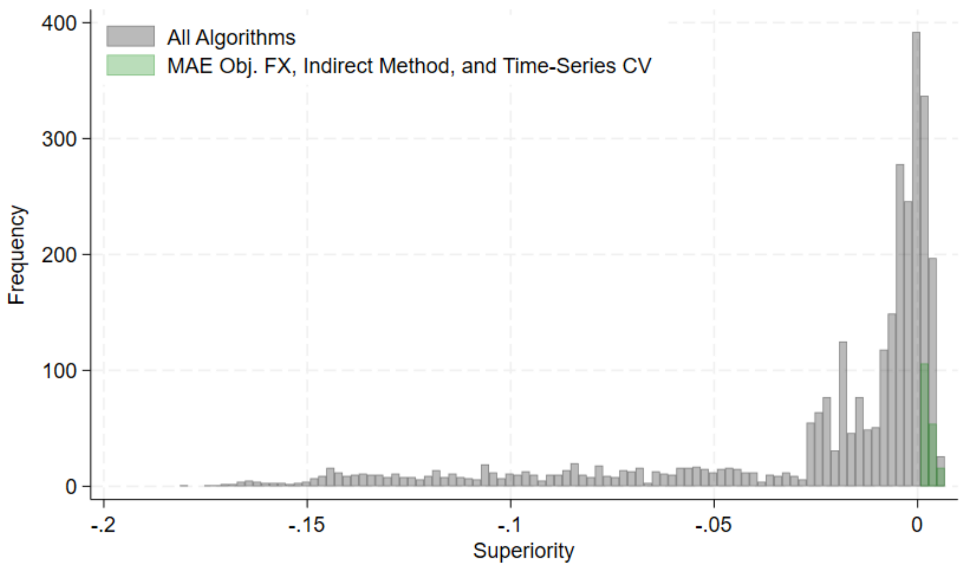
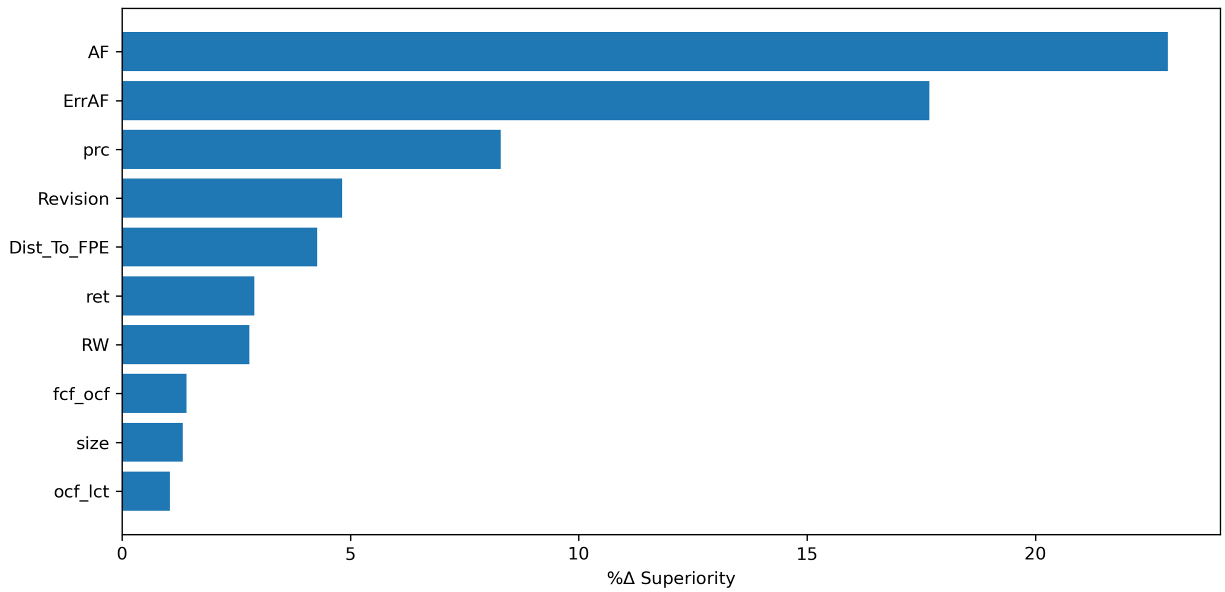
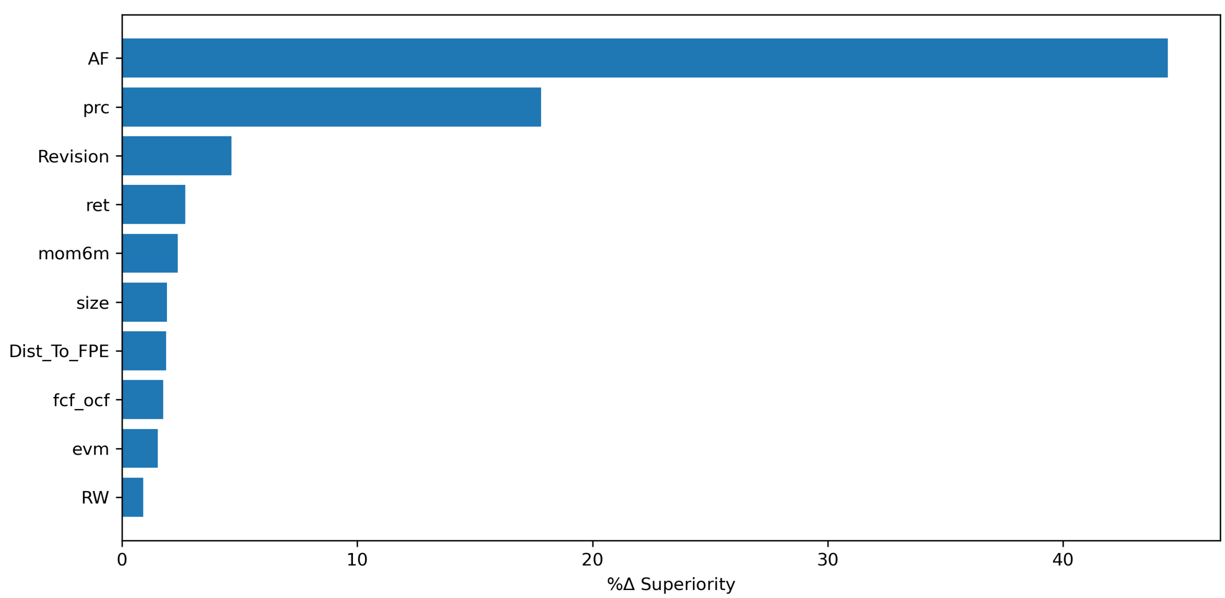


Figure 2: Feature Importance Analysis

This figure presents the top 10 features of the top ML model for the FQ and FY2 forecasts in panels (a) and (b), respectively. The feature importance (see Eq. 3) is based on the percentage change in ML Superiority when a feature is excluded from the predictor set. The reported numbers are in percentage points.



