Fundamental Analysis via Machine Learning

Kai Cao
School of Business and Management
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
kcaoab@connect.ust.hk

Haifeng You*
School of Business and Management
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
achy@ust.hk

This draft: September 20, 2020

Abstract:

We examine the efficacy of machine learning in one of the most important tasks in fundamental analysis, forecasting corporate earnings. Our analyses show that machine learning models, especially those that accommodate nonlinearities, generate significantly more accurate and informative forecasts than a host of state-of-the-art earnings prediction models in the extant literature. Further analysis suggests that machine learning models uncover economically sensible relationships between historical financial information and future earnings. We also find that the new information uncovered by machine learning models is of considerable economic significance to investors. The new information component of the machine learning-based forecasts is significantly associated with future stock returns. Stocks in the quintiles with the most favorable new information outperform those in the least favorable quintiles by approximately 70 bps per month, suggesting that the new information is not well understood by investors. Finally, insights from machine learning models are useful for improving the extant models.

JEL Classification: G10, G11, G14, G17, M40, M41

Keywords: machine learning, earnings forecasts, fundamental analysis, equity valuation, market efficiency

* We thank Utpal Bhattacharya, Sean Cao, Kevin Chen, Peter Chen, Zhihong Chen, Allen Huang, Mingyi Hung, Arthur Morris, Kevin Li, Xinlei Li, Richard Sloan, Rongfei Wang, Amy Zang, Yue Zheng, and workshop participants at HKUST for their helpful comments and suggestions. Research funding support from Hong Kong RGC project no. T31-604/18-N is highly appreciated.

Fundamental Analysis via Machine Learning

Abstract:

We examine the efficacy of machine learning in one of the most important tasks in fundamental analysis, forecasting corporate earnings. Our analyses show that machine learning models, especially those that accommodate nonlinearities, generate significantly more accurate and informative forecasts than a host of state-of-the-art earnings prediction models in the extant literature. Further analysis suggests that machine learning models uncover economically sensible relationships between historical financial information and future earnings. We also find that the new information uncovered by machine learning models is of considerable economic significance to investors. The new information component of the machine learning-based forecasts is significantly associated with future stock returns. Stocks in the quintiles with the most favorable new information outperform those in the least favorable quintiles by approximately 70 bps per month, suggesting that the new information is not well understood by investors. Finally, insights from machine learning models are useful for improving the extant models.

JEL Classification: G10, G11, G14, G17, M40, M41

Keywords: machine learning, earnings forecasts, fundamental analysis, equity valuation, market efficiency

1. Introduction

Expectation about future earnings is an important determinant of security values (e.g., Ohlson, 1995; Ohlson and Juettner-Nauroth, 2005). Accurate earnings forecasts not only allow investors to make more informed investment decisions, but also facilitate the efficient allocation of capital in society (Hayek, 1945; Arrow, 1964). A large body of research has attempted to develop accurate earnings forecasting models but achieved little success (e.g., Ball and Watts, 1972; Albrechet, Lookabill, and McKeown, 1977; Watts and Leftwich, 1977; Fama and French, 2006; Hou, van Dijk, and Zhang, 2012; So, 2013; Li and Mohanram, 2014). As Monahan (2018) summarizes in his survey, *the extant models are too inaccurate,* and *the extant results lead to seemingly absurd conclusions regarding best practice*. This study aims to examine the efficacy of machine learning in forecasting corporate earnings and to shed light on the usefulness of financial statement information for earnings prediction and equity valuation.

Machine learning algorithms offer several advantages in forecasting earnings. First, machine learning algorithms can efficiently handle high dimensional data. The generation of corporate earnings (or losses) is a complex process involving numerous business transactions. The effects of these transactions, as summarized in financial statement line items, often have different implications for future earnings (e.g., cash sales vs. credit sales). However, for tractability, extant earnings prediction models only focus on highly aggregated measures, such as bottom line earnings and book value of equity, while neglecting potentially rich information in financial statement line items (e.g., Fairfield, Sweeney, and Yohn, 1996; Chen, Miao, and Shevlin, 2015). By accommodating a large set of financial statement line items, machine learning

algorithms can potentially better model the differential effects of these items and generate more accurate and informative earnings forecasts.

Second, in contrast to traditional linear models, machine learning algorithms can accommodate more complex relationships between financial statement line items and future earnings. Economic theories and empirical evidence suggest the existence of nonlinear relationships between financial statement line items and future earnings. For example, the law of diminishing returns predicts a nonlinear relationship between capital investment and future earnings. Prior literature also shows that the relationship between current and future earnings is nonlinear (e.g., Freeman and Tse, 1992; Chen and Zhang, 2003) and varies with other financial metrics, such as firm size and capital intensity (e.g., Lev, 1983; Baginski, Lorek, Willinger, and Branson, 1999). Machine learning algorithms based on decision trees and neural networks are rather flexible in modeling nonlinear relationships and interaction effects, providing another advantage in forecasting earnings.

The advantage associated with the ability to accommodate high dimensional data and nonlinearity does come at a cost. More flexible/complex models are also more susceptible to in-sample overfitting, which can lead to poor out-of-sample performance. To mitigate this problem, machine learning algorithms often adopt "parameter regularization" to penalize overly complex models. Furthermore, cross-validation is often performed to evaluate and select models based on their performance on the holdout validation samples. Given these advantages and pitfalls, whether machine learning algorithms perform better in earnings prediction tasks remains an empirical question. We test this research question by comparing earnings forecasts generated using a comprehensive list of machine learning models with those generated using several state-of-the-art earnings prediction models developed in the finance and accounting literature.

Specifically, we examine three linear machine learning models, namely ordinary least squares regression (OLS), least absolute shrinkage and selection operator (LASSO), and Ridge regression, and three nonlinear machine learning models of which two are based on decision trees (namely random forest (RF) and gradient boosting regression (GBR)) and one is based on artificial neural networks (ANNs). We supply these algorithms with 56 input variables (or predictors/features), including 28 major financial statement line items and their respective first-order differences to predict the target variable, that is, future earnings. We compare the out-of-sample earnings forecasts generated using the six machine learning models with the forecasts generated using the benchmark random walk (RW) model and five models proposed in the literature, which are the (first-order) autoregressive model (AR); two models (HVZ and SO) developed by Hou, van Dijk, and Zhang (2012) and So (2013), respectively; and the earnings persistence (EP) model and the residual income (RI) model proposed by Li and Mohanram (2014). In addition, we examine several composite forecasts, namely COMP_EXT, COMP_LR, COMP_NL, and COMP_ML, which are the mean forecasts of the five extant models, the three linear machine learning models, the three nonlinear machine learning models, and all six machine learning models, respectively.

We compare the accuracy and information content of the above earnings forecasts for 134,154 firm-year observations over the period from 1975 to 2019. Consistent with the literature, we find that the earnings forecasts generated using the extant models are not consistently more accurate than that generated using the naïve RW model. For example, although the mean absolute forecast error of the most accurate extant model, the RI model, is approximately 3.07% lower than that of the RW model, its median absolute forecast error is higher. The earnings forecasts generated using the machine learning models are generally more accurate. The three

linear models are approximately 5.83%–6.31% more accurate than the RW model. The mean absolute forecast errors of the three nonlinear machine learning models, namely ANN, RF, and GBR, are approximately 6.67%, 8.64%, and 8.86% lower than that of the RW model, respectively.

Composite forecasts that combine predictions from different models are more accurate. The mean absolute forecast errors of the composite forecasts obtained with COMP_EXT, COMP_LR, and COMP_NL are approximately 3.58%, 6.16%, and 9.87% lower than that of the RW model, respectively. Moreover, cross-sectional analyses indicate that the nonlinear machine learning models lead to even greater improvements in forecast accuracy in the case of firms with more difficult-to-forecast earnings. For example, among the top quintiles of firms with the most volatile return on assets (ROA) and those with the highest magnitude of total accruals, COMP_NL improves the forecast accuracy relative to the RW model by 15.21% and 17.71%, respectively.

Further analyses suggest that the improvement in the performance of the nonlinear machine learning models can be attributed to at least the following reasons. First, the models can identify a set of economically sensible predictors, even without explicit theoretical guidance. For example, in addition to current earnings and operating cash flows, income tax expenses and their first-order differences are among the most important predictors, which prior studies find are closely related to tax-based earnings and contain important information about earnings quality and persistence (e.g., Lev and Nissim, 2004; Hanlon, 2005; Thomas and Zhang, 2014). Second, these models are able to detect subtle, yet sensible nonlinear relationships and interaction effects. For example, both the RF and GBR models correctly predict that loss is less persistent than profits, which is assumed in most extant models. Furthermore, these models can detect and use

the interaction effects between economically linked variables, such as cost of goods sold and inventories, properties, plants and equipment and depreciation and amortization in forecasting future earnings.

Next, we evaluate the information content of various models by investigating their (out-of-sample) predictive power with respect to future changes in actual earnings. Our analyses show that the machine learning models have greater predictive power than the extant models. Predicted earnings changes (FECH) based on COMP_NL explain 18.57% of the variation in future actual earnings changes (ECH), which is not only higher than that of COMP_EXT (12.73%), but also higher than that of COMP_LR (15.09%). Additional analysis shows that the machine learning models subsume the information content of all extant models, except that of the SO model, which incorporates forward-looking stock price information.

The above results suggest that machine learning models, especially nonlinear models, help to uncover new information neglected by the extant models. We test the economic significance of the results by examining whether the new information uncovered by the machine learning models is useful for making investment decisions. We orthogonalize the machine-learning-based forecasts against the current earnings and the forecasts generated by the five extant models and use the resulting residuals to measure the new information obtained from the machine-learning-based forecasts. Our analyses show that these residual forecasts have significant predictive power with respect to future stock returns. For example, the top 20% of the stocks with the most favorable new information in COMP_NL outperform those with the least favorable new information by approximately 41–77 bps per month on the risk-adjusted basis. Additional analysis also suggests that the new information in machine learning forecasts predicts

analyst earnings forecast errors, indicating that both investors and analysts underreact to the predictable new information uncovered by the machine learning algorithms.

The contributions of our paper are as follows. First, we contribute to the extensive literature on financial statement analysis and value relevance of accounting information. The analysis of historical financial statement data to forecast future earnings is a central task in fundamental analysis and security valuation. The literature on earnings prediction is extensive, but the extant models fail to consistently outperform the naïve RW model. As summarized in Monahan (2018), these results contradict one of the fundamental tenets of financial statement analysis that financial ratios and financial statement line items should be incrementally informative about future earnings beyond bottom line earnings. Our paper sheds light on this puzzle by showing that the inability of extant models to handle high dimensional data with nonlinear relationships, together with the omission of several important financial statement line items, contributes to these results. Furthermore, by demonstrating that significant new value-relevant information can be uncovered from financial statement line items, our paper provides unambiguous evidence of the usefulness of financial statements and fundamental analysis for equity valuation and investment decision-making.

Second, our findings are relevant to investors. Earnings forecasts are critical inputs in determining the value of securities by investors. However, the increasingly complex financial reporting (e.g. Dyer, Lang, and Stice-Lawrence, 2017) imposes significant costs for investors to understand complicated financial statements and to make accurate earnings forecasts. By developing a model that can produce significantly more accurate and informative earnings forecasts for a large sample of firms at a low cost, our paper is clearly of considerable interest to investors who aspire to identify mispriced securities through equity valuation. Accurate earnings

forecasts are of social value as well. When investors make investment decisions based on equity valuation, they tend to allocate more capital to firms with better predicted future fundamental performance. Thus, more accurate earnings forecasts facilitate efficient resource allocation (to more productive firms) in society.

Finally, we contribute to the burgeoning literature on the application of machine learning in the finance and accounting research domains. For example, several recent studies adopt machine learning algorithms for asset pricing (e.g., Rapach, Strauss, Tu, and Zhou, 2019; Gu, Kelly, and Xiu, 2020) and show that these algorithms outperform traditional models in predicting future returns and asset risk premia. In another study, Bao, Ke, Li, Yu, and Zhang (2020) adopt the AdaBoost algorithm to predict accounting frauds and demonstrate that their model substantially outperforms the model developed by Dechow, Ge, Larson, and Sloan (2011). Ding, Lev, Peng, Sun, and Vasarhelyi (2020) run "horseraces" of four machine learning algorithms to predict insurance losses and find that machine learning estimates are superior to managers' estimates. We contribute to the literature by providing a comparative study on one of the most important research questions in fundamental analysis and equity valuation, i.e., earnings forecasting, and by documenting the significant benefits of adopting machine learning algorithms in this setting.

2. Related Literature and Extant Earnings Forecasting Models

As Monahan (2018) summarizes in his comprehensive survey of the earnings forecasting literature, early research mostly adopts the time-series approach to forecast future earnings (e.g., Ball and Watts, 1972; Albrecht et al., 1977; Watts and Leftwich, 1977). Overall, their results

suggest that the simple RW model, which predicts that expected future earnings are equal to current earnings, generates more accurate out-of-sample forecasts than more sophisticated autoregressive integrated moving average (ARIMA) models (e.g., Brown, 1993; Kothari, 2001). Subsequent research demonstrates that the RW model performs well even when compared with analyst forecasts. For example, Bradshaw, Drake, Myers, and Myers (2012) find that analysts' earnings forecasts are *not* economically more accurate than those obtained using the naïve RW model, and those for horizons longer than one year are consistently less accurate than those obtained using the naïve RW model. The superiority and simplicity of the RW model make it a natural benchmark to evaluate other earnings forecasting models.

There are several potential reasons for the poor out-of-sample performance of sophisticated ARIMA models. First, these models require a long time series to yield reliable parameter estimates, but the earnings process may not be stationary over a long period. Second, these firm-specific time series models ignore the rich information in other financial statement line items. To overcome these limitations, subsequent studies turn to cross-sectional approaches (or panel data approaches, as in Monahan, 2018), which use a pooled cross-section of firms to estimate forecasting models. Following recent studies, we adopt several state-of-art cross-sectional models developed in the literature as alternative benchmarks (e.g., Gerakos and Gramacy, 2013; Call, Hewitt, Shevlin, and Yohn, 2016).

The first cross-sectional model that we examine is the first-order AR model, which uses only one-year lagged earnings as the predictor:

$$E_{i,t+1} = \alpha_0 + \alpha_1 E_{i,t} + \varepsilon_{i,t+1} \quad (1)$$

where $E_{i,t}$ is firm i's earnings in year t. Gerakos and Gramacy (2013) show that the AR model performs well relative to the RW model and that it is more accurate than most of the more sophisticated models.

The second cross-sectional model, the HVZ model, is proposed by Hou, van Dijk, and Zhang (2012). The model extends the Fama and French (2000, 2006) models and uses total assets, accruals, and dividends to forecast future earnings:

$$E_{i,t+1} = \alpha_0 + \alpha_1 A_{i,t} + \alpha_2 D_{i,t} + \alpha_3 DD_{i,t} + \alpha_4 E_{i,t} + \alpha_5 NegE_{i,t} + \alpha_6 AC_{i,t} + \varepsilon_{i,t+1} \quad (2)$$

where $A_{i,t}$ is total assets, $D_{i,t}$ is the dividend payment, $DD_{i,t}$ is a dummy variable indicating dividend payers, $E_{i,t}$ is earnings for year t, $NegE_{i,t}$ is a dummy variable indicating negative earnings, and $AC_{i,t}$ is accruals. Detailed variable definitions are provided in Appendix 1.

The third cross-sectional model, the SO model, is developed by So (2013). So modifies the profitability forecasting model proposed by Fama and French (2006) to forecast future earnings per share (EPS), $EPS_{i,t+1}$, using the following regression:

$$EPS_{i,t+1} = \alpha_0 + \alpha_1 EPS_{i,t}^{+} + \alpha_2 NegE_{i,t} + \alpha_3 AC_{i,t}^{-} + \alpha_4 AC_{i,t}^{+} + \alpha_5 AG_{i,t} + \alpha_6 NDD_{i,t} + \alpha_7 DIV_{i,t} + \alpha_8 BTM_{i,t} + \alpha_9 Price_{i,t} + \varepsilon_{i,t+1} \quad (3)$$

where $EPS_{i,t}^{+}$ is EPS for positive earnings, and is zero otherwise; $NegE_{i,t}$ is an indicator variable for negative earnings; $AC_{i,t}^{-}$is accruals per share for negative accruals, and is zero otherwise; $AC_{i,t}^{+}$ is accruals per share for positive accruals, and is zero otherwise; $AG_{i,t}$ is the percentage change in total assets; $NDD_{i,t}$ indicates zero dividend; $DIV_{i,t}$ is dividends per share; $BTM_{i,t}$ is the book-to-market ratio; and $Price_{i,t}$ is the stock price at the end of the third month after the

end of fiscal year t. In addition to financial statement items, the SO model uses stock price and book-to-market ratio, potentially allowing it to capture more forward-looking information.

The final two cross-sectional models, namely, the EP and RI models, are proposed by Li and Mohanram (2014). Specifically, the EP model allows loss firms to have different earnings persistence from profitable firms and predicts earnings with the following regression:

$$E_{i,t+1} = \alpha_0 + \alpha_1 NegE_{i,t} + \alpha_2 E_{i,t} + \alpha_3 NegE_{i,t} * E_{i,t} + \varepsilon_{i,t+1} \quad (4)$$

The RI model, which is based on the residual income model proposed by Feltham and Ohlson (1996), further augments the EP model with the book value of equity ($BVE_{i,t}$) and total accruals ($TACC_{i,t}$) from Richardson, Sloan, Soliman, and Tuna (2005):

$$E_{i,t+1} = \alpha_0 + \alpha_1 NegE_{i,t} + \alpha_2 E_{i,t} + \alpha_3 NegE_{i,t} * E_{i,t} + \alpha_4 BVE_{i,t} + \alpha_5 TACC_{i,t} + \varepsilon_{i,t+1} \quad (5)$$

Although the above models represent state-of-art earnings forecasting models in the literature (Call et al., 2016), prior studies conclude that these models still fail to consistently outperform the RW model (e.g., Monahan, 2018; Easton, Kelly, and Neuhierl, 2018).[1] Given that "*the question of whether historical accounting numbers are useful for forecasting earnings is central to accounting research*" (Monahan, 2018, p. 183), both Monahan (2018) and Easton et al. (2018) call for further research in this area. We believe that machine learning offers several important advantages in earnings forecasting, and a study on the efficacy of machine learning algorithms can not only help us better understand the limitations of the extant models, but also

[1] For example, Monahan (2018) concludes that the forecasts obtained using these models "*are not substantially more accurate than forecasts obtained from the random-walk model*" (p. 182). Easton, Kelly, and Neuhierl (2018) also state that "*all these models offer forecasts of earnings that are less accurate than a random walk*" (p. 2).

shed light on the important research question pertaining to the usefulness of financial statement information for earnings forecasting and equity valuation.

3. Machine-Learning-Based Earnings Forecasting Models

Due to several limitations of the extant models, they do not make the best use of information in financial statements to forecast future earnings. First, the extant models focus on a small number of aggregate financial statement items, such as bottom line earnings and total assets, and fail to fully consider many other financial statement line items that could be highly valuable for earnings prediction (e.g., Fairfield et al., 1996; Chen et al., 2015). Second, even though economic theories and empirical evidence suggest the prevalence of nonlinear relationships between historical accounting information and future earnings, these models mostly adopt linear functional forms (some with simple interactions) and are therefore unlikely to be able to capture these subtle yet important relationships. Machine learning algorithms are designed to handle high dimensional data and are rather flexible with respect to the functional forms of the underlying relationships. Therefore, they can potentially overcome the above limitations and generate better earnings forecasts. In the following, we describe the development of machine learning models for earnings forecasting. Specifically, we first discuss our choice of financial statement line items as the input variables/predictors and then introduce the six machine learning algorithms, followed by a description of the model estimation procedure.

3.1. Financial statement line items as predictors

Although machine learning algorithms are designed to handle high dimensional data, the inclusion of many irrelevant features increases the occurrence probability of overfitting noise and

reduces the effectiveness of machine learning algorithms. Thus, we need to select a set of sufficiently disaggregated financial statement line items without overwhelming the algorithms with excessive irrelevant noise. Following Chen et al. (2015), we construct our predictor variables based on the "Balancing Model" for the balance sheet and income statement provided by the Compustat Fundamental Annual database (Compustat database hereinafter). Specifically, we select 28 major financial statement line items from the Compustat database. Except for operating cash flow, which is obtained from the cash flow statement, the other 27 line items are obtained from the balance sheet and income statement. A detailed list of these items is provided in Appendix 1. In addition to the 28 financial statement line items, we include their respective first-order differences as the input features for machine learning. We add these variables because the literature demonstrates that changes in financial statement items often contain incremental information beyond the levels of these items (e.g., Kothari, 1992; Ohlson and Shroff, 1992; Richardson, Sloan, Soliman, and Tuna, 2005). We use the same set of 56 predictors for the 6 machine learning algorithms, as follows, to forecast future earnings.[2]

3.2. Machine learning algorithms

We study a fairly comprehensive set of machine learning algorithms, including three linear machine learning algorithms and three machine learning algorithms that accommodate nonlinear relationships. The mathematical details of these algorithms can be found in Appendix B of Gu et al. (2020). In the following two subsections, we discuss these algorithms briefly.

[2] We limit the input variables to financial statement data because we are interested in understanding the efficacy of machine learning in extracting information from financial statements. Furthermore, we compare the machine learning models with the extant models, most of which also only use financial statement items (with the exception of the So model). Because stock prices incorporate forward-looking information, the inclusion of stock prices likely improves the predictive power of the models. However, earnings forecasts inferred from stock prices are unlikely to be useful in identifying mispriced stocks through equity valuation.

3.2.1 Linear machine learning algorithms (OLS, LASSO, and Ridge models)

The first model we estimate uses OLS, the least complex algorithm. In this algorithm, the parameters are estimated by minimizing the following objective/loss function:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(E_{i,t+1} - f\left(x_{i,t};\boldsymbol{\theta}\right)\right)^2 \quad (6)$$

where f is a linear function of the predictor variables $x_{i,t}$, which include the aforementioned 56 predictors and the parameter vector $\boldsymbol{\theta}$. $E_{i,t+1}$ denotes the earnings of firm *i* in year t + 1.

When forecasting future earnings using a large number of historical financial statement line items, the abundance of predictors makes OLS prone to overfitting. To alleviate this problem, we adopt two penalized linear models: LASSO and Ridge regressions.

The LASSO regression adds a convex penalty term (i.e., L_1 regularization) to the objective function of OLS:

$$\mathcal{L}(\boldsymbol{\theta};\ \lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(E_{i,t+1} - f\left(x_{i,t};\boldsymbol{\theta}\right)\right)^2 + \lambda\sum_{j=1}^{P}|\theta_j| \quad (7)$$

where θ_j is the j^{th} element of parameter vector $\boldsymbol{\theta}$, λ is the regularization parameter, and all other variables are the same as defined previously. With an appropriate parameter value of λ, the LASSO model sets the coefficients of some predictors to zero and uses only the remaining predictors for forecasting. The optimal value of the regularization parameter is determined using the cross-validation technique, as will be discussed in greater detail later in section 3.2.3.

Ridge regression differs from LASSO regression in that it uses a L_2 regularization term:

$$\mathcal{L}(\boldsymbol{\theta};\ \lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(E_{i,t+1} - f\left(x_{i,t};\boldsymbol{\theta}\right)\right)^2 + \lambda\sum_{j=1}^{P}\theta_j^2 \quad (8)$$

Unlike LASSO, Ridge regression pushes all regression coefficients closer to zero instead of setting some to exactly zero. By shrinking the regression coefficients toward zero, the Ridge model mitigates the risk that the regression coefficients are unduly affected by in-sample noise and collinearity.

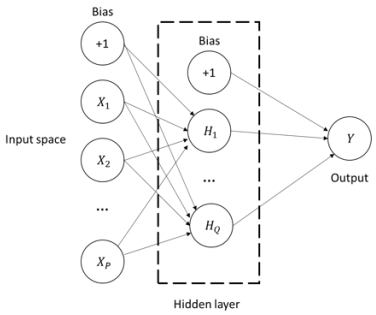
3.2.2 Nonlinear machine learning algorithms (RF, GBR, and ANN models)

We further investigate three more complex models that accommodate nonlinearity. Two of these models are based on decision trees and one is based on artificial neural networks. The detailed theory of decision tree models can be found in Breiman, Friedman, Stone, and Olshen (1984). In decision-tree-based earnings prediction models, the algorithm attempts to separate (the training) samples into relatively homogenous groups (i.e., with similar earnings levels) by making a sequence of binary decisions based on the given lists of predictors (e.g., historical earnings and cash flows). In each step, the algorithm looks for the binary split (i.e., splitting predictor and threshold) that reduces the sum of the squared residuals of all of the resulting subgroups at the fastest rate (i.e., largest information/purity gain), and this process is repeated recursively until all data are processed or a pre-specified stopping criterion is reached. The predicted earnings for an observation equal the average future earnings of all observations in the same leaf node.

Due to the flexible structure, an individual decision tree can easily overfit the data, especially when the number of input features or the depth of the tree is large. To mitigate this problem, we use two decision-tree-based ensemble learning models: the RF model and the GBR model to forecast future earnings. Specifically, the RF model draws a number of different bootstrap samples from the original training set, trains a decision tree for each sample by using a

random subset of predictors, and averages their forecasts to generate predictions. Because bootstrap sampling and predictor dropout weaken the correlation between individual decision trees, the RF model tends to have lower variance and better out-of-sample generalizability than individual decision trees. Boosting is another ensemble technique that enhances the predictive power of (weak) decision tree regression models. The GBR model starts with a simple decision tree regression model and then recursively boosts the model by adding a new decision tree regression model that fits the residuals of the prior model (multiplied by a learning rate), until a certain stopping criterion is fulfilled. Predictions for new observations are then generated using the combined model.

Our final model, that is, the ANN model, is based on Artificial Neural Networks. A simple neural network is illustrated in the following figure:



The basic element in a neural network is the neuron. The neural network in the above figure consists of one hidden layer with Q neurons. Each neuron in the hidden layer receives signals from *P* connected neurons in the previous layer, aggregates them linearly by using a set of weight parameters and a bias term, transforms the resulting value with a nonlinear activation function, and outputs a signal that is used as an input feature for the neurons in the next layer. Neural networks with multiple hidden layers and a large number of neurons are also prone to overfitting. Therefore, we adopt the bootstrap aggregating (i.e., bagging) technique by

constructing 10 bootstrap samples with each sample randomly drawing 60% of the observations from the training set. Thereafter, we train an ANN model for each bootstrapped sample and then average the 10 models to generate predictions.

3.2.3 Cross-validation and hyperparameter tuning

In machine learning tasks, it is imperative to select a model with an appropriate level of complexity because overly simple models tend to underfit the data while overly complex models tend to have the overfitting problem, and both lead to poor out-of-sample predictability. The level of model complexity is largely determined by the value of certain hyperparameters, which must be set optimally before the start of the learning process to estimate other parameter values, such as regression coefficients and neural network weights.[3] We search for the "optimal" hyperparameter values through hyperparameter tuning by using cross-validation. Specifically, for each of the machine learning models (except for OLS), we provide a set of reasonable candidate values for the key hyperparameters. We use five-fold cross-validation to identify the optimal hyperparameter values that generate the most accurate forecasts on the validation samples. Then, we estimate the model by using the optimal hyperparameter values obtained from cross-validation and apply the model to the prediction set to generate out-of-sample forecasts. The key hyperparameters and the corresponding candidate values for each of the five algorithms are provided in Appendix 2 of this paper.[4]

[3] For example, larger values of the regularization hyperparameter in LASSO tend to lead to a relatively less complex model that effectively uses fewer input features for prediction. Similarly, different values of the hyperparameters, such as the maximum depth of decision trees in the RF and GBR models and the number of hidden layers in the ANN model, may also lead to models with drastically different structures and complexities.

[4] The use of cross-validation and hyperparameter tuning marks an important distinction between our work and that of Gerakos and Gramacy (2013). As discussed earlier, randomly selected hyperparameter values likely result in either overfitted or underfitted models. Gerakos and Gramacy (2013) do not explain clearly how they determine the structure of their models, but the software packages they use (lars, MASS, and randomForest packages in R) do not seem to implement cross-validation. Our paper also differs from that paper in several other dimensions. First, we use

4. Data, Sample Selection, and Model Estimation Procedure

Our initial sample comprises 267,777 firm-year observations obtained from the intersection of the Compustat fundamentals annual file and the Center for Research in Security Prices (CRSP) data up to fiscal year 2019. We further impose the following data requirements: 1) the following financial statement items must be non-missing: total assets, sales revenue, income before extraordinary items, and common shares outstanding; 2) the stocks must be ordinary common shares listed on the NYSE, AMEX, or NASDAQ; 3) the firms cannot be in the financial (SIC 6000-6999) and regulated utilities (SIC 4900-4999) industries; and 4) the stock prices at the end of the third month after the end of the fiscal year must be greater than US$1. Among the remaining firm-year observations, we first set the missing values of some line items[5] to zero before computing the first-order differences of the 28 items in Appendix 1. We then delete all firm-year observations with missing values for the 56 input features. This leaves us with a final sample of 156,256 observations from 1965 to 2019. Because we need data from the past 10 years to estimate the models, our testing sample (i.e., prediction set) starts from 1975 and consists of 142,592 firm-year observations. Table 1 presents the number of firms in the final testing sample by year, where the number of annual observations ranges from 2,299 in 2019 to 4,976 in 1997.

We generate out-of-sample forecasts of one-year-forward earnings E_{t+1} for the above testing sample using the 6 machine learning algorithms and the 56 predictors, as discussed

a more comprehensive list of financial statement items and include their first-order differences in model training. Our analyses show that the change variables are often among the top 10 most important input features. Furthermore, we investigate a more comprehensive set of advanced machine learning algorithms, such as GBR and ANNs. These models are not examined in Gerakos and Gramacy (2013).