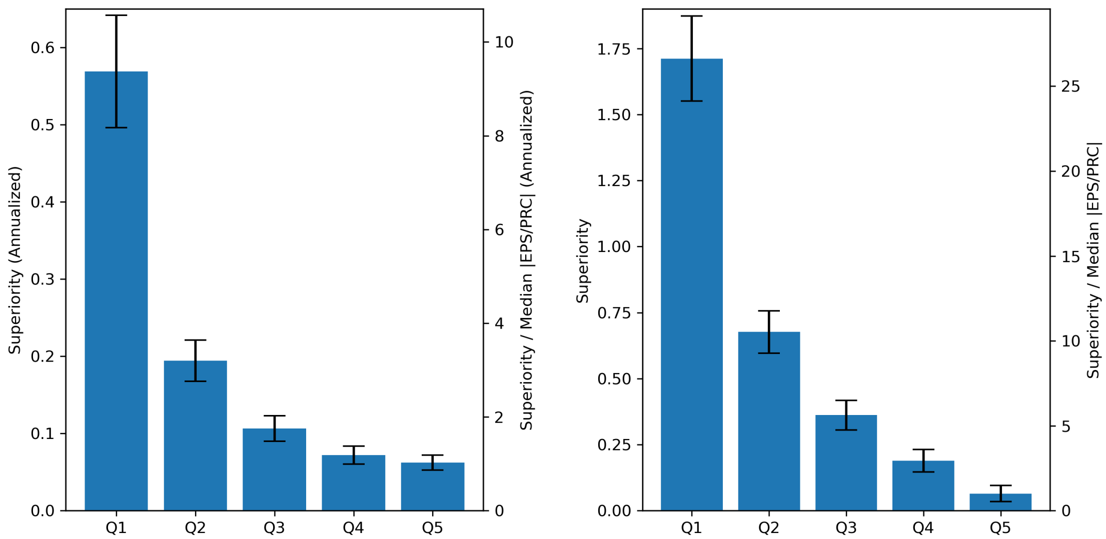
(a) FQ Forecast



(b) FY2 Forecast

Figure 3: ML Superiority by Size Quintiles

This figure presents the ML Superiority by size quintiles for FQ and FY2 in panels (a) and (b), respectively. We cross-sectionally sort the firms into size quintiles based on their market capitalization at the end of each month and calculate the panel average ML Superiority within each quintile. The ML Superiority is in percentage points, and for economic magnitude, we also report it as a percentage of the full-sample median |EPS/PRC| on the right y-axis. Whiskers denote 95% confidence bands. Standard errors of the resulting regression coefficients are two-way clustered at the firm and month level.

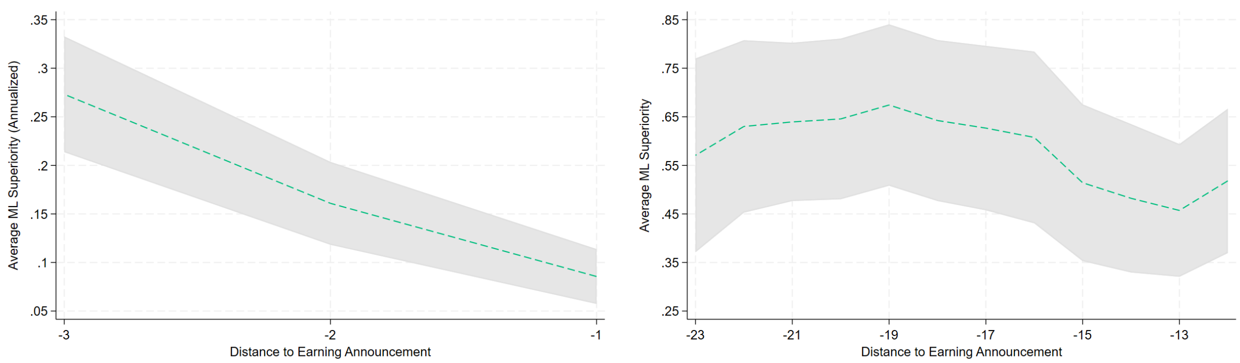


(a) FQ Forecast

(b) FY2 Forecast

Figure 4: Horizon Effect: Distance to Earnings Announcement

Panels (a) and (b) plot the ML Superiority of the statistically optimal forecast for the months prior to the earnings announcement. Panel (a) consists of the annualized ML Superiority of FQ forecasts for the 1 to 3 months prior to the earnings announcement. Panel (b) consists of the ML Superiority of FY2 forecasts for the 23 to 12 months prior to the earnings announcement. The 95% confidence intervals are generated using standard errors clustered by firm and the year of fiscal period end. The ML Superiority is shown in percentage points.

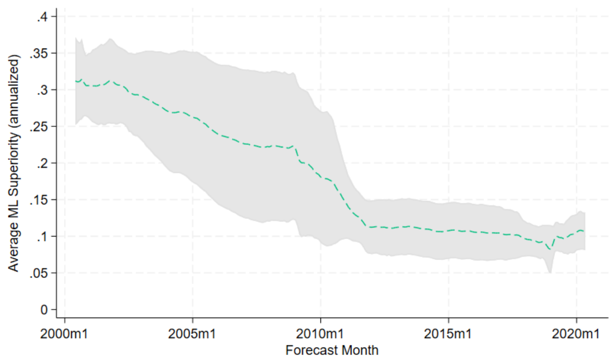


(a) FQ

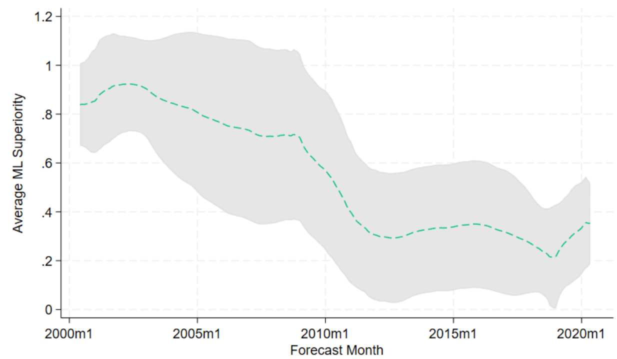
(b) FY2

Figure 5: ML Superiority over Time

Panels (a) and (b) plot the 10-year rolling average (between months $t-119$ and t) of the cross-sectional mean of the ML Superiority for the statistically optimal forecast. The 95% confidence intervals are generated using Newey and West (1987) standard errors with 24 lags. Forecast month t is shown on the x-axis. The ML Superiority is in percentage points.



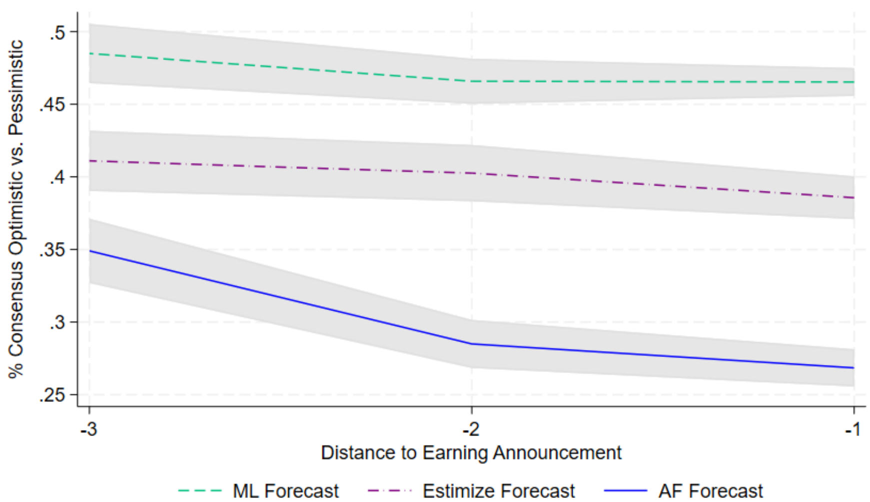
(a) FQ 10 Yr Rolling Average



(b) FY2 10 Yr Rolling Average

Figure 6: The Performance of Estimize Forecasts

This figure presents a comparison of analysts' forecasts, Estimize consensus forecasts, and the statistically optimal forecast (ML). In Panel (a), we plot the percentage of firms with consensus forecasts that exceed realized earnings for the three months leading up to the earnings announcement. Panel (b) plots the superiority measure of the Estimize forecasts, ML forecasts, and a combined forecast (the average of Estimize and analysts' consensus forecasts) for the three months leading up to the earnings announcement. The superiority measure is calculated as the difference in the mean absolute forecast error of a forecast relative to the analysts' consensus forecast, scaled by the price. This analysis focuses on the firm-quarter observations for which the Estimize estimates are available and thus the sample period is for forecasts made from October 2011 to May 2020.

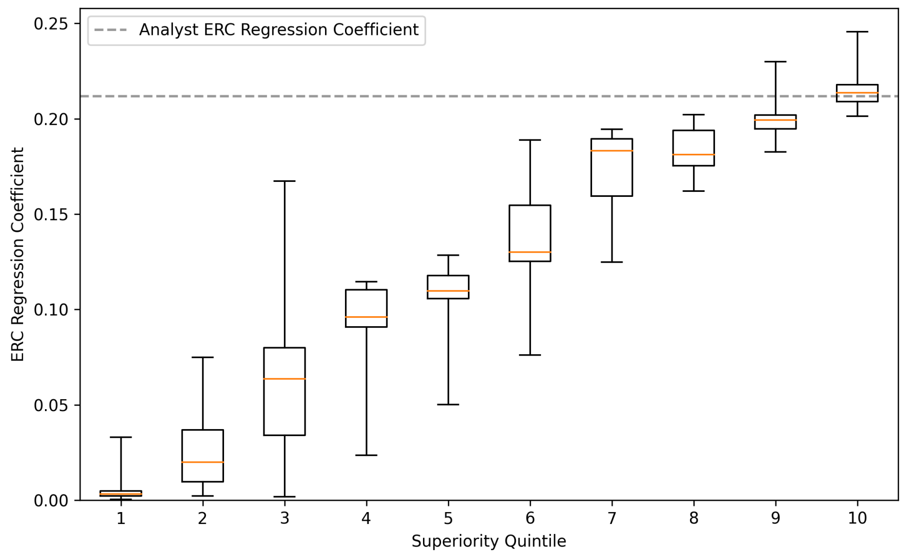


(a) Walkdown of Analyst, ML, and Estimize

(b) Performance Difference between Estimize and ML

Figure 7: Earnings Regression Coefficients

This figure reports the regression coefficients (ERCs) from regressing 3-day size adjusted announcement returns onto standardized unexpected earnings (SUE) following Equation 4. We group FQ forecasts derived from 1512 ML models by their ML Superiority into deciles, and within each decile, we present the box plot of the ERCs. Announcement returns, SUE^{AF}, and SUE^{ML} are trimmed at the 1% and 99% levels for each calendar quarter.



(a) Earnings Response Coefficients

Table 1: Literature Review

Panel A: Superiority of Analysts' Forecasts Relative to Statistical Model Forecasts