[5] The list of items includes special items; accounts payable; income taxes payable; interest and related expenses; investments and advances-other; selling, general, and administrative expenses; intangible assets; short-term investments; research and development (R&D) expense; advertising expenses; current liabilities; current assets; and dividends per share. Cash flow from operating activities, if missing, is computed using the balance sheet approach (Sloan, 1996).

above. We scale both the predictors and the target variable by the common shares outstanding at the fiscal year end to ensure that the estimation procedure (or the loss function) is not dominated by a small number of extremely large firms. Following the prior literature (e.g., Hou et al., 2012, Li and Mohanram, 2014), for each year t between 1975 and 2019, we use all observations from the previous 10 years (i.e., year t - 10, t - 9, …, t - 1) as the training sample to estimate the models, then apply the models to the predictors of year t to generate earnings forecasts for year t+1.[6] For consistency, all of the extant models are also estimated using the data of the same previous 10 years, and the resulting linear models are applied to their respective predictors in year t to generate earnings forecast for year t + 1.

5. Empirical Analysis

In this section, we compare the quality of the earnings forecasts generated using the six machine learning models (namely OLS, LASSO, Ridge, RF, GBR, and ANN) against the forecasts obtained using the benchmark RW model and the other extant models (AR, HVZ, SO, EP, and RI). In particular, we investigate the accuracy and information content of these forecasts and analyze the economic significance of the findings.

5.1. Comparison of forecast accuracy

To evaluate the forecast accuracy of the different models, we first compare the mean and median absolute forecast errors. Following the literature, we define forecast error as the

[6] Following the literature, we assume that financial statement data are available at the end of the third month after the end of the previous fiscal year. For machine learning algorithms that require hyperparameter tuning, we first apply the aforementioned five-fold cross-validation procedure on the same training sample to determine the optimal hyperparameter values before estimating the parameter values for the "optimal" models.

difference between the predicted earnings and the actual earnings deflated by the market value of equity at the end of three months after the fiscal year end. A larger absolute forecast error indicates less accurate earnings forecasts. Table 2 reports the time series average of the out-of-sample annual mean and median absolute forecast errors of all models. The benchmark RW model has an average mean absolute forecast error of 0.0764 and an average median absolute forecast error of 0.0309. Consistent with the literature, we find that the extant models struggle to generate significantly more accurate forecasts than the naïve RW model. Even two of the most accurate traditional models, namely EP and RI, fail to consistently beat the RW model. For example, although the mean absolute forecast error of the RI model (0.0741) is approximately 3.07% lower than that of the RW model, its median absolute forecast error is higher. Among the other extant models, HVZ is slightly less accurate than EP and RI,[7] followed by the AR model and the SO model.

The machine learning models, especially those that accommodate nonlinear relationships, tend to generate more accurate earnings forecasts. The average mean absolute forecast errors of the three linear machine learning models are 0.720, 0.716, and 0.718, which are 5.83%, 6.31%, and 6.11% lower than those of the RW model, respectively. The average median absolute forecast errors of the three linear machine learning models are also lower than that of the RW model, but the differences are not statistically significant. The machine learning models that accommodate nonlinear relationships further improve forecast accuracy. The average mean absolute forecast errors of the RF and GBR models are 0.0698 and 0.0697, which are approximately 8.64% and 8.86% lower than that of the RW model, respectively, and the

[7] If we estimate the HVZ model at the dollar level with unscaled data, as in Hou et al. (2012) and Li and Mohanram (2014), the results are similar to those of Li and Mohanram (2014), which demonstrates that the HVZ model is significantly less accurate than the EP and RI models, which are estimated at the per-share level.

differences are statistically significant. The average median absolute forecast errors of the RF and GBR models are also significantly lower than that of the RW model. While the average mean absolute forecast error of the ANN model is lower than that of the RW and the other extant models, its median absolute forecast error is slightly higher.

Figure 1 plots the rank of the average deciles of the absolute forecast errors of different individual models over the 45-year sampling period. A lower rank indicates a more accurate prediction model. As Figure 1 shows, the GBR and RF models are the two most accurate models across all of the nine deciles. Although the ANN model performs well at the extremes, its two middle percentiles are less accurate than those of the RW model and some of the extant models. The linear machine learning models also show reasonably stable performance across the spectrum, indicating that regularization helps in most parts of the distribution. The plots of the extant models are intertwined with that of the RW model, which is consistent with the prior literature (e.g., Monahan, 2018; Easton et al., 2018) that these models do not consistently outperform the RW model.

For the four composite models, the results at the bottom of Table 2 show that composite forecasts obtained by combining predictions from different models help increase forecast accuracy, especially when the individual models are not strongly correlated. In terms of the time series average, the average mean absolute forecast errors of the composite forecasts generated by combining the extant models (COMP_EXT), the linear machine models (COMP_LR), and the nonlinear machine models (COMP_NL) are 0.0737, 0.0717, and 0.0689, representing improvements of 3.58%, 6.16%, and 9.87% relative to the RW model, respectively. The average median absolute forecast error of COMP_NL is also approximately 5.55% lower than that of the RW model, and the difference is statistically significant. The results suggest that the ability to

handle high dimensional financial statement data and accommodate nonlinear subtle relationships allows the machine learning models to produce significantly more accurate out-of-sample earnings forecasts. The results also show that both the mean and median absolute forecast errors of COMP_ML, which combines the six machine learning algorithms, are greater than those of COMP_NL, suggesting that much of the information in the linear models may have already been incorporated in the nonlinear models.

5.2. Cross-sectional analysis

The above results suggest that the machine learning models generate significantly more accurate earnings forecasts than the RW model. We posit that the benefit of considering more financial statement line items and more complex forms of relationships would be more important for firms with more difficult-to-forecast earnings. We partition our sample along the following dimensions, ROA volatility, magnitude of accruals, R&D expense, and an indicator variable of loss firms, and report the percentage improvement in the forecast accuracy of COMP_LR and COMP_NL relative to the benchmark RW model for each subgroup. The results are presented in Table 3.

Panel A presents the results for the subsamples partitioned on ROA volatility. ROA volatility is calculated as the standard deviation of earnings scaled by total assets (ROA) over the past five years with non-missing values for at least three years. The benefits of both considering more financial statement line items and accommodating nonlinear relationships increase with earnings volatility. For the highest ROA volatility quintile, COMP_LR improves forecast accuracy by 10.26%, whereas COMP_NL improves forecast accuracy by 15.21%. For the lowest ROA volatility quintile, COMP_LR reduces forecast accuracy by approximately 2% compared

with the RW forecasts, but COMP_NL improves forecast accuracy by 4.44% compared with the RW forecasts. The results suggest that when earnings are stable, the addition of other financial statement line items to a linear model does not improve its predictive power. However, considering nonlinear relationships still leads to substantial improvements in forecast accuracy.

Panel B and Panel C report the results for the subsamples partitioned on the magnitude of total and working capital accruals, respectively. Both panels show the greater benefits of using machine learning models for firms with higher accruals. COMP_NL is 17.71% and 13.68% more accurate than the RW model for firms with the highest magnitude of total accruals and firms with the highest magnitude of working capital accruals, respectively. Among these firms, the linear machine learning models improve forecast accuracy by 10%–12%. For firms with the lowest magnitude of accruals, the linear machine learning models do not seem to improve forecast accuracy significantly, although they consider a large number of financial statement line items. However, there is still a significant benefit in accommodating nonlinear relationships in these firms. Panel D shows that although the linear machine learning models improve the accuracy of earnings forecasts in most quartiles of R&D expense scaled by total assets, there are no consistent cross-sectional variations. However, it is more important to accommodate nonlinear relationships for firms with higher R&D expense. Finally, Panel E shows that the benefits of considering more financial statement line items and accommodating nonlinear relationships are greater among loss firms.

5.3. A peek into the "black box"

We conduct several additional analyses to better understand the underlying reasons for the superiority of the nonlinear machine learning models. As discussed earlier, the ability to

handle high dimensional data and accommodate more complex relationships are two important advantages of nonlinear machine learning models. We first plot the feature importance charts of the two tree-based machine learning models (i.e., RF and GBR) to check whether these models use economically sensible features to generate predictions. Figure 2 shows the top 10 features that on average make the largest contributions to the model predictions over the entire sampling period. The results indicate that past earnings and operating cash flows are important predictors of future earnings in both algorithms, ranked 1st and 3rd on the top 10 lists, respectively. Interestingly, total income tax and its first-order difference are the 2nd and 4th most important features, respectively, which is consistent with the recent literature on the important role of tax income or expenses in capturing the quality of earnings and predicting future fundamentals and stock returns (e.g., Lev and Nissim, 2004; Hanlon, 2005; Hanlon, Laplante, and Shevlin, 2005; Thomas and Zhang, 2011, 2014). Other influential predictors include common equity and its change and changes in total assets and receivables.

Panels A–E of Figure 3 present the accumulated local effects (ALE) plots, which provide a visualization of the effects of the predictors (Apley and Zhu, 2020), of the top five most important features of the RF prediction models for 1975, 1985, 1995, 2005, and 2015, respectively.[8] A few observations stand out from the plots. First, current earnings are the single most important feature across all years, and the RF models uncover a lower persistence for loss than for profits, a feature that is explicitly assumed by most extant models. Second, cash flow

[8] Partial dependence plot (PDP) is another commonly used method to depict the functional relationship between a particular predictor and the predicted value of the target variables. PDP looks at a particular predictor across a specified range. At each value of the predictor, the model is evaluated for *all* observations of the other model inputs, and the output is then averaged. Thus, the plotted relationship is only valid if the variable of interest does not correlate strongly with other model inputs. In contrast, ALE plots estimate the local effects of a predictor of interests over a number of small intervals using only observations falling into that intervals and then accumulate them. Thus, the ALE plot is an unbiased alternative to PDP.

from operating activities is in the top five lists in all five years. Although future earnings increase with current operating cash flow, the predicted relationship exhibits an obvious nonlinear pattern. Third, although the book value of equity (CEQ) is in the top five lists in 1975 and 1985, it disappears from the lists in subsequent years. Fourth, both the level and changes in tax expenses are among the top five most important features in almost all years. They both share an increasing but nonlinear relationship with future earnings. Panels F–J of Figure 3 report similar results for the GBR models, except that the figures of some of the change variables are more irregular.

In addition to the nonlinear relationships between individual predictors and the target variable, the nonlinear machine learning models accommodate the interaction effects between predictors. Our untabulated analyses show that the interaction effects between the following pairs are the top five contributors to the explanatory power of the GBR model[9]: change in sales revenue (SALE) and change in cost of goods sold (COGS), change in debts in current liabilities (DLC) and change in total current liabilities (LCT), sales revenue (SALE) and accounts payable (AP), cost of goods sold and inventories (INVT), depreciation and amortization expense (DP) and net property, plant, and equipment (PPENT). The first two pairs identify the interaction effects between two closely related accounts on the income statement and balance sheet items, separately. The other three pairs pick up the interactions between income statement items and their closely related balance sheet gross accrual items. The finding that the machine learning models pick up the interaction between income statement items and the corresponding gross

[9] We use Friedman's H-statistic to measure the pairwise feature interaction strength (Friedman and Popescu, 2008). To implement the calculation in Python, we use sklearn-gbmi (for details, refer to https://pypi.org/project/sklearn-gbmi/). For the GBR model we build each year, we first calculate the H-statistic for each pair and average the H-statistic of the 45 models to determine the top five interaction terms in general.

rather than net accrual items echoes remarkably well with the recent call for research by Dichev (2020) on the role of gross accruals in determining the quality of earnings.

5.4. Information content analysis

Forecast accuracy is not a sufficient statistic for the decision usefulness of earnings forecasts. For example, although the RW forecast is relatively more accurate than other forecasts, it provides no information with respect to future earnings changes. In this section, we evaluate the information content of various models by investigating their (out-of-sample) predictive power with respect to future earnings change, ECH. ECH is computed as the difference between earnings in year t + 1 and those in year t, scaled by market capitalization at the end of the third month after the end of fiscal year t. We calculate forecasted earnings change, or FECH, as the predicted earnings for year t + 1 minus the actual earnings for year t, scaled by market capitalization at the third month end after the end of fiscal year t.

We first compare the mean Pearson and Spearman correlation coefficients between FECH calculated from various models and ECH in the 45 years. Table 4 Panel A shows that the Pearson correlation coefficients between the forecasted earnings changes derived from the extant models and ECH range from 0.199 to 0.321, whereas the correlation coefficients for the three linear and three nonlinear machine learning models increase to approximately 0.37 and 0.40, respectively. For the composite forecasts, the Pearson correlation coefficients increase from 0.333 for COMP_EXT, to 0.372 for COMP_LR, and 0.413 for COMP_NL. Consistent with earlier results, further combining linear models with COMP_NL does not increase the information content because the correlation coefficient for COMP_ML is smaller than that for COMP_NL. We find a similar pattern for the Spearman correlation coefficients, with

COMP_EXT, COMP_LR, COMP_NL, and COMP_ML having correlation coefficients of 0.188, 0.247, 0.300, and 0.286, respectively.

We also run the univariate Fama–MacBeth regression of ECH on FECH calculated using different models. To facilitate the comparison of the coefficients, we follow the literature to standardize the forecasted earnings changes so that they have zero mean and unit variance each year. The three columns on the right in Panel A of Table 4 present the regression results. The coefficients on FECH computed using the extant models range from 0.0304 to 0.048, explaining between 8.07% and 12.22% of the cross-sectional variation in realized earnings changes. The coefficients increase to about 0.055 and the explanatory power improves between 14.87% and 15.12% for the three linear machine learning models. For the three nonlinear machine learning models, the coefficients further increase to about 0.058 and the adjusted R-squares increase to a level between 16.95% and 17.36%. Among the composite forecasts, FECH based on COMP_NL has the largest regression coefficient of 0.0605 and explanatory power of 18.57%, which are not only higher than those of COMP_EXT (0.0497 and 12.73%), but also higher than those of COMP_LR (0.0550 and 15.09%) and COMP_ML (0.0601 and 18.09%).

Next, we run a multivariate Fama–MacBeth regression of ECH on FECH based on the six machine learning models and the composite models by controlling for FECH predicted using all extant models. All independent variables are standardized to have zero mean and unit variance each year. Panel B of Table 4 shows the results. All FECH coefficients determined using the machine learning models are significantly positive, with Newey–West t-statistic greater than 10. The nonlinear machine models in general yield more incremental information content than the linear machine learning models. When the machine learning models are used, the FECH coefficients based on all extant models become not significant, except for that of the SO model,

which uses forward-looking non-financial statement predictors, such as stock price and book-to-market ratio. The results suggest that machine learning models efficiently aggregate information from financial statements, the information content of extant models.

5.5. Earnings response coefficient

Following the prior literature (e.g., Hou et al., 2012; Li and Mohanram, 2014; Easton, Kelly, and Neuhierl, 2018), we examine the earnings response coefficients (ERCs) by using the above forecasts. As prior studies argue, a higher ERC indicates that the market reacts more strongly to unexpected earnings generated using that model, which implies that the earnings forecasts obtained using that model resemble the market expectations (of future earnings) more closely. Following Hou et al. (2012), we estimate the ERC of the various models in two ways. First, we estimate the ERC by running annual cross-sectional regressions of the sum of future quarterly earnings announcement window market-adjusted stock returns on standardized unexpected earnings. This type of ERC is defined as "announcement ERC." Unexpected earnings are calculated as the difference between future actual earnings and the forecasts obtained using the above models, deflated by market capitalization three months after the end of the previous fiscal year. Then, we standardize these unexpected earnings to have zero mean and unit variance each year to facilitate comparisons across different models.