Paper	Journal	Analysts' Superiority		Paper's comments on Machine Superiority		Consistency with our paper
		Conclusion	Evaluation Metric	Qualitative	Quantitative	
Bradshaw et al. (2012)	Review of Accounting Studies	Analyst superiority in short horizon (1 year) and RW superiority in longer horizons (2 and 3 year)	Mean of differences in absolute values of forecast error scaled by price (Eq. 3)	Analysts earnings forecasts consistently beat RW earnings forecasts over short windows, for longer forecast horizons, analysts superiority declines, and at certain horizons, analysts forecasts are dominated by RW forecasts.	For forecasts made in the same month as the earnings announcement, analysts forecasts are more accurate than RW forecasts by 282 basis points. In contrast, 11 months prior, analysts superiority is only 35 basis points...analysts forecasts are significantly more accurate than RW forecasts from 12 through 21 months prior... at month 21, analysts superiority is only 3 basis points, and by months 22 and 23, the RW forecast is significantly more accurate than analysts forecasts on average. However, the difference in accuracy is economically trivial, at 7 and 14 basis points respectively....mean analysts superiority from 24 through 35 months prior. Again, analysts superiority falls monotonically, from 66 basis points at 24 months prior to -41 basis points at 35 months prior	Consistent: RW underperforms AF in short horizons. Differ(Potential Reason): RW no longer outperforms analysts forecasts over longer horizons (Bradshaw et al. focuses on an earlier sample period)
So (2013)	Journal of Financial Economics	Machine superiority (1 year horizon)	Mean forecast errors (EPS level not scaled by price)	The median analyst forecast is generally above the median characteristic forecast, consistent with analysts facing incentives to issue optimistic forecasts.	The average characteristic forecast error per share is 0.112 (t-statistic=1.587), which is consistent with the average difference between realized and forecasted earnings being insignificantly different than zero. In contract, the average analyst error is -0.216 (t-statistic=-4.846) which is consistent with the average analyst forecast being optimistic.	Differ (Potential Reason): AF outperforms OLS when no analyst variables are included in predictor set (Evaluation metric: So examines the mean error rather than the accuracy measures such as MSE or MAE.
van Binsbergen et al. (2022)	Review of Financial Studies	Machine superiority (from 1 quarter to 2 year horizon)	Time series average of squared forecast errors (EPS level not scaled by price)	The mean squared errors of the machine learning forecast are smaller than the analysts mean squared errors (all horizons), demonstrating that our forecasts are more accurate than the forecasts provided by analysts.	The realized analysts forecasts errors...increase in the forecast horizon, ranging from 0.028 to 0.384 on average. All of these are statistically significantly different from zero....the time-series averages of the differences between the machine learning forecast and realized earnings are statistically indistinguishable from zero, with an average absolute value of around 0.001 for the quarterly earnings forecasts, 0.027 for the 1-year-ahead forecast, and -0.004 for the 2-year-ahead forecast.	Consistent: AF underperforms RF for FY2 Differ (Potential Reason): AF outperforms RF for FQ (Evaluation metric: van Binsbergen et al. do not scale MSE such that firms with large EPS likely dominate)
de Silva and Thesmar (2022)	WP	Analyst superiority for the short horizon 1 and 2 quarter and 1 year) and machine superiority in longer horizons (from 2 year to 4 year)	Mean squared forecast errors (scaled by price) normalized by mean squared realized EPS (scaled by price) (p.8)	At quarterly and the one-year horizons, we find the combined forecast cannot meaningfully beat the analyst consensus. At longer horizons, however, the combined forecast dominates by a large amount.	At longer horizons, however, the combined forecast dominates by a large amount: 9 and 21 percentage points of realized MSS at the three- and four-year horizons, respectively. However, comparing these results to those in Table 2 shows the improvement relative to pure econometric forecasts is small: around 1 to 3 percentage points	Consistent: we find consistent results when using MSE (with trimmed y-variable) as the objective function Differ (Potential Reason): AF's superiority over ML for the short horizon depends on the objective function (MSE vs. MAE) and whether we use direct or indirect methods.
Ball and Ghysels (2018)	Management Science	Machine superiority (1 quarter horizon)	Median absolute forecast errors (EPS level not scaled by Price) (Eq. 7)	Forecasts are more accurate...than analysts when forecast dispersion is high and when the firm size is smaller. In addition, we find that combining our MIDAS forecasts with analysts forecasts systematically outperforms analysts alone	Our findings are surprisingly sharp as we find that we are always better off combining MIDAScombination forecasts with those of analysts. At the beginning (end) of the target quarter, the combination of model-based and analyst forecasts reduced the forecast error by 21% (11%) relative to analysts forecasts alone. This means that the MIDAS and analyst forecasts feature complementary information.	Consistent: Our findings are consistent with Ball and Ghysels, despite using different models
Chen et al. (2022)	Journal of Accounting Research	Machine superiority (1 year horizon)	Area Under Curve (predicting the direction of changes in EPS)	Model outperforms conventional benchmarks, RW, and analyst forecasts.	The area under the receiver operating characteristics curve ranges from 67.52% to 68.66%, significantly higher than the 50% of a random guess. The annual size-adjusted returns to hedge portfolios formed based on the prediction of our models range from 5.02% to 9.74%.	Cannot compare (Chen et al. does not forecast EPS)
Cao and You (2021)	WP	Machine superiority (1-3 year horizon)	Statistical difference between time series average of mean absolute analyst error and mean absolute machine error (scaled by MVE)	First COMP_NL has significantly lower mean absolute forecast errors than analyst consensus earnings forecasts for all three forecast horizons	COMP_NL has significantly lower mean absolute forecast errors than analyst consensus earnings forecasts for all three forecast horizons (0.0541, 0.0679, and 0.0776 for COMP_NL vs. 0.0588, 0.0742, and 0.0922 for analyst forecasts).	Consistent: Our findings are consistent with Cao and You, despite forecasting EPS instead of Earnings
Uddin et al. (2022)	Quantitative finance	Machine superiority (1 quarter horizon)	Mean squared forecast errors (not scaled by price), mean absolute percentage error and R2 (for the last two, the minimum of the denominator is set to 0.005)	Combined with our data estimation technique, advanced machine learning algorithms provide a superior prediction of firms earnings.	With the help of CMF estimated data, XGBoost beats analysts consensus forecast by 34%.	Hard to compare: Uddin et al. focuses on a smaller dataset of 117 firms for out-of-sample model performance evaluation

Panel B: Specification Choices Directly Related to Model Training

			Hyper-Parameter Tuning				Model Fitting		
Paper	Model(s)	Loss Function	Method	Frequency	Parameter Range	Dataset(s)	Frequency	Forecast Horizon	Dataset Size
Bradshaw et al. (2012)	RW	N/A	N/A	N/A	N/A	N/A	Annual	1-3 years	1 year rolling window
So (2013)	OLS	MSE	N/A	N/A	N/A	N/A	Annual	1 year	1 year rolling window
van Binsbergen et al. (2022)	Modified RF	MSE	Temporal Training-Validation Split	Beginning of sample	Number Trees (1-2000), Depth (1-20), and Fraction of observations to sample (0.01-1)	Training: 1 year window Validation: 1 month window	Monthly	1-3 quarters and 1-2 years	12 month rolling window
de Silva and Thesmar (2022)	EN, RF, GBR	MSE	5 Fold Cross-Validation	Annual	EN: L1 ratio (0.1,0.99) RF: Number Trees (1000), Depth (4-8), Feature Fraction (0.3-1), Minimum samples in leaf (1-5), Samples to split node (2-10) GBT: Number Trees (500-10k), Depth (1-3), Learning rate (0.001-0.01)	5 year rolling window	Annual	1-4 quarters and 1-4 years	5 year rolling window
Ball and Ghysels (2018)	MIDAS (Mixed Data Sampling)	MSE	N/A	N/A	N/A	N/A	Quarterly	1 quarter	40 quarter rolling window
Chen et al. (2022)	RF, GBR	AUC	Temporal Training-Validation Split	Annual	RF: Number Variables (110-120), Number Trees (500-2000),Minimum samples in leaf (1-4), Fraction of observations to sample (0.5) GBR: Number Trees (500-2000), Learning Rate (0.005, 0.01, 0.05), Max Depth (1-4), Minimum samples in leaf (10), Fraction of observations to sample (0.5)	Training: 2 year rolling window Validation: 1 year rolling window	Annual	1 year	1 year rolling window
Cao and You (2021)	OLS, LASSO, Ridge, RF, GBR, ANN	MSE (All but GBR) Huber (GBR)	5 Fold Cross-Validation	ND	Lasso: L1 penalty (1e-3-1e-1) Ridge: L2 penalty (5e1-1e3) RF: Depth (20, 25, 30, 35), Minimum samples in leaf (15, 20, 25, 50) GBR: Depth (1, 3, 5) Minimum samples in leaf (75, 100, 125, 150) ANN: Activation (RELU, Tanh) Alpha (1e-3, 1e-4, 1e-5), Hidden layers ([64,32,16,8], [32,16,8,4], [16,8,4,2], [64,32,16], [32,16,8], [16,8,4], [8,4,2], [63,32], [32,16], [16,8], [8,4], [4,2], [64], [32], [16], [8], [4])	ND	Annual	1-3 years	10 year rolling window
Uddin et al. (2022)	LASSO, XGB, SVR	MSE	10 Fold Cross Validation	Beginning of sample	Lasso: L1 penalty (1e-5, 1e-4,1e-3, 1e-2, 1e-1) XGBoost: Learning rate (1e-5, 1e-4, 1e-3, 1e-2), Regularization (1e-1, 1e-4, 1e-3, 1e-2) SVR: Epsilon (1e-5, 1e-4, 1e-3, 1e-2), Constant (1, 2, 3, 5, 10)	60 Quarters	Quarterly	1 quarter	Expanding window starting with 59 quarters