The left panel of Table 5 presents the announcement ERCs of the various models. Consistent with Li and Mohanram (2014), we find that the announcement ERCs based on the EP and RI forecasts are higher than that based on the RW and HVZ models. In contrast, the announcement ERCs based on the forecasts generated using the AR and SO models are lower than that based on the RW model. The announcement ERCs obtained using the machine-

learning-based forecasts are also higher than that based on the RW model, but are not reliably higher than those of the extant models. The announcement ERCs of the three linear machine learning models (approximately 0.041) are similar to those of the EP and RI models (0.0410 and 0.0411, respectively). Although the announcement ERC of the RF model is slightly higher at 0.0412, the announcement ERC of the GBR model is lower than those of the extant models. The ERCs based on the composite machine learning models (COMP_LR, COMP_NL, and COMP_ML) are not significantly higher than that of COMP_EXT.

Second, we estimate the ERC by regressing the buy-and-hold returns over the next one year starting from the fourth month after the end of fiscal year t on standardized unexpected earnings over the same period, which is denoted as "annual ERC." The right panel of Table 5 presents the results. None of the individual models have significantly higher annual ERCs than that of the RW model. Only the RF and ANN models generate annual ERCs higher than that of the RW model, but the difference is not statistically significant. Moreover, none of the composite forecasts have a significantly higher annual ERC than that of the RW model.

The overall results suggest that although the machine-learning-based forecasts are more accurate and informative about future earnings changes, the market does not seem to give them more weight when forming expectations about future earnings. This suggests that the market may underreact to the new information content of machine-learning-based forecasts, and we provide more evidence of this phenomenon in the following section.

5.6. Economic significance analysis

In this section, we examine whether the new information uncovered by the machine learning models is economically significant for investors. To capture this new information, we

orthogonalize the machine-learning-based forecasts against the forecasts generated using the RW model and the extant models. Specifically, we run a cross-sectional regression of the machine-learning-based forecasts on the RW model and the five extant models each year and use the resulting residual forecasts to measure the new information uncovered by the machine learning models. Then, we estimate the following regression models and examine whether the residual forecasts are associated with future stock returns:

$$EXRET12M_{i,t+1} = \beta_0 + \beta_1 ML\_RESD_{i,t} + \beta_2 SIZE_{i,t} + \beta_3 BM_{i,t} + \beta_4 MOM_{i,t} + \beta_5 ROE_{i,t} + \beta_6 INV_{i,t} + \beta_7 ACC_{i,t} + IndustryFE + \varepsilon_{i,t+1} \quad (9)$$

where $EXRET12M_{i,t+1}$ is the one-year ahead excess return over 12 months starting from the fourth month after the end of fiscal year t for firm i. $ML\_RESD_{i,t}$ is the residual from the aforementioned regression that orthogonalizes the machine-learning-based forecasts against the RW model and the five extant models; $SIZE_{i,t}$ is the logarithm of market capitalization at the end of the third month after the end of fiscal year t; $BM_{i,t}$ is the book-to-market ratio; $MOM_{i,t}$ is the momentum calculated as the cumulative return of firm i's stock during the 11-month period starting 12 months ago; $ROE_{i,t}$ is profitability, defined as earnings divided by the book value of common equity; $INV_{i,t}$ is the growth rate of total assets; and $ACC_{i,t}$ is accruals scaled by total assets. We also include the three-digit SIC as industry fixed effects in the regression.

Table 6 presents the Fama–MacBeth regression results of Model (9). All of the residual machine learning forecasts have significant positive associations with future stock returns, even after controlling for various return-predicting factors. The coefficients of the other return-predicting factors mostly bear signs that are consistent with those in the literature. For example,

future stock returns are negatively associated with firm size, total asset growth, and accruals, and are positively associated with book-to-market ratio and profitability.

Next, we conduct portfolio analysis on the return predictive power of the new information component. Specifically, at the beginning of each month, we estimate the new information component as the residual from the regression of the machine-learning-based forecasts on the forecasts generated using the RW model and the five extant models. We then sort all stocks into quintiles based on the resulting residual forecasts for each three-digit SIC industry. We construct a hedge portfolio that takes long positions in quintiles with the most favorable new information and short positions in quintiles with the least favorable new information. Table 7 reports the average return, CAPM alpha, Fama–French three-factor alpha, Carhart four-factor alpha, and Fama–French five-factor alpha for the equal-weighted and value-weighed hedge portfolios. The results in Panel A of Table 7 indicate that all residual forecasts (linear or nonlinear, individual or composite) based on the machine learning models generate significantly positive alphas for the equal-weighted portfolios. In particular, the hedge portfolios created using the three composite forecasts, COMP_LR, COMP_NL, and COMP_ML, generate monthly returns of approximately 64, 72, and 77 bps, respectively. The monthly alphas of the three portfolios computed using the Fama–French five-factor model remain statistically significant and are above 50 bps per month.

Panel B of Table 7 reports the mean monthly returns and alpha of the value-weighted hedge portfolios. The overall results are weaker compared to the equal-weighted portfolios, especially when using the linear machine learning models. All Fama–French five-factor alphas of the residual forecasts obtained using the linear machine learning models are not statistically significant. However, the residual forecasts obtained using all nonlinear machine learning

models generate significant risk-adjusted returns. For example, the Fama–French five-factor alpha of the residual forecast based on COMP_NL still amounts to 41 bps, which is significant both statistically and economically.

6. Additional Analysis and Robustness Check

6.1. Alternative deflator

As a robustness check, we calculate the absolute errors of earnings forecasts using two alternative deflators, namely total assets (at) and common shares outstanding (csho). The results are presented in Table 8. The left panel reports the results obtained using total assets as the deflator. The time series average of the mean absolute errors of the RW model is 0.0593. None of the extant models have lower mean absolute forecast errors than that of the RW model. All of the linear machine learning models, including the composite forecasts, have lower mean absolute forecast errors, but not significantly so. In contrast, all of the nonlinear machine learning models yield significantly more accurate earnings forecasts than the RW model, with a decrease in mean absolute forecast errors ranging from 3.82% for the ANN model to 8.59% for the COMP_NL model. The superior prediction performance of the nonlinear machine learning models remains robust when we compare the per share forecast error, which is reported in the right panel of Table 8. The results again show that the extant models do not generate significantly more accurate forecasts than the RW model. The mean absolute forecast errors of the six machine learning models are significantly lower than that of the RW model. In both cases, the GBR model and COMP_NL are still the most accurate individual and composite models.

6.2. Longer horizon forecasts

In the main analyses, we focus on the forecast horizon of one year. All our results hold and, sometimes, become stronger when we extend the forecast horizon to two or three years ahead. Table 9 presents the results for two- and three-year ahead earnings forecasts, respectively. The overall results are largely similar, showing that the RF and GBR models are the most accurate individual forecast models, while COMP_NL is generally the most accurate composite forecast model. Judging from the mean absolute forecast errors, these forecasts are approximately 11% more accurate than that of RW model for two-year ahead forecasts. The median absolute forecast errors of COMP_NL are also approximately 6% lower than those of the RW model over the two forecast horizons.

6.3. Prediction of analyst forecast errors

Our overall results show that the machine learning models help to extract new information about future earnings. A natural question is whether this new information is well understood by sell-side financial analysts. We answer this question by examining whether the new information component helps predict errors in consensus analyst earnings forecasts made after the availability of financial statement information. Specifically, we examine the errors in analyst consensus forecasts made in the fourth month after the end of fiscal year t. Following So (2013), we estimate the following models:

$$FERR_{i,t+1} = \beta_0 + \beta_1 ML\_RESD_{i,t} + \beta_2 SIZE_{i,t} + \beta_3 BM_{i,t} + \beta_4 MOM_{i,t} + \beta_5 ACC_{i,t} + \beta_6 LTG_{i,t} + \varepsilon_{i,t+1} \quad (10)$$

where $FERR_{i,t+1}$ is the analyst forecast error defined as the realized difference between EPS as reported in IBES and the consensus EPS forecast made in the fourth month after the end of fiscal year t, scaled by the stock price of firm i on the day the consensus forecast is formed (i.e.,

statpers); $ML\_RESD_{i,t}$, $SIZE_{i,t}$, $BM_{i,t}$, $MOM_{i,t}$, $ACC_{i,t}$ have the same definitions as those in Model (9); and $LTG_{i,t}$ is the consensus long-term growth forecast in IBES.

If analysts perfectly incorporate the new information extracted by the machine learning models into their forecasts, their forecast errors will be uncorrelated with the proxy for the new information component. However, the results in Table 10 indicate that the proxies of all machine learning models are significantly correlated with analyst forecast errors, suggesting that analysts do not fully understand the new information uncovered by the machine learning algorithms.

6.4. Improving extant models using machine learning

The overall results show that even without the explicit guidance of economic theories, the nonlinear machine learning models extract information from a set of economically sensible predictors and capture the interaction effects between some economically linked pairs of predictors. Nevertheless, the models may be too complex for some researchers to embrace wholeheartedly. Thus, we test whether the insights obtained from the nonlinear machine learning models can be used to enhance the performance of the extant models.

First, we test whether augmenting the extant models with the level and change in total income tax expenses helps improve their performance. Tax expenses are closely related to taxable income, which according to prior studies is an important determinant of the quality of earnings and earnings persistence (e.g., Lev and Nissim, 2004; Hanlon, 2005). One may justify the inclusion of these variables in the earnings prediction model, just like the inclusion of other variables, such as accruals. The results presented in Panel A of Table 11 show that the augmented models always have significantly lower mean absolute forecast errors than the

corresponding extant models, with a decrease in average mean absolute forecast errors range from 2.53% to 4.90%.

Second, we examine whether the linear models that use the sets of predictors identified by the nonlinear machine learning models as the most important features yield better performing earnings prediction models. Random forests are frequently used by data scientists for variable selection (e.g., Hapfelmeier and Ulm, 2013). Our earlier results show that RF identifies a set of economically sensible features for predicting future earnings. For example, if we use the top five most influential features identified by the RF algorithm, we can formulate the following model:

$$E_{i,t+1} = \alpha_0 + \alpha_1 CEQ_{i,t} + \alpha_2 E_{i,t} + \alpha_3 CFO_{i,t} + \alpha_4 TXT_{i,t} + \alpha_5 \Delta TXT_{i,t} + \varepsilon_{i,t+1} \quad (11)$$

where all variables are the same as defined earlier. This model is very intuitive and fits very well with economic theory. It uses information about the book value of equity, current earnings and operating cash flows, current level, and change in tax expenses to forecast future earnings. We estimate the above model using the OLS, LASSO, and Ridge regressions. The results reported in Panel B of Table 11 suggest that the three models are not only significantly more accurate than the RW model, but also outperform all extant models. For example, the average mean absolute errors of the forecasts generated using Model (11) based on OLS is 0.0713, which is lower than that of the more accurate extant RI model, which has an average mean absolute forecast error of 0.0741, as shown in Table 2. Furthermore, the results show that the earnings predictions generated by the three models are more accurate than the predictions generated by the corresponding models using all 56 inputs. The results suggest that although supplying the linear models with the full set of inputs allow the models to accommodate a considerably richer

information set, doing so may also exacerbate the overfitting problem. Regularization using LASSO and Ridge models does not eliminate the overfitting risk.

6.5. Other robustness tests

We also conduct a host of additional analyses to test the robustness of our results. The tests are briefly summarized below. For brevity, the results are not tabulated in the paper, but they are available upon request:

1) We conduct model training by deflating both the target variable and the predictors with total assets or market capitalization and the results are similar, showing that the machine learning models generate more accurate and informative earnings forecasts and that the new information in these forecasts predicts both future stock returns and analyst forecast errors.

2) We rerun all analyses by further excluding all firms with market capitalization in the bottom 25% of the annual distribution. All results are slightly weaker, but still significant both statistically and economically.

3) Instead of implementing the ANN with bagging, we attempt to generate earnings forecasts using plain vanilla ANN and the results are similar, *albeit* slightly weaker.

4) Instead of the key financial statement line items from the Compustat Balancing Model, we use the top 100 fundamental signals, which are considered as the economic drivers in Yan and Zheng (2017), as the predictors to forecast future earnings, but the resulting earnings forecasts are not as accurate or as informative.

7. Conclusions

As a crucial input for equity valuation, earnings forecasts are of central importance to both academics and practitioners. We posit that the ability to handle high dimensional data and accommodate more flexible relationships benefits machine learning algorithms to forecast future earnings using financial statement information. Consistent with this notion, our analyses show that machine learning algorithms, especially those that accommodate nonlinear relationships, generate more accurate forecasts than both the RW model and the state-of-the-art earnings prediction models developed in the accounting and finance literature. Furthermore, machine-learning-based earnings forecasts contain more information about future earnings changes than the extant models.

Despite their superiority in terms of accuracy and information content, investors do not seem to give more weight to machine-learning-based forecasts when forming expectations about future earnings. Earnings response coefficients for earnings surprises based on machine learning forecasts are not reliably larger than those based on the extant earnings prediction models. Furthermore, we find that the new information uncovered by the machine learning algorithms is economically significant for investors. In particular, the new information component of earnings forecasts generated using machine learning models, especially those that accommodate nonlinear relationships, are significantly associated with future stock returns. Stocks with the most favorable new information outperform those with the most unfavorable new information by approximately 41 to 77 bps per month. The overall results thus suggest that the market appears to underreact to the new information uncovered by machine learning models. Our paper contributes to the literature by providing robust evidence of the usefulness of machine learning algorithms in forecasting corporate earnings, which is one of the most critical tasks in fundamental analysis

and equity valuation. The results also shed light on the limitations of the extant earnings prediction models and provide unambiguous evidence of the usefulness of financial statement information in earnings forecasting and equity valuation.

References:

Albrecht, W. S., L. L. Lookabill, and J. C. McKeown. 1977. The time-series properties of annual earnings. *Journal of Accounting Research* 15 (2):226-244.

Apley, D. W., and J. Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B* 82 (4):1059-1086.

Arrow, K. J. 1964. Optimal capital policy, the cost of capital, and myopic decision rules. *Annals of the Institute of Statistical Mathematics* 16 (1):21-30.

Baginski, S. P., K. S. Lorek, G. L. Willinger, and B. C. Branson. 1999. The relationship between economic characteristics and alternative annual earnings persistence measures. *The Accounting Review* 74 (1):105-120.

Ball, R., and R. Watts. 1972. Some time series properties of accounting income. *The Journal of Finance* 27(3):663-681.

Bao, Y., B. Ke, B. Li, Y. J. Yu, and J. Zhang. 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research* 58 (1):199-235.

Bradshaw, M. T., M. S. Drake, J. N. Myers, and L. A. Myers. 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies* 17 (4):944-968.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. *CRC Press.*

Brown, L. D. 1993. Earnings forecasting research: Its implications for capital markets research. *International Journal of Forecasting* 9 (3):295-320.

Call, A. C., M. Hewitt, T. Shevlin, and T. L. Yohn. 2016. Firm-specific estimates of differential persistence and their incremental usefulness for forecasting and valuation. *The Accounting Review* 91 (3):811-833.

Chen, P. F., and G. Zhang. 2003. Profitability, earnings and book value in equity valuation: A geometric view and empirical evidence. *Hong Kong University of Science and Technology Working Paper*. Available at SSRN: https://ssrn.com/abstract=442260 or http://dx.doi.org/10.2139/ssrn.442260

Chen, S., B. Miao, and T. Shevlin. 2015. A new measure of disclosure quality: The level of disaggregation of accounting data in annual reports. *Journal of Accounting Research* 53 (5):1017-1054.

Dechow, P. M., W. Ge, C. R. Larson, and R. G. Sloan. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28 (1):17-82.

Dichev, I. D. 2020. Fifty years of capital markets research in accounting: Achievements so far and opportunities ahead. *China Journal of Accounting Research*. In press.

Ding, K., B. Lev, X. Peng, T. Sun, and M. A. Vasarhelyi. 2020. Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies*. https://doi.org/10.1007/s11142-020-09546-9.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64 (2-3):221-245.

Easton, P. D., P. Kelly, and A. Neuhierl. 2018. Beating a random walk. Available at SSRN: https://ssrn.com/abstract=3040354 or http://dx.doi.org/10.2139/ssrn.3040354

Fairfield, P. M., R. J. Sweeney, and T. L. Yohn. 1996. Accounting classification and the predictive content of earnings. *The Accounting Review* 71 (3):337-355.

Fama, E. F., and K. R. French. 2000. Forecasting profitability and earnings. *The Journal of Business* 73 (2):161-175.

Fama, E. F., and K. R. French. 2006. Profitability, investment and average returns. *Journal of Financial Economics* 82 (3):491-518.

Feltham, G. A., and J. A. Ohlson. 1996. Uncertainty resolution and the theory of depreciation measurement. *Journal of Accounting Research* 34 (2):209-234.

Freeman, R. N., and S. Y. Tse. 1992. A nonlinear model of security price responses to unexpected earnings. *Journal of Accounting Research* 30 (2):185-209.

Friedman, J. H., and B. E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2 (3):916-954.

Gerakos, J. J., and R. Gramacy. 2013. Regression-based earnings forecasts. *Chicago Booth Research Paper* (12-26). Available at SSRN: https://ssrn.com/abstract=2112137 or http://dx.doi.org/10.2139/ssrn.2112137

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33 (5):2223-2273.

Hanlon, M. 2005. The persistence and pricing of earnings, accruals, and cash flows when firms have large book-tax differences. *The Accounting Review* 80 (1):137-166.

Hanlon, M., S. Kelley Laplante, and T. Shevlin. 2005. Evidence for the possible information loss of conforming book income and taxable income. *The Journal of Law and Economics* 48 (2):407-442.

Hapfelmeier, A., and K. Ulm. 2013. A new variable selection approach using random forests. *Computational Statistics & Data Analysis* 60:50-69.

Hayek, F. A. 1945. The use of knowledge in society. *The American Economic Review* 35 (4):519-530.

Hou, K., M. A. Van Dijk, and Y. Zhang. 2012. The implied cost of capital: A new approach. *Journal of Accounting and Economics* 53 (3):504-526.

Kothari, S. 1992. Price-earnings regressions in the presence of prices leading earnings: Earnings level versus change specifications and alternative deflators. *Journal of Accounting and Economics* 15 (2-3):173-202.

Kothari, S. 2001. Capital markets research in accounting. *Journal of Accounting and Economics* 31 (1-3):105-231.

Lev, B. 1983. Some economic determinants of time-series properties of earnings. *Journal of Accounting and Economics* 5:31-48.

Lev, B., and D. Nissim. 2004. Taxable income, future earnings, and equity values. *The Accounting Review* 79 (4):1039-1074.

Li, K. K., and P. Mohanram. 2014. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies* 19 (3):1152-1185.

Monahan, S. J. 2018. Financial statement analysis and earnings forecasting. *Foundations and Trends® in Accounting* 12 (2):105-215.

Ohlson, J. A. 1995. Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research* 11 (2):661-687.

Ohlson, J. A., and B. E. Juettner-Nauroth. 2005. Expected EPS and EPS growth as determinants of value. *Review of Accounting Studies* 10 (2-3):349-365.

Ohlson, J. A., and P. K. Shroff. 1992. Changes versus levels in earnings as explanatory variables for returns: Some theoretical considerations. *Journal of Accounting Research* 30 (2):210-226.

Rapach, D. E., J. K. Strauss, J. Tu, and G. Zhou. 2019. Industry return predictability: A machine learning approach. *The Journal of Financial Data Science* 1 (3):9-28.

Richardson, S. A., R. G. Sloan, M. T. Soliman, and I. Tuna. 2005. Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics* 39 (3):437-485.

Sloan, R. G. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review* 71 (3):289-315.

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.

Thomas, J., and F. Zhang. 2011. Tax expense momentum. *Journal of Accounting Research* 49 (3):791-821.

Thomas, J., and F. Zhang. 2014. Valuation of tax expense. *Review of Accounting Studies* 19 (4):1436-1467.

Watts, R. L., and R. W. Leftwich. 1977. The time series of annual accounting earnings. *Journal of Accounting Research* 15 (2):253-271.

Yan, X., and L. Zheng. 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. *The Review of Financial Studies* 30 (4):1382-1423.

Appendix 1: Variable definitions

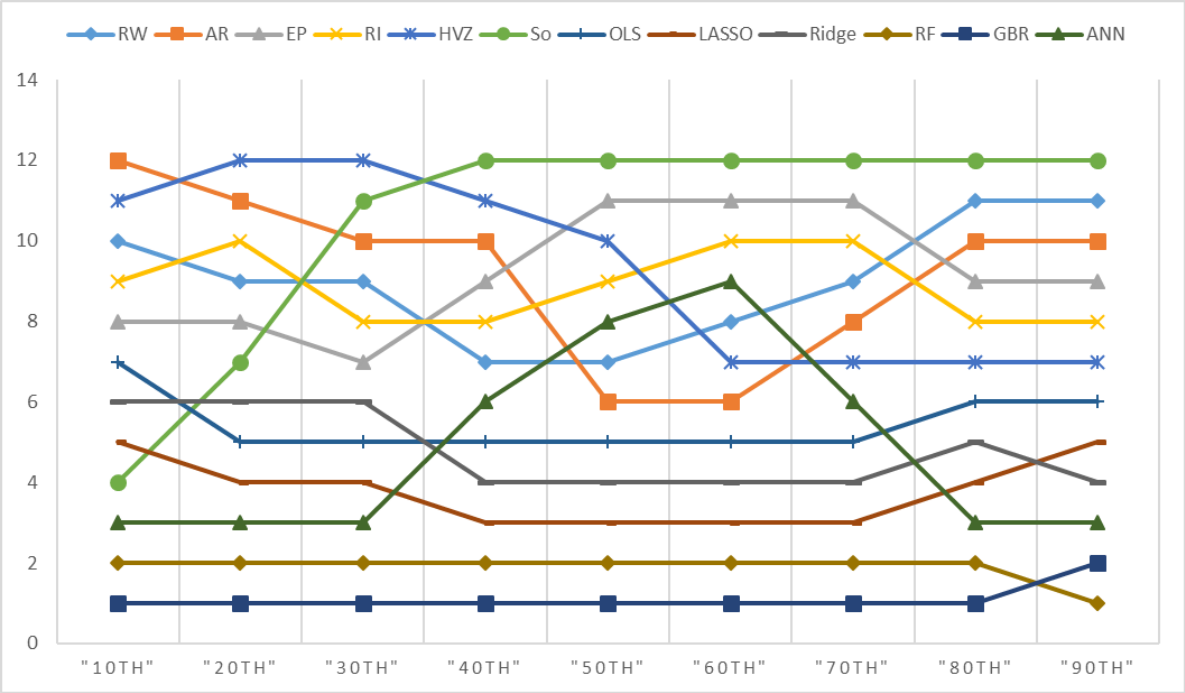
Variable	Definition
Earnings to be forecasted	
E_{t+1}	Earnings (ib - spi) in year t + 1
EPS_{t+1}	Earnings (ib - spi) in year t + 1 scaled by shares outstanding (csho)
Input features for extant models	
E_t	Earnings (ib - spi) in year t
A_t	Total assets (at)
D_t	Dividend payment (dvc)
DD_t	Dummy variable indicating dividend payers
$NegE_t$	Dummy variable indicating negative earnings
AC_t	Accruals calculated as change in non-cash current assets (act - che) minus change in current liabilities excluding short-term debt and taxes payable (lct - dlc - txp) minus depreciation and amortization (dp)
EPS_t^+	Earnings per share when earnings are positive, and zero otherwise
AC_t^-	Accruals per share when accruals are negative, and zero otherwise
AC_t^+	Accruals per share when accruals are positive, and zero otherwise
AG_t	Percentage change in total assets
NDD_t	Dummy variable indicating zero dividend per share
DIV_t	Dividend per share (dvpsx_f)
BTM_t	Book-to-market ratio, calculated as the book value of equity divided by the market equity as of three months after the end of the last fiscal year
$Price_t$	Stock price as of three months after the end of fiscal year t
BVE_t	Book value of equity (ceq)
$TACC_t$	Total accruals defined in Richardson et al. (2005), which is the sum of the change in WC (i.e., (act - che) - (lct - dlc)), change in NCO (i.e., (at - act - ivao) - (lt - lct - dltt)), and change in FIN (i.e., (ivst + ivao) - (dltt + dlc + pstk))
Input features for machine learning models	
Income statement items (# = 12):	
$SALE_t$	Sales (sale)
$COGS_t$	Cost of goods sold (cogs)
$XSGA_t$	Selling, general, and administrative expenses (xsga)
XAD_t	Advertising expense (xad)
XRD_t	Research and development (R&D) expense (xrd)
DP_t	Depreciation and amortization (dp)
$XINT_t$	Interest and related expense (xint)
$NOPIO_t$	Non-operating income (expense) – other (nopio)
TXT_t	Income taxes (txt)
$XIDO_t$	Extraordinary items and discontinued operations (xido)
E_t	Earnings (ib - spi)
DVC_t	Common dividend (dvc)
Balance sheet items (# = 15):	
CHE_t	Cash and short-term investments (che)
$INVT_t$	Inventories (invt)
$RECT_t$	Receivables (rect)
ACT_t	Total current assets (act)

$PPENT_t$	Property, plant, and equipment – Net (ppent)
$IVAO_t$	Investments and advances – other (ivao)
$INTAN_t$	Intangible assets (intan)
AT_t	Total assets (at)
AP_t	Accounts payable (ap)
DLC_t	Debt in current liabilities (dlc)
TXP_t	Income taxes payable (txp)
LCT_t	Total current liabilities (lct)
$DLTT_t$	Long-term debt (dltt)
LT_t	Total liabilities (lt)
CEQ_t	Common/Ordinary equity (ceq)
Cash flow statement items (# = 1):	
CFO_t	Cash flow from operating activities (oancf - xidoc); if missing, it is computed using the balance sheet approach (ib - accruals)
First-order differences of the above 28 items (# = 28):	
$\Delta CHE_t \sim \Delta CFO_t$	Computed as the corresponding item in year t less the same item in year t - 1
Dependent variables in the regressions	
$EXRET12M_{t+1}$	One-year ahead excess return, computed as the 12-month cumulative return less that of the risk-free rate, starting from the fourth month after the end of the last fiscal year
$FERR_{t+1}$	Analyst forecast error, computed as the realized difference between earnings per share (EPS) as reported in IBES and the consensus forecast made in the fourth month after the end of the last fiscal year, scaled by the stock price on the day of formation of the consensus forecast
Controls	
$SIZE_t$	Logarithm of market capitalization at the end of the third month after the end of the last fiscal year
BM_t	Book-to-market ratio, calculated as the book value of equity divided by market equity at the end of three months after the end of the last fiscal year
MOM_t	Momentum calculated as the cumulative return during the 11-month period starting 12 months ago
ROE_t	Earnings (ib - spi) divided by common equity (ceq)
INV_t	Growth rate of total assets ($\mathrm{at_t}/at_{t-1} - 1$)
ACC_t	Accruals scaled by total assets
LTG_t	Consensus long-term growth forecast in IBES

Appendix 2: Tuning of hyperparameters for the machine learning models.

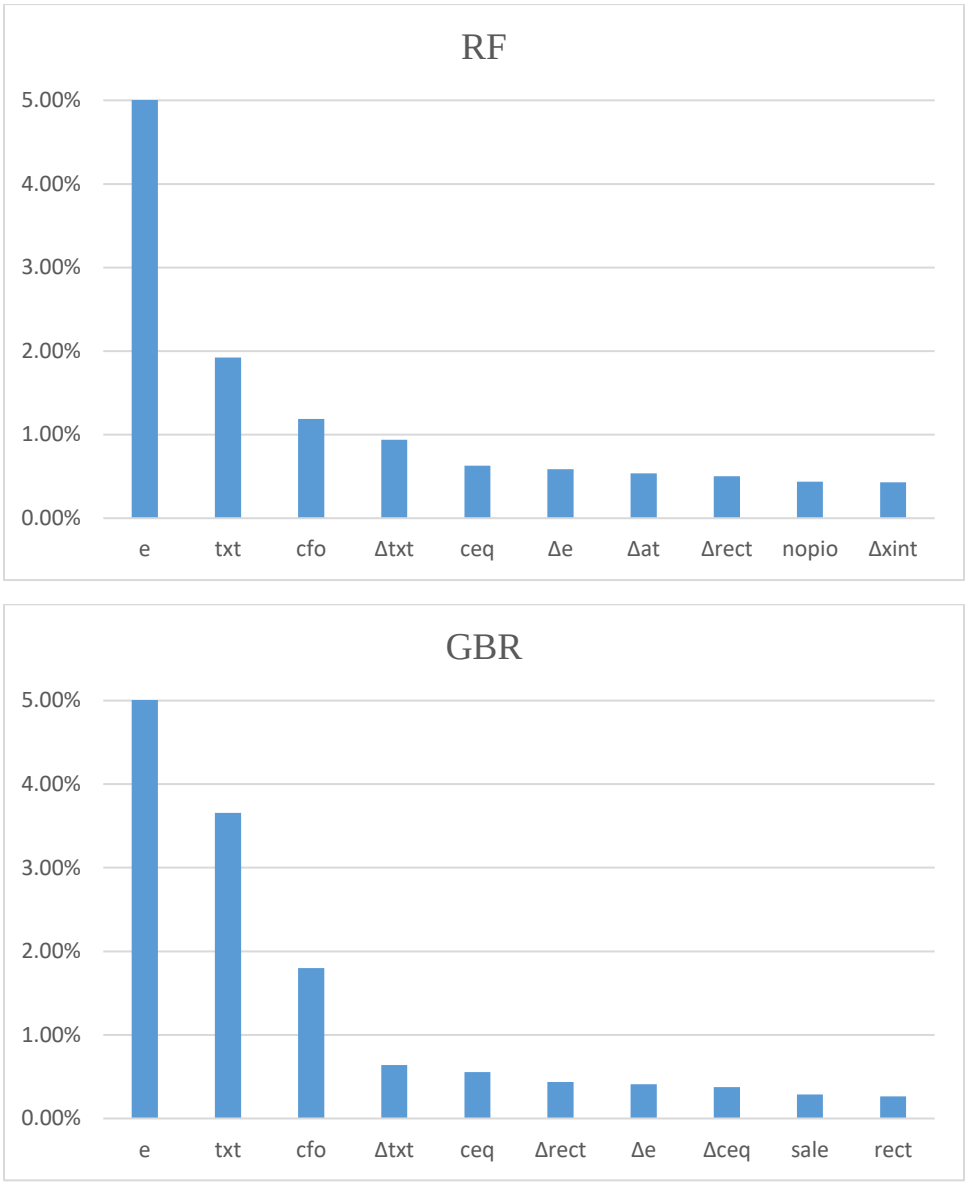
Model	Candidate values	Algorithms in sklearn
LASSO	alphas = np.linspace(1e-3,1e-1,1000)	*LassoCV(alphas=np.linspace(1e-3,1e-1,1000),fit_intercept=False,max_iter=25000,n_jobs=-1)*
Ridge	alphas = np.linspace(5e1,1e3,500)	*RidgeCV(alphas=np.linspace(5e1,1e3,500),fit_intercept=False,cv=5)*
RF	parameters = {'max_features':['auto'],'max_depth':[20,25,30,35],'min_samples_leaf':[15,20,25,50]}	*GridSearchCV(RandomForestRegressor(n_estimators=500,criterion='mse',oob_score=True,n_jobs=-1,random_state=10), parameters, cv=5, n_jobs=-1, scoring='neg_mean_squared_error')*
GBR	parameters = {'max_features':['auto'],'max_depth':[1,3,5],'min_samples_leaf':[75,100,125,150]}	*GridSearchCV(GradientBoostingRegressor(learning_rate=0.1,n_estimators=500,loss='huber',alpha=0.7,subsample=0.9,random_state=10), parameters, cv=5, n_jobs=-1, scoring='neg_mean_squared_error')*
ANN	parameters = {'activation':['relu','tanh'],'hidden_layer_sizes':[(64,32,16,8),(32,16,8,4),(16,8,4,2),(64,32,16),(32,16,8),(16,8,4),(8,4,2),(64,32),(32,16),(16,8),(8,4),(4,2),(64,),(32,),(16,),(8,),(4,)],'alpha':[1e-3,1e-4,1e-5]}	*BaggingRegressor(regr,random_state=0,n_estimators=10,max_samples=0.6,n_jobs=-1),where regr = GridSearchCV(MLPRegressor(max_iter=1000,random_state=10,early_stopping=True,tol=1e-6), parameters, cv=5, n_jobs=-1, scoring='neg_mean_squared_error')*

Figure 1: Rank of the average deciles of absolute forecast errors among different models.