Panel C: Other Specifications Choices

Paper	Y Variable	Predictor Set	Missing Values Imputation	Data Transformation	
				X Variables	Y Variables
Bradshaw et al. (2012)	EPS	EPS	ND	ND	ND
So (2013)	EPS	Financial statement variables (Earnings per share, loss indicator, pos./neg. accruals per share, asset growth, dividend indicator, book to market, share price, dividend per share)	Drop observation if missing variables	ND	ND
van Binsbergen et al. (2022)	EPS	WRDS Financial Ratio Suite, Realized Earnings, stock price, stock returns, analysts forecasts, and macro mariables	Replace with industry median	Winsorize and then standardize	ND
de Silva and Thesmar (2022)	Earnings/Price	WRDS Financial Ratio Suite, SIC, stock Returns, stock price, 5-year monthly return volatility, analysts forecasts, number of forecasts, total assets, fiscal year dummies. Include 2 lags of variables	Fill Missing with Zero	Trim at 5x IQR and then standardize	Winsorize forecasts of EPS and EPS at 10 times their interquartile range
Ball and Ghysels (2018)	Change in EPS	Change in inventory, A/R, CAPEX, gross margin, and SG&A, abnormal stock return, return volatility, macro economic variables, analysts forecast	ND	ND	ND
Chen et al. (2022)	Indicator for change in EPS (-1 decrease and 1 increase)	10-K XBRL items, lag of 10-K XBRL items, percentage change with lag of 10-K XBRL items, Nissim and Penman (2001) variables, Ou and Penman (1989), analysts forecast, stock price data	Exclude all custom and uncommon tags. Fill missing values with zero	ND	ND
Cao and You (2021)	Earnings	30 Financial statement variables and their first difference. (Historical earnings and major components (8), Income statement items (5), Balance sheet items (16), OCF)	Drop observation if missing key variables and fill some line items with zero. Then drop any observations if missing variables	ND	ND
Uddin et al. (2022)	EPS	Change in Inventory, A/R, CAPEX, gross margin, and SG&A, abnormal stock return, return volatility, stock price, leverage, total assets, dividends, dividend dummy, net income, negative net income dummy, change in total assets, BTM, MVE, accruals, analysts forecasts	Test different methods: Matrix factorization and coupled matrix factorization	ND	ND

Table 2: Nine Sets of Specification Choices Evaluated

This table lists the nine sets of specification choices we evaluate. The complete combination of these specification choices results in 3,024 ML models, with 576 configurations for Lasso, Ridge, Elastic Net, RF, and GBRT and 144 configurations for OLS because it does not require hyperparameter tuning. The sample contains forecasts made between June 1990 and May 2020.

Specification	Choices
Loss Function	MAE MSE MSE with trimmed y-variable (1% and 99% in the train set)
Cross-Validation	Time-series Panel
Estimation Window	Rolling Expanding
Hyper-Parameter Tune Frequency	Beginning Annual
Refit Frequency	Yearly Monthly
Forecasting Approach	Direct Indirect*
Predictor Set	With Analyst Variables Without Analyst Variables
Scaling in Forecasting Step	None Price
Forecast Period	FQ FY2

*The indirect forecasting approach predicts EPS in two steps. First, we forecast analysts' forecast error; second, we adjust analysts' forecasts for the predicted errors to arrive at the final EPS forecasts.

Table 3: The Impact of Specification Choices

This table analyzes the impact of the 9 specifications choice sets summarized in Table 2. For each option within a given specification choice set, we compute the average ML Superiority and machine runtime (in hours) over all the possible combinations of the specifications from the remaining eight sets of specifications. For example, when evaluating the estimation window, we take the average ML Superiority and machine runtime over all models with an expanding window and then do the same for all models with a rolling window. For brevity, we present the results for options with the best and the worst ML Superiority for each choice set and report the resulting difference in the Diff column. The %Diff column reports the values in the Diff column as percentage points of the median $|EPS/PRC|$. ML Superiority is in percentage points, and a higher value means greater outperformance relative to analysts' forecasts.

			Superiority				Run Time (Hrs)		
	Best Sup.	Worst Sup.	Best	Worst	Diff	% Diff	Best	Worst	Diff
Loss Function	MAE	MSE	-0.37	-6.69	6.32	101.1	226.4	77.8	148.6
Cross-Validation	Time-series	Panel	-1.70	-3.17	1.47	23.5	113.0	158.8	-45.8
Estimation Window	Rolling	Expanding	-2.40	-2.67	0.27	4.4	91.4	168.0	-76.6
Refit Frequency	Monthly	Yearly	-2.49	-2.59	0.10	1.7	151.9	107.4	44.5
Param. Frequency.	Annual	Beginning	-2.39	-2.48	0.09	1.5	241.9	29.9	212.0
Forecasting Approach	Indirect	Direct	-1.51	-2.72	1.21	19.4	126.2	133.4	-7.2
Predictor Set	w/ Analyst	w/o Analyst	-2.72	-3.39	0.68	10.8	133.4	129.4	4.0
Scaling in Forecasting Step	PRC	None	-2.30	-2.78	0.48	7.7	126.2	133.2	-7.0
Forecast Period	FY2	FQ	-2.21	-2.86	0.65	12.8	125.8	133.5	-7.7

Table 4: The Best Performing Specifications

This table shows the best-performing choices for specifications related to ML model training for each ML algorithm. Panel A (B) reports the best-performing specifications for the FQ (FY2) forecasts. The associated ML Superiority is shown in percentage points and the associated runtime is shown in hours.