This figure plots the rank of the time series average of the deciles of absolute forecast errors among the 12 individual models from 1975 to 2019. The absolute forecast error is defined as the absolute value of the difference between the actual one-year ahead earnings and the forecasted earnings of the various models, deflated by market equity three months after the end of the last fiscal year. The value of the rank ranges from 1 to 12. Rank = 1 and rank = 12 represent the most accurate and least accurate in terms of absolute forecast errors.

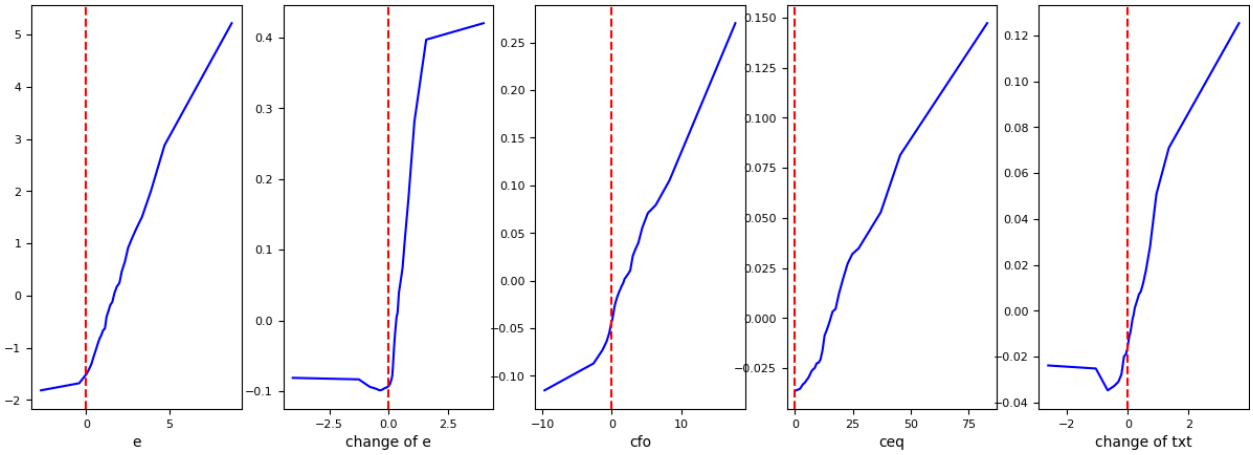
Figure 2: Top 10 influential features of random forest and gradient boosting regression.



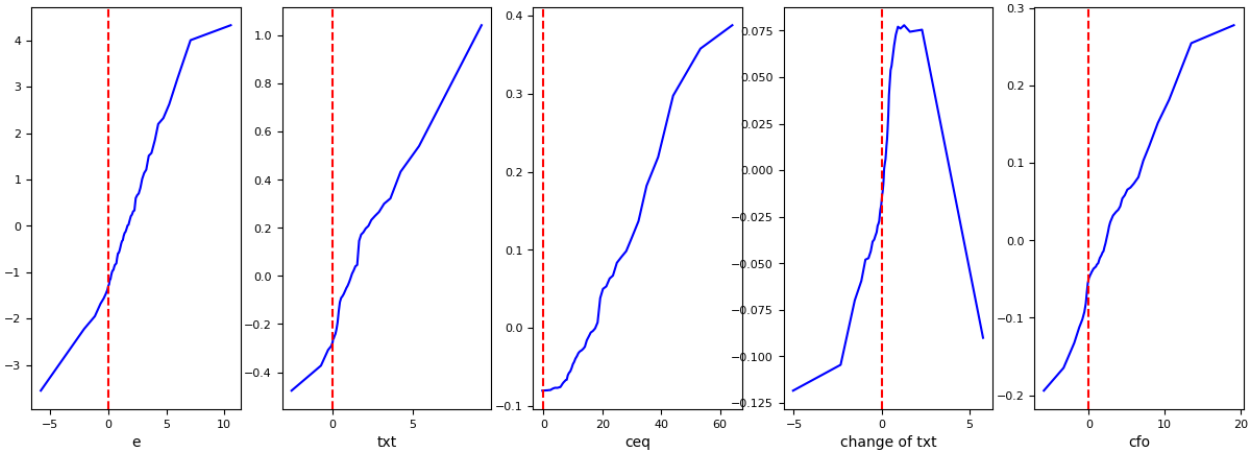
This figure plots the average feature importance extracted from the fitted models of random forest (RF) and gradient boosting regression (GBR) that we train with data from the 1975–2019 period. The higher the importance score, the more important the feature. To facilitate representation, we set the maximum of the y axis at 0.05, while the average feature importance values for earnings ("e") are 0.8187 and 0.8519 for RF and GBR, respectively.

Figure 3: A visualization of the relationships between important features and future earnings.

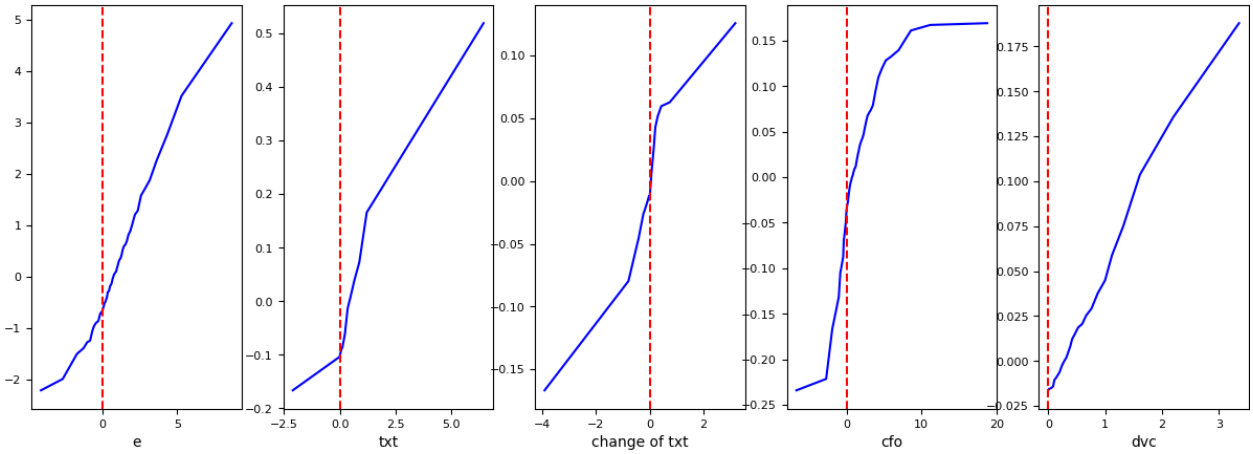
Panel A: Accumulated local effects (ALE) of the top five most influential features of the RF model for 1975.



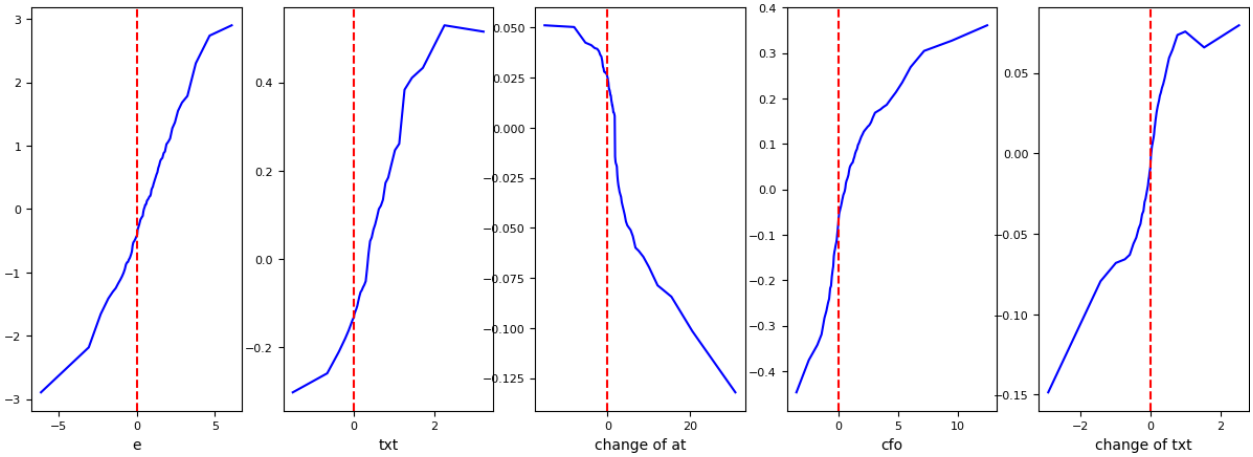
Panel B: Accumulated local effects (ALE) of the top five most influential features of the RF model for 1985.



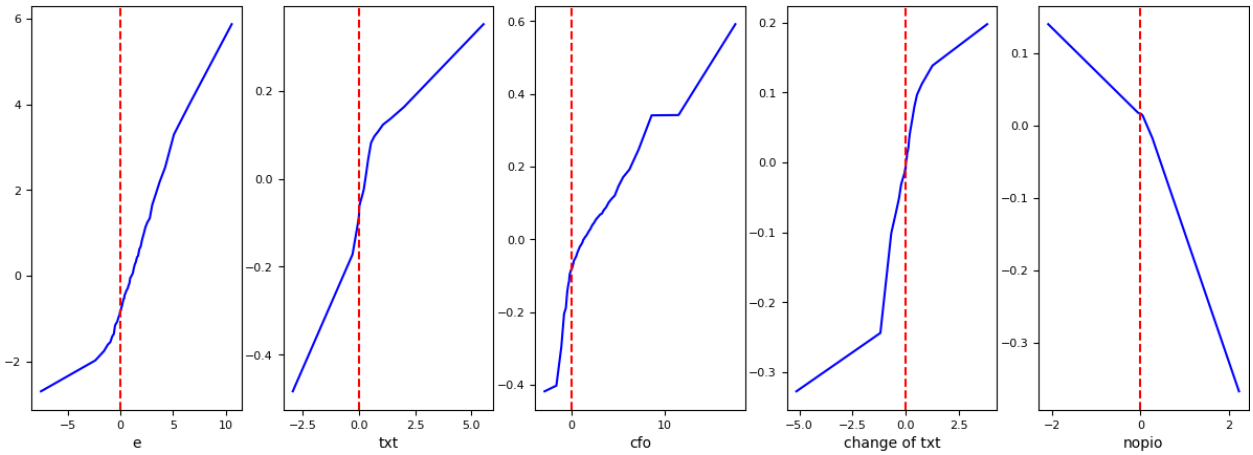
Panel C: Accumulated local effects (ALE) of the top five most influential features of the RF model for 1995.



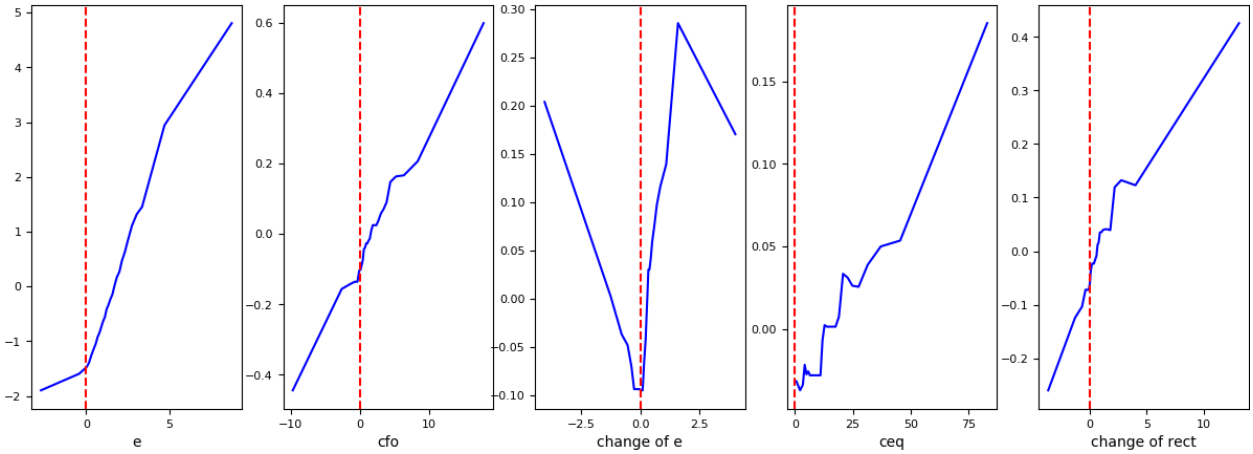
Panel D: Accumulated local effects (ALE) of the top five most influential features of the RF model for 2005.



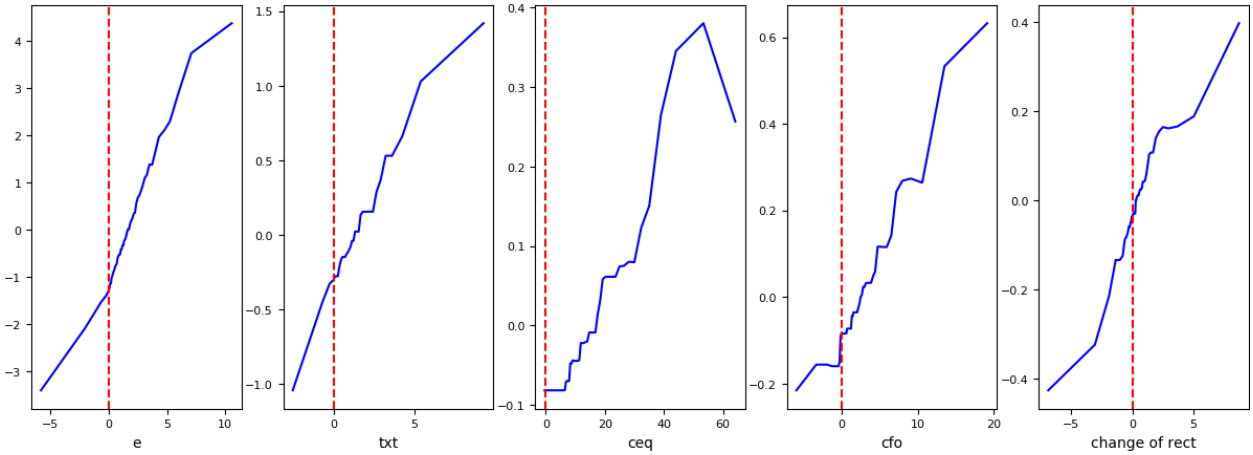
Panel E: Accumulated local effects (ALE) of the top five most influential features of the RF model for 2015.



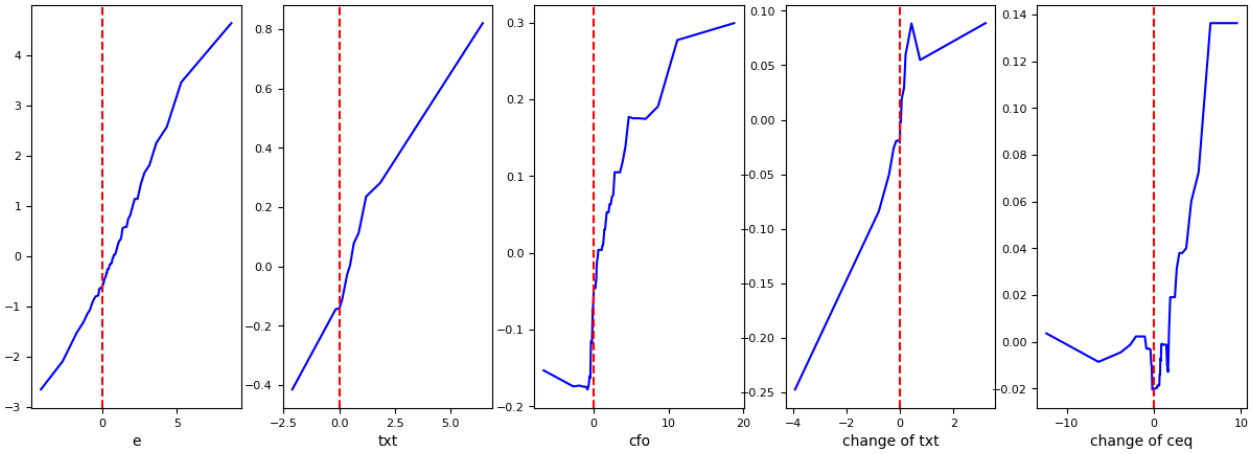
Panel F: Accumulated local effects (ALE) of the top five most influential features of the GBR model for 1975.



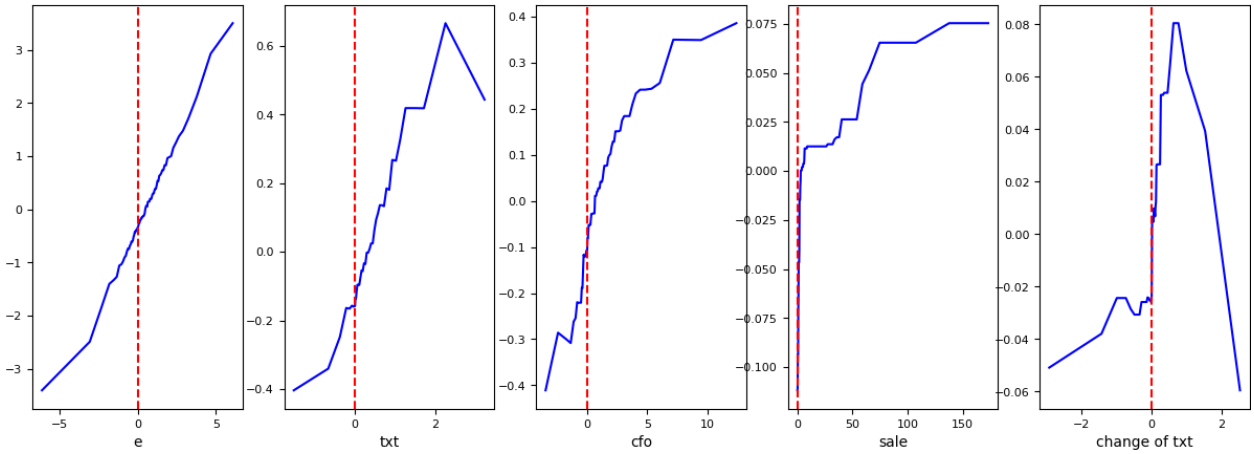
Panel G: Accumulated local effects (ALE) of the top five most influential features of the GBR model for 1985.



Panel H: Accumulated local effects (ALE) of the top five most influential features of the GBR model for 1995.



Panel I: Accumulated local effects (ALE) of the top five most influential features of the GBR model for 2005.



Panel J: Accumulated local effects (ALE) of the top five most influential features of the GBR model for 2015.

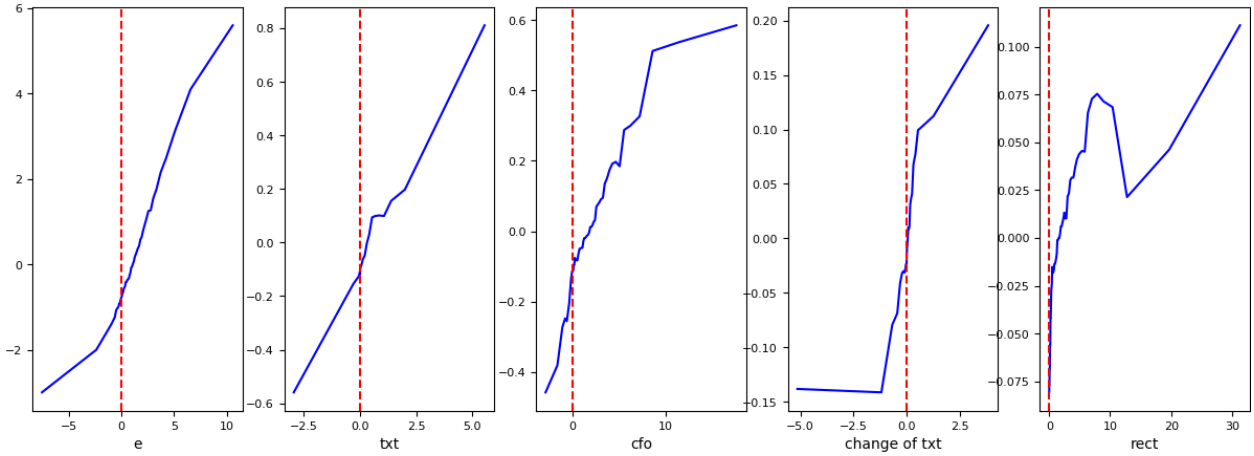


Table 1: Sample distribution by year.

year	# obs	year	# obs	year	# obs
1975	2,550	1990	3,029	2005	3,303
1976	2,558	1991	3,140	2006	3,259
1977	2,578	1992	3,484	2007	3,166
1978	2,593	1993	3,816	2008	2,945
1979	2,679	1994	4,236	2009	2,583
1980	2,694	1995	4,373	2010	2,747
1981	2,685	1996	4,690	2011	2,673
1982	2,689	1997	4,976	2012	2,538
1983	2,830	1998	4,930	2013	2,499
1984	2,892	1999	4,620	2014	2,522
1985	3,047	2000	4,540	2015	2,497
1986	3,087	2001	3,969	2016	2,419
1987	3,083	2002	3,595	2017	2,383
1988	3,219	2003	3,310	2018	2,349
1989	3,140	2004	3,378	2019	2,299

This table reports the number of firms with non-missing input features for all of the models from 1975 to 2019.

Table 2: Comparison of forecast accuracy.

	Mean absolute forecast errors				Median absolute forecast errors			
	Average	Comparison with RW			Average	Comparison with RW		
		DIFF	t-stat	%DIFF		DIFF	t-stat	%DIFF
Benchmark model								
RW	0.0764				0.0309			
Extant models								
AR	0.0755	-0.0009	-2.51	-1.15%	0.0308	-0.0001	-0.22	-0.24%
HVZ	0.0743	-0.0022	-3.63	-2.82%	0.0311	0.0002	0.64	0.76%
EP	0.0742	-0.0022	-2.79	-2.85%	0.0313	0.0004	1.02	1.42%
RI	0.0741	-0.0023	-3.15	-3.07%	0.0311	0.0002	0.66	0.74%
SO	0.0870	0.0105	5.19	13.78%	0.0347	0.0039	5.50	12.56%
Linear machine learning models								
OLS	0.0720	-0.0045	-5.04	-5.83%	0.0306	-0.0002	-0.60	-0.73%
LASSO	0.0716	-0.0048	-5.32	-6.31%	0.0304	-0.0004	-1.11	-1.43%
Ridge	0.0718	-0.0047	-5.19	-6.11%	0.0305	-0.0003	-0.87	-1.08%
Nonlinear machine learning models								
RF	0.0698	-0.0066	-6.44	-8.64%	0.0296	-0.0012	-3.10	-3.97%
GBR	0.0697	-0.0068	-6.08	-8.86%	0.0292	-0.0016	-4.23	-5.34%
ANN	0.0713	-0.0051	-5.38	-6.67%	0.0310	0.0001	0.24	0.38%
Composite models								
COMP_EXT	0.0737	-0.0027	-3.89	-3.58%	0.0311	0.0002	0.56	0.66%
COMP_LR	0.0717	-0.0047	-5.25	-6.16%	0.0305	-0.0004	-1.02	-1.33%
COMP_NL	0.0689	-0.0075	-6.99	-9.87%	0.0292	-0.0017	-3.92	-5.55%
COMP_ML	0.0693	-0.0071	-7.12	-9.35%	0.0294	-0.0015	-3.75	-4.81%

This table reports the time series average of the mean and median absolute forecast errors for the 12 individual models and the 4 composite models and their comparisons with the benchmark model (i.e., the RW model). The absolute forecast error is calculated as the absolute value of the difference between the actual one-year ahead earnings and the model-based earnings forecasts, scaled by market equity at the end of three months after the end of the last fiscal year. DIFF is the time series average of the difference calculated as the mean (median) absolute forecast error of each model minus that of the benchmark model. A negative DIFF value indicates an improvement in the forecast accuracy of the specific model relative to the benchmark model, and vice versa. The Newey–West t-statistic of DIFF is adjusted using three lags and reported accordingly. The percentage difference (%DIFF) is DIFF divided by the time series average of the mean (median) absolute forecast error of the benchmark model.

Table 3: Cross-sectional analysis of improvement in forecast accuracy (in percentage).

Panel A: Partition variable – ROA volatility					
	Low	2	3	4	High
COMP_LR vs. RW	-1.97	2.32	3.24	6.83	10.26
	(-1.96)	(2.98)	(4.84)	(6.78)	(8.53)
COMP_NL vs. RW	4.44	5.68	5.70	8.99	15.21
	(5.81)	(6.20)	(7.17)	(7.64)	(9.54)
Panel B: Partition variable – |Total accruals|/Total assets					
	Low	2	3	4	High
COMP_LR vs. RW	0.19	1.45	3.11	6.77	11.47
	(0.21)	(1.51)	(3.72)	(7.60)	(7.52)
COMP_NL vs. RW	3.25	4.32	6.00	10.07	17.71
	(4.71)	(6.13)	(7.47)	(8.43)	(9.52)
Panel C: Partition variable – |Working capital accruals|/Total assets					
	Low	2	3	4	High
COMP_LR vs. RW	1.47	2.56	4.02	4.99	10.34
	(1.18)	(2.94)	(4.62)	(5.74)	(8.73)
COMP_NL vs. RW	6.59	7.44	7.69	8.34	13.68
	(5.69)	(6.16)	(6.96)	(8.60)	(8.90)
Panel D: Partition variable – R&D expense/Total assets					
	MISSING	Low	2	3	High
COMP_LR vs. RW	6.09	4.95	7.24	6.67	3.52
	(6.38)	(5.03)	(8.24)	(6.21)	(1.56)
COMP_NL vs. RW	8.90	8.54	10.17	10.51	11.27
	(7.07)	(6.89)	(10.92)	(9.02)	(8.43)
Panel E: Partition variable – Loss dummy					
	Non-loss				Loss
COMP_LR vs. RW	3.19				8.41
	(3.61)				(5.29)
COMP_NL vs. RW	6.41				12.89
	(5.38)				(7.93)

This table presents a cross-sectional analysis of the percentage improvement in the forecast accuracy of COMP_LR and COMP_NL compared with that of the RW model. The percentage improvement is defined as the time series average of the difference in the mean absolute forecast errors between the pairs (i.e., the composite model versus the RW model) divided by the mean absolute forecast error of the RW model. A positive number indicates an improvement in the accuracy of the composite model. In panels A, B, and C, we sort all firms into quintiles for each year based on the magnitude of the partition variable (ROA volatility, absolute value of total accruals divided by total assets, and absolute value of working capital accruals divided by total assets, respectively). In Panel D, we classify all firms with missing R&D expense into a separate group and sort the remaining firms into quartiles for each year based on their R&D expense divided by total assets. In Panel E, we divide all firms into two groups for each year depending on whether their earnings are negative. Then, we calculate the percentage improvement for each subgroup along with the Newey–West t-statistics (in brackets) with three lags.

Table 4: Information content analysis.

Panel A: Relative information content of the earnings prediction models.