Panel A: FQ Forecasts

Algorithm	Window	Loss Function	Param Freq	Refit Freq	CV Scheme	Superiority	Run Time
OLS	Rolling	MAE		Monthly		0.113	1.16
Lasso	Rolling	MAE	Beginning	Monthly	Time-series	0.113	1.38
Ridge	Rolling	MAE	Beginning	Monthly	Time-series	0.113	1.39
EN	Rolling	MAE	Annual	Monthly	Time-series	0.113	14.89
RF	Rolling	MAE	Annual	Monthly	Time-series	0.116	23.83
GBRT	Rolling	MAE	Annual	Monthly	Time-series	0.201	73.42

Panel B: FY2 Forecasts

Algorithm	Window	Loss Function	Param Freq	Refit Freq	CV Scheme	Superiority	Run Time
OLS	Rolling	MAE		Monthly		0.267	1.05
Lasso	Rolling	MAE	Beginning	Monthly	Time-series	0.300	1.01
Ridge	Rolling	MAE	Beginning	Monthly	Time-series	0.306	1.59
EN	Rolling	MAE	Beginning	Monthly	Time-series	0.309	1.97
RF	Rolling	MAE	Beginning	Monthly	Time-series	0.365	2.47
GBRT	Expanding	MAE	Annual	Monthly	Time-series	0.601	121.35

Table 5: The Role of the Analysts' Forecasts

This table provides the ML Superiority (averaged over all firm-month observations) by ML algorithm using the direct method (w/o analysts), direct method (w/ analysts), and indirect method for FQ and FY2 forecasts. The ML Superiority is shown in percentage points. Standard errors are clustered by firm and the year of fiscal period end. Statistical significance is denoted as ***, **, and * for p<0.10, p<0.05, and p<0.01, respectively.

	(1) N	(2) RW	(3) OLS	(4) LASSO	(5) Ridge	(6) EN	(7) RF	(8) GBRT
FQ Direct w/o Analysts	1080680	-1.868***	-1.722***	-1.721***	-1.724***	-1.731***	-2.629***	-1.132***
		(-19.34)	(-25.12)	(-25.14)	(-25.10)	(-25.56)	(-25.75)	(-18.42)
FY2 Direct w/o Analysts	973417	-1.211***	-0.460***	-0.434***	-0.461***	-0.420***	-0.558***	0.205**
		(-7.25)	(-4.51)	(-4.50)	(-4.52)	(-4.51)	(-6.14)	(2.39)
FQ Direct w/ Analysts	1080680		-0.113***	-0.117***	-0.120***	-0.113***	-2.369***	0.0921***
			(-4.81)	(-5.04)	(-5.03)	(-5.01)	(-23.06)	(3.51)
FY2 Direct w/ Analysts	973417		0.167**	0.165**	0.168**	0.168**	-0.452***	0.536***
			(2.41)	(2.38)	(2.43)	(2.44)	(-4.84)	(7.20)
FQ Indirect	1080680		0.113***	0.113***	0.113***	0.113***	0.116***	0.201***
			(7.52)	(7.48)	(7.47)	(7.76)	(7.95)	(8.83)
FY2 Indirect	973417		0.267***	0.300***	0.306***	0.309***	0.365***	0.601***
			(4.73)	(5.70)	(5.75)	(5.82)	(6.23)	(7.66)

Table 6: Cross-Sectional Variation in ML Superiority

This table presents Fama-MacBeth regressions of the ML Superiority of the top ML model on size, count of business segments, institutional ownership, idiosyncratic volatility, net external financing, bid-ask spread, R&D, and accrual volatility. All independent variables are the normalized rank (i.e., the rank based on the variable of interest scaled by the number of stocks in the cross-section) between 0 and 1. The ML Superiority is shown in percentage points. Standard errors are computed based on Newey and West (1987) with 24 monthly lags. Statistical significance is denoted as ***, **, and * for p<0.10, p<0.05, and p<0.01, respectively.

	FQ			FY2		
	(1)	(2)	(3)	(4)	(5)	(6)
1/Size	0.419***	0.456***	0.359***	1.265***	1.713***	1.032***
	(4.0)	(4.3)	(4.3)	(5.3)	(7.5)	(6.4)
Count of Business Segments	0.138***	0.180***	0.108***	0.407***	0.576***	0.243***
	(4.2)	(3.7)	(2.9)	(4.0)	(4.8)	(3.0)
Institutional Ownership	-0.056*	-0.072*	-0.037	-0.189	-0.417***	-0.209**
	(-1.7)	(-1.7)	(-0.9)	(-1.6)	(-3.7)	(-2.1)
Idiosyncratic Volatility	0.219***	0.218***	0.212***	0.844***	0.440***	0.705***
	(3.8)	(3.4)	(6.0)	(5.7)	(4.0)	(5.8)
Net external financing	-0.019	-0.081***	0.035	0.499***	0.536***	0.453***
	(-1.0)	(-3.1)	(1.6)	(7.9)	(6.8)	(6.0)
Bid-Ask Spread	0.123*	0.096	0.122**	0.217*	0.283**	0.172**
	(1.9)	(1.5)	(2.2)	(1.9)	(2.6)	(2.1)
R&D		0.046			0.111	
		(1.1)			(0.7)	
Accrual Volatility			0.016			0.173***
			(0.4)			(3.2)
Observations	869025	482985	513376	793705	436214	477771

Table 7: Investors' Expectations Implied by ERC Tests

This table shows the bivariate regression results of size adjusted 3-day earnings announcement returns ($R_{d-1,d+1}$) on $\text{SUE}^{AF}_{t-1,d}$ and $\text{SUE}^{ML}_{t-1,d}$, as defined in Eq. (5). $Bias$ is the difference between the statistically optimal forecast and the analyst ($Bias^{AF} \equiv \frac{\text{ML}_{t-1}-\text{AF}_{t-1}}{P_{t-1}}$). Excess weight is calculated following Eq. (8). IO and Size are the normalized cross-sectional rank of 13F institutional ownership and the market value of equity known at month $t-1$. We trim announcement returns, SUE^{AF}, $SUE^{Estimize}$, and SUE^{ML} at the 1% and 99% level each calendar quarter. Standard Errors are clustered by firm and the calendar year-quarter. Statistical significance is denoted as ***, **, and * for p<0.10, p<0.05, and p<0.01, respectively.

	(1)	(2)	(3)	(4)	(5)
SUE^{ML}	0.245***		0.250***	0.185***	0.176***
	(6.81)		(6.53)	(7.02)	(9.25)
SUE^{AF}		0.212***			
		(6.84)			
Bias^{AF}			0.096***	0.093***	0.102***
			(6.03)	(6.05)	(4.98)
SUE^{ML}*Size				0.221***	
				(3.48)	
Bias^{AF}*Size				-0.097**	
				(-2.47)	
Size				0.834***	
				(10.55)	
SUE^{ML}*IO					0.200***
					(2.84)
Bias^{AF}*IO					-0.071
					(-1.45)
IO					1.083***
					(13.34)
Const.	-0.202***	-0.123***	-0.164***	-0.606***	-0.723***
	(-6.48)	(-3.88)	(-5.29)	(-9.99)	(-12.85)
R-Squared (%)	2.352	2.213	2.483	2.693	2.775
# Obs.	328398	328398	328398	328398	327081
Excess Weight on AF			0.382		
Δ Excess Weight Across IQR				-0.228	-0.223

Table 8: Investors' Expectations Implied by ERC Tests (Estimize Sample)

This table shows the bivariate regressions of size adjusted 3-day earnings announcement returns ($R_{d-1,d+1}$) on $\text{SUE}^{ML}_{t-1,d}$ and various bias measures. $Bias^{Estimize}$ and $Bias^{Combined}$ are defined similar to $Bias^{AF}$ but replace AF with the Estimize forecast and the combined forecast, respectively. The Estimize data begins in October 2011, but we extend the sample in Column 4 to January 2010 to allow a control group and pre-coverage period. *Post* is an indicator equal to one for periods following the initial Estimize forecast for a given firm. Time is a continuous variable that increases linearly by calendar quarter over the course of the Estimize forecast sample, starting at 0 in Q1 2010 and ending at 1 in Q2 2020. We trim announcement returns, SUE^{AF}, $SUE^{Estimize}$, and SUE^{ML} at the 1% and 99% level each calendar quarter. Standard Errors are clustered by firm and the calendar year-quarter. Statistical significance is denoted as ***, **, and * for p<0.10, p<0.05, and p<0.01, respectively.

	(1)	(2)	(3)	(4)
SUE^{ML}	0.677***	0.675***	0.676***	0.485***
	(4.73)	(4.72)	(4.71)	(7.31)
Bias^{AF}	0.014			0.210***
	(0.12)			(4.29)
$\text{Bias}^{Estimize}$		-0.002		
		(-0.20)		
$\text{Bias}^{Combine}$			-0.004	
			(-0.17)	
SUE^{ML}*Post				0.199***
				(4.77)
Bias^{AF}*Post				-0.055
				(-0.81)
Post				0.136
				(1.37)
SUE^{ML}*Time				-0.112
				(-1.24)
Bias^{AF}*Time				-0.219*
				(-1.68)
Const.	-0.124	-0.130	-0.129	-0.275***
	(-1.14)	(-0.99)	(-0.99)	(-4.75)
R-Squared (%)	4.082	4.082	4.082	4.572
# Obs.	27758	27758	27758	100890
Excess Weight	0.021	-0.003	-0.006	
ΔExcess Weight Post Estimize Coverage				-0.207

Table 9: Investors' Expectations Implied by the Return Predictability of Analysts' Biases

This table examines the return predictability of analysts' biases in four non-overlapping subsamples. We form monthly portfolios by first dividing stocks in the cross section into four groups based on the cross-sectional medians of firm size and institutional ownership. Then, within each group we divide the firms into quintiles based on analysts' bias and compute the value weighted quintile portfolio returns. Finally, we report the average excess return, FF5 alpha, and q^5alpha of the long short Q5-Q1 portfolio for each group. Panels A and B show results based on biases in analysts' FQ forecasts and FY2 forecasts, respectively. Statistical significance is denoted as ***, **, and * for p<0.10, p<0.05, and p<0.01, respectively. Standard errors of the test statistics are computed based on Newey and West (1987) with 12 lags.

Panel A: FQ

	Small Low IO	Small High IO	Large Low IO	Large High IO
Q5-Q1 Excess Return	0.872***	0.872***	0.872***	0.872***
	(5.23)	(5.23)	(5.23)	(5.23)
Q5-Q1 FF5 Alpha	0.985***	0.985***	0.985***	0.985***
	(5.05)	(5.05)	(5.05)	(5.05)
Q5-Q1 q^5 Alpha	0.623***	0.623***	0.623***	0.623***
	(2.98)	(2.98)	(2.98)	(2.98)

Panel B: FY2

	Small Low IO	Small High IO	Large Low IO	Large High IO
Q5-Q1 Excess Return	1.155***	0.545*	0.750**	0.238
	(4.07)	(1.80)	(2.24)	(0.96)
Q5-Q1 FF5 Alpha	1.257***	0.937***	0.822***	0.511**
	(4.57)	(3.39)	(3.26)	(2.13)
Q5-Q1 q^5 Alpha	0.820***	0.396	0.368	0.007
	(2.94)	(1.60)	(1.39)	(0.03)

A. Appendix

Table A1: WRDS Financial Ratio Variables

This table provides the definitions of WRDS Financial Ratio Variables. Following van Binsbergen et al. (2022), we exclude Forward P/E to 1-year Growth (PEG) ratio, Forward P/E to Long-term Growth (PEG) ratio, Price/Operating Earnings (Basic, Excl. Extraordinary Income), and Price/Operating Earnings (Diluted, Excl. Extraordinary Income) from the WRDS Financial Suite Ratios due to the large number of missing observations.

Acronym	Definition	Acronym	Definition
accrual	Accruals/Average Assets	int_totdebt	Interest/Average Total Debt
adv_sale	Advertising Expenses/Sales	inv_turn	Inventory Turnover
aftret_eq	After-tax Return on Average Common Equity	invt_act	Inventory/Current Assets
aftret_equity	After-tax Return on Total Stockholders Equity	lt_debt	Long-term Debt/Total Liabilities
aftret_invcapx	After-tax Return on Invested Capital	lt_ppent	Total Liabilities/Total Tangible Assets
at_turn	Asset Turnover	npm	Net Profit Margin
bm	Book/Market	ocf_lct	Operating Cash Flow/Current Liabilities
capei	Shiller's Cyclically Adjusted P/E Ratio	opmad	Operating Profit Margin After Depreciation
capital_ratio	Capitalization Ratio	opmbd	Operating Profit Margin Before Depreciation
cash_conversion	Cash Conversion Cycle (Days)	pay_turn	Payables Turnover
cash_debt	Cash Flow/Total Debt	pcf	Price/Cash Flow
cash_lt	Cash Balance/Total Liabilities	pe_exi	P/E (Diluted, Excl. EI)
cash_ratio	Cash Ratio	pe_inc	P/E (Diluted, Incl. EI)
cfm	Cash Flow Margin	peg_trailing	Trailing P/E to Growth (PEG) ratio
curr_debt	Current Liabilities/Total Liabilities	pretret_earnat	Pre-tax Return on Total Earning Assets
curr_ratio	Current Ratio	pretret_noa	Pre-tax Return on Net Operating Assets
de_ratio	Total Debt/Total Equity	profit_lct	Profit Before Depreciation/Current Liabilities
debt_assets	Total Debt/Total Assets	ps	Price/Sales
debt_at	Total Debt/Total Assets	ptb	Price/Book
debt_capital	Total Debt/Total Capital	ptpm	Pre-Tax Profit margin
debt_ebitda	Total Debt/EBITDA	quick_ratio	Quick Ratio
debt_invcap	Long-term Debt/Invested Capital	rd_sale	Research and Development/Sales
divyield	Dividend Yield	rect_act	Receivables/Current Assets
dltt_be	Long-term Debt/Book Equity	rect_turn	Receivables Turnover
dpr	Dividend Payout Ratio	roa	Return on Assets
efftax	Effective Tax Rate	roce	Return on Capital Employed
equity_invcap	Common Equity/Invested Capital	roe	Return on Equity
evm	Enterprise Value Multiple	sale_equity	Sales/Stockholders Equity
fcf_ocf	Free Cash Flow/Operating Cash Flow	sale_invcap	Sales/Invested Capital
gpm	Gross Profit Margin	sale_nwc	Sales/Working Capital
gprof	Gross Profit/Total Assets	short_debt	Short-Term Debt/Total Debt
int_debt	Interest/Average Long-term Debt	staff_sale	Labor Expenses/Sales
intcov	After-tax Interest Coverage	totdebt_invcap	Total Debt/Invested Capital
intcov_ratio	Interest Coverage Ratio		

Table A2: Other Variables

This table provides the definitions of the other variables used in generating our ML predictions that are not included in the WRDS Financial Ratio Suite. EPS and ErrAF are target variables, while all other variables are additional predictors.

Acronym	Definition
EPS (FY2/FQ)	Realized Earnings per Share
ErrAF (FY2/FQ)	Realized EPS-Analysts' forecast as of current month
medest2	Analysts' consensus forecast for FY2 horizon
medestqtr	Analysts' consensus forecast for FQ horizon
ibes_earnings_ann	Most recently realized annual earnings as of current month
ibes_earnings_qtr	Most recently realized quarterly earnings as of current month
last_F2ana_fe_med	Most recently realized FY2 horizon analysts' forecast error as of current month
last_Fqtrana_fe_med	Most recently realized FQ horizon analysts' forecast error as of current month
rev_FY2_3m	Revision of analysts' FY2 horizon forecast between current month and 3 months prior
rev_FYqtr_3m	Revision of analysts' FQ horizon forecast between current month and 3 months prior
dist2	Distance between FY2 fiscal period end and current month
distqtr	Distance between FQ fiscal period end and current month
ret	Stock Return
prc	Stock Price
size	LN(Market Capitalization)
mom6m	6 month momentum
indmom	Industry weighted 6 month momentum

Table A3: Sample Construction

This table describes how we arrive at our final sample and shows the effect of each data filter. We identify abnormal forecast period end dates in I/B/E/S (fpedats) following Bordalo et al. (2019), who provide supplementary information and replication codes for the procedure in their online appendix.

Panel A: FQ

Data Filter	Firm-Months
CRSP US common stocks merged with Compustat monthly observations Jan. 1985-Dec. 2020	**2,145,510**
Less: missing analysts' forecasts for all forecast horizons (FY1, FY2, FQ1, FQ2, FQ3)	-669,267
Less: missing FQ analysts' forecasts	-114,483
Less: missing most recently realized quarterly earnings	-16,525
Less: missing stock price, return, market capitalization, the two momentum variables, and price-to-sales	-38,045
Less: abnormal forecast period end	-37,732
Less: announcement month less than or equal to current month	-12,018
Final Dataset	**1,257,440**

Panel B: FY2

Data Filter	Firm-Months
CRSP US common stocks merged with Compustat monthly observations Jan. 1983-Dec. 2020	**2,261,904**
Less: missing analysts' forecasts for all forecast horizons (FY1, FY2, FQ1, FQ2, FQ3)	-727,524
Less: missing FY2 analysts' forecasts	-77,218
Less: missing most recently realized annual earnings	-69,629
Less: missing stock price, return, market capitalization, the two momentum variables, and price-to-sales	-21,069
Less: abnormal forecast period end	-17,757
Less: announcement month less than or equal to current month	-2
Final Dataset	**1,348,705**

Table A4: Variables used in ML Superiority Analysis

This table provides the definition of the variables used in the analysis of the cross-sectional variation in ML Superiority.

Variable	Definition	Category	Citation
Size	Ln(Market Value of Equity)	Information Uncertainty	Kross et al. (1990); Lys and Soo (1995); Das et al. (1998); Frankel et al. (2006); Lehavy et al. (2011)
Idiosyncratic Volatility	Standard deviation of residuals from CAPM regressions using the past year of daily data.		
Count of Business Segments	Count of the firms' business segments*	Firm Complexity	Amir et al. (2003); Gu and Wang (2005); Frankel et al. (2006); Lehavy et al. (2011)
R&D	Research and Development Expense scaled by market value*		
Bid-Ask Spread	Effective bid ask spread based on Corwin-Schulz scaled by stock price.+	Price Informativeness	Kerr et al. (2020)
Institutional Ownership	Percent shares held by institutional owners**	Analysts' Incentives	Frankel et al. (2006); Ljungqvist et al. (2007); Lehavy et al. (2011); Bradshaw et al. (2016)
Net external financing	Sale of common stock (sstk) minus dividends (dv) minus purchase of common stock (prstkc) plus long-term debt issuance (dltis) minus long-term debt reductions (dltr). Scaled by total assets (at).*+		
Accrual Quality	Estimate a regression for each year and industry of total current accruals on the current value and one year lag and lead of cash flow from operations, change in revenues, and gross value of PPE. Save the regression residuals and replace with missing if there are not at least 20 observations per year and industry. Calculate accrual quality (AQ) as the standard deviation of residuals over 4 years. If more than one observation is missing set AQ to missing.*+	Earnings Management	Dechow and Dichev (2002); Francis et al. (2004); Lobo et al. (2012)

*Data as of the most recently realized fiscal year end.
**Data as of the most recently realized fiscal quarter end.
+Data obtained from website associated with Chen and Zimmermann (2022).