	Correlation analysis		Univariate regression $ECH = \beta_0 + \beta_1 FECH_{.} + \varepsilon$		
	Pearson	Spearman	β_1	t-stat (β_1)	Avg. R^2 (%)
Extant models					
AR	0.199	0.117	0.0304	3.43	8.07
HVZ	0.283	0.179	0.0422	8.21	9.98
EP	0.321	0.154	0.0480	9.14	12.22
RI	0.313	0.148	0.0467	9.34	11.68
SO	0.291	0.153	0.0440	9.60	9.66
Linear machine learning models					
OLS	0.369	0.245	0.0546	10.25	14.87
LASSO	0.372	0.247	0.0550	10.18	15.12
Ridge	0.372	0.247	0.0550	10.27	15.12
Nonlinear machine learning models					
RF	0.396	0.279	0.0581	11.02	17.36
GBR	0.395	0.283	0.0582	10.99	17.15
ANN	0.396	0.276	0.0580	12.01	16.95
Composite models					
COMP_EXT	0.333	0.188	0.0497	9.44	12.73
COMP_LR	0.372	0.247	0.0550	10.28	15.09
COMP_NL	0.413	0.300	0.0605	11.45	18.57
COMP_ML	0.408	0.286	0.0601	10.94	18.09

Panel B: Incremental information content of the machine learning models

Multivariate regression: $ECH = \beta_0 + \beta_1 FECH_{ML} + \beta_2 FECH_{AR} + \beta_3 FECH_{HVZ} + \beta_4 FECH_{EP} + \beta_5 FECH_{RI} + \beta_6 FECH_{SO} + \varepsilon$								
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	Avg. R^2 (%)
Linear machine learning models								
OLS	0.0016	0.0432	0.0107	-0.0058	0.0004	-0.0098	0.0251	18.99
	(0.57)	(11.90)	(1.56)	(-1.42)	(0.03)	(-0.82)	(8.82)	
LASSO	0.0016	0.0458	0.0085	-0.0072	0.0017	-0.0111	0.0251	19.09
	(0.57)	(15.45)	(1.28)	(-1.72)	(0.13)	(-0.87)	(8.72)	
Ridge	0.0016	0.0453	0.009	-0.0068	0.0019	-0.0113	0.0251	19.09
	(0.57)	(12.19)	(1.36)	(-1.66)	(0.14)	(-0.89)	(8.71)	
Nonlinear machine learning models								
RF	0.0016	0.049	0.0105	-0.0072	-0.0043	-0.0014	0.0146	19.53
	(0.57)	(16.83)	(1.60)	(-1.71)	(-0.30)	(-0.12)	(3.89)	
GBR	0.0016	0.0497	0.0086	-0.0079	-0.0005	-0.006	0.0183	19.63
	(0.57)	(16.40)	(1.42)	(-1.91)	(-0.03)	(-0.54)	(5.54)	
ANN	0.0016	0.0466	0.0078	-0.0047	0.0111	-0.0137	0.0176	20.20
	(0.57)	(16.24)	(1.29)	(-1.17)	(0.78)	(-1.17)	(5.15)	
Composite models								
COMP_LR	0.0016	0.045	0.0094	-0.0068	0.0016	-0.011	0.025	19.08
	(0.57)	(12.27)	(1.41)	(-1.64)	(0.12)	(-0.88)	(8.82)	
COMP_NL	0.0016	0.059	0.0075	-0.0087	0.0053	-0.0144	0.0132	20.84
	(0.57)	(17.91)	(1.30)	(-2.11)	(0.36)	(-1.25)	(3.92)	
COMP_ML	0.0016	0.0593	0.0071	-0.0104	0.0081	-0.0199	0.0175	20.80
	(0.57)	(16.22)	(1.20)	(-2.44)	(0.63)	(-1.71)	(6.24)	

Panel A reports the average Person and Spearman correlation coefficients between the forecasted earnings changes calculated using various models and the actual earnings changes over 45 years from 1975 to 2019, as well as the univariate Fama–MacBeth regression results. In the regression, all forecasted earnings changes are standardized to have zero mean and unit variance each year. Panel B reports the multivariate Fama–MacBeth regression results. Specifically, we regress ECH on FECH using the six machine learning models and the three composite models and controlling for all earnings changes predicted using the extant models. All independent variables are standardized to have zero mean and unit variance each year. All earnings changes are scaled by market equity at the end of three months after the end of the last fiscal year. The table presents the average coefficients along with the Newey–West t-statistics (in brackets) with three lags and the average adjusted R-square. The subscripts are omitted for brevity.

Table 5: ERC analysis.

	Announcement ERC				Annual ERC			
			Comparison with RW				Comparison with RW	
	ERC	t-stat	DIFF	t-stat	ERC	t-stat	DIFF	t-stat
Benchmark model								
RW	0.0399	28.70			0.1479	9.92		
Extant models								
AR	0.0398	27.32	-0.0001	-0.18	0.1456	10.10	-0.0023	-1.98
HVZ	0.0399	26.15	0.0000	0.05	0.1429	9.37	-0.0050	-1.89
EP	0.0410	25.12	0.0011	1.52	0.1443	9.64	-0.0036	-0.97
RI	0.0411	26.64	0.0012	1.85	0.1450	9.66	-0.0029	-0.75
SO	0.0372	19.82	-0.0027	-2.42	0.1385	9.03	-0.0094	-2.04
Linear machine learning models								
OLS	0.0410	26.43	0.0011	1.56	0.1415	10.40	-0.0064	-1.55
LASSO	0.0412	26.76	0.0013	1.91	0.1430	10.41	-0.0049	-1.29
Ridge	0.0411	26.88	0.0012	1.76	0.1422	10.33	-0.0057	-1.44
Nonlinear machine learning models								
RF	0.0412	24.49	0.0013	1.71	0.1494	10.60	0.0015	0.55
GBR	0.0404	26.05	0.0005	0.73	0.1447	10.77	-0.0032	-0.82
ANN	0.0410	29.56	0.0011	1.74	0.1511	12.25	0.0032	0.68
Composite models								
COMP_EXT	0.0416	25.43	0.0017	2.81	0.1510	10.19	0.0031	1.16
COMP_LR	0.0411	26.80	0.0012	1.80	0.1422	10.39	-0.0056	-1.41
COMP_NL	0.0415	26.03	0.0016	2.34	0.1508	11.13	0.0029	0.91
COMP_ML	0.0418	26.96	0.0019	2.81	0.1488	11.14	0.0009	0.26

This table reports the time series average of the announcement ERC and the annual ERC computed using the 12 individual models and the 4 composite models and compares them with the values obtained using the benchmark RW model) along with Newey–West t-statistics with 3 lags. The announcement ERC is estimated by regressing the sum of the quarterly earnings announcement returns (market-adjusted, from day -1 to day +1) over the next fiscal year on standardized unexpected earnings. Standardized unexpected earnings are calculated as the difference between future actual earnings and the model forecasts, deflated by market capitalization at the end of three months after the end of the last fiscal year and then standardized to have zero mean and unit variance each year. The annual ERC is estimated by regressing the buy-and-hold returns over the next year starting from the fourth month after the end of the last fiscal year on standardized unexpected earnings over the same period. The pairwise difference (DIFF) is calculated as the time series average of the difference between the ERC values of various models and that of the RW model each year from 1975 to 2019.

Table 6: Regression analysis of future excess stock returns on the new information uncovered using the machine learning models.

Multivariate regression: $EXRET12M = \beta_0 + \beta_1 ML\_RESD + \beta_2 SIZE + \beta_3 BM + \beta_4 MOM + \beta_5 ROE + \beta_6 INV + \beta_7 ACC + IndustryFE + \varepsilon.$									
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	Avg. R^2
Linear machine learning models									
OLS	0.077	0.821	-0.005	0.068	-0.026	0.104	-0.056	-0.113	4.35
	(0.62)	(7.21)	(-0.88)	(6.97)	(-1.29)	(3.18)	(-8.68)	(-2.01)	
LASSO	0.107	0.908	-0.005	0.068	-0.027	0.105	-0.057	-0.108	4.36
	(0.80)	(6.67)	(-0.88)	(6.96)	(-1.31)	(3.17)	(-8.79)	(-1.91)	
Ridge	0.141	0.890	-0.005	0.068	-0.027	0.105	-0.057	-0.110	4.37
	(1.10)	(7.11)	(-0.88)	(6.92)	(-1.32)	(3.18)	(-8.77)	(-1.96)	
Nonlinear machine learning models									
RF	0.227	0.721	-0.006	0.068	-0.028	0.101	-0.050	-0.111	4.30
	(1.32)	(5.22)	(-1.00)	(7.01)	(-1.40)	(3.23)	(-7.80)	(-1.86)	
GBR	0.085	0.790	-0.006	0.067	-0.028	0.099	-0.054	-0.106	4.35
	(0.53)	(6.08)	(-1.02)	(6.89)	(-1.38)	(3.18)	(-7.89)	(-1.79)	
ANN	0.172	0.679	-0.006	0.067	-0.027	0.104	-0.055	-0.108	4.40
	(1.33)	(5.24)	(-1.00)	(6.91)	(-1.37)	(3.18)	(-8.85)	(-1.87)	
Composite models									
COMP_LR	0.112	0.887	-0.005	0.068	-0.027	0.104	-0.057	-0.110	4.36
	(0.96)	(7.03)	(-0.88)	(6.94)	(-1.31)	(3.18)	(-8.81)	(-1.95)	
COMP_NL	0.226	0.958	-0.006	0.067	-0.029	0.100	-0.053	-0.100	4.42
	(1.32)	(5.33)	(-1.04)	(6.90)	(-1.49)	(3.21)	(-8.22)	(-1.67)	
COMP_ML	0.106	1.072	-0.006	0.067	-0.029	0.103	-0.056	-0.100	4.42
	(0.86)	(6.43)	(-0.98)	(6.86)	(-1.44)	(3.18)	(-8.52)	(-1.72)	

This table reports the Fama–MacBeth regression results that regress one-year ahead excess returns starting from the fourth month after the end of fiscal year t on the new information uncovered using the machine learning models (ML_RESD), controlling for various known return-predicting factors and industry fixed effects (3-digit SIC): $EXRET12M = \beta_0 + \beta_1 ML\_RESD + \beta_2 SIZE + \beta_3 BM + \beta_4 MOM + \beta_5 ROE + \beta_6 INV + \beta_7 ACC + IndustryFE + \varepsilon$. ML_RESD is estimated as the residual by regressing the machine-learning-based forecasts on the RW model and the five extant models each year. The definitions of the control variables are given in Appendix 1. All independent variables are winsorized at 1% and 99% each year. The table presents the average coefficients with the Newey–West t-statistics (in brackets) with three lags and the average adjusted R-square. The subscripts are omitted for brevity.

Table 7: Portfolio analysis of the new information uncovered using the machine learning models.

Panel A: Equal-weighted portfolios

	OLS	LASSO	Ridge	RF	GBR	ANN	COMP_LR	COMP_NL	COMP_ML
Mean Return	0.6185	0.6262	0.6346	0.5962	0.6795	0.7185	0.6402	0.7203	0.7720
	(8.65)	(8.89)	(8.85)	(7.49)	(8.73)	(8.12)	(9.29)	(8.05)	(9.50)
CAPM Alpha	0.6817	0.6856	0.6989	0.6328	0.7110	0.7784	0.7022	0.7695	0.8372
	(9.96)	(10.46)	(10.48)	(7.82)	(9.07)	(8.89)	(10.87)	(8.78)	(10.73)
FF3 Alpha	0.6538	0.6597	0.6758	0.6062	0.6733	0.7247	0.6761	0.7279	0.8033
	(9.71)	(9.88)	(10.18)	(8.54)	(9.90)	(9.63)	(10.46)	(9.61)	(11.39)
Carhart4 Alpha	0.5938	0.5921	0.6178	0.5166	0.5934	0.6558	0.6137	0.6448	0.7134
	(9.08)	(9.03)	(9.49)	(7.29)	(8.57)	(8.50)	(9.66)	(8.35)	(10.23)
FF5 Alpha	0.5371	0.5488	0.5655	0.4312	0.4828	0.5286	0.5613	0.5143	0.6096
	(7.96)	(8.21)	(8.48)	(5.97)	(7.08)	(7.18)	(8.64)	(6.63)	(8.59)

Panel B: Value-weighted portfolios

	OLS	LASSO	Ridge	RF	GBR	ANN	COMP_LR	COMP_NL	COMP_ML
Mean Return	0.2239	0.2484	0.2674	0.3177	0.4163	0.4747	0.2677	0.4568	0.3831
	(1.99)	(2.19)	(2.27)	(2.74)	(3.50)	(4.08)	(2.29)	(3.74)	(3.60)
CAPM Alpha	0.3571	0.3778	0.3969	0.3775	0.4797	0.5914	0.3954	0.5490	0.4884
	(3.30)	(3.57)	(3.53)	(3.05)	(4.01)	(5.07)	(3.58)	(4.34)	(4.66)
FF3 Alpha	0.3237	0.3552	0.3667	0.4478	0.5505	0.6325	0.3663	0.6217	0.5289
	(3.34)	(3.53)	(3.54)	(3.75)	(4.60)	(5.52)	(3.65)	(5.19)	(5.15)
Carhart4 Alpha	0.2829	0.2999	0.3320	0.3081	0.4316	0.5605	0.3247	0.4768	0.4558
	(3.08)	(3.06)	(3.41)	(3.07)	(3.70)	(4.70)	(3.37)	(4.49)	(4.23)
FF5 Alpha	0.1222	0.1205	0.1634	0.2810	0.4142	0.4358	0.1575	0.4119	0.3715
	(1.42)	(1.40)	(1.90)	(2.57)	(3.80)	(4.40)	(1.85)	(3.54)	(3.89)

This table summarizes the return spread between the extreme quintiles sorted based on the new information uncovered using the machine learning models. At the beginning of each month, we estimate the new information component as the residual from the regression of the machine-learning-based forecasts on the forecasts generated using the RW model and the five extant models across all firms. We sort the stocks into quintiles based on the resulting residual forecasts for each 3-digit SIC industry and report the return performance of the hedge portfolio, which takes long positions in quintiles with the most favorable new information and short positions in quintiles with the least favorable new information. Panel A reports the results for the equal-weighted portfolios. Panel B reports the results for the value-weighted hedge portfolios.

Table 8: Comparison of forecast accuracy with alternative deflators.

	Forecast errors deflated by total assets				Per share forecast errors			
	Average	Comparison with RW			Average	Comparison with RW		
		DIFF	t-stat	%DIFF		DIFF	t-stat	%DIFF
Benchmark model								
RW	0.0593				0.7378			
Extant models								
AR	0.0595	0.0002	1.03	0.38%	0.7436	0.0058	1.25	0.79%
HVZ	0.0597	0.0005	1.02	0.77%	0.7350	-0.0028	-0.52	-0.38%
EP	0.0612	0.0020	1.78	3.30%	0.7283	-0.0095	-1.50	-1.28%
RI	0.0609	0.0016	1.67	2.78%	0.7272	-0.0106	-1.73	-1.44%
SO	0.0773	0.0180	6.11	30.34%	0.7594	0.0216	1.82	2.93%
Linear machine learning models								
OLS	0.0590	-0.0003	-0.35	-0.49%	0.7077	-0.0301	-4.92	-4.08%
LASSO	0.0588	-0.0005	-0.56	-0.79%	0.7045	-0.0333	-5.46	-4.52%
Ridge	0.0589	-0.0004	-0.42	-0.60%	0.7060	-0.0318	-5.28	-4.31%
Nonlinear machine learning models								
RF	0.0546	-0.0047	-7.74	-7.90%	0.6815	-0.0563	-8.54	-7.63%
GBR	0.0543	-0.0050	-6.80	-8.40%	0.6712	-0.0666	-8.62	-9.02%
ANN	0.0570	-0.0023	-3.08	-3.82%	0.7048	-0.0330	-5.09	-4.48%
Composite models								
COMP_EXT	0.0597	0.0005	0.73	0.79%	0.7198	-0.0180	-3.12	-2.44%
COMP_LR	0.0588	-0.0004	-0.50	-0.71%	0.7055	-0.0323	-5.25	-4.38%
COMP_NL	0.0542	-0.0051	-7.22	-8.59%	0.6701	-0.0677	-8.92	-9.17%
COMP_ML	0.0553	-0.0040	-5.50	-6.68%	0.6770	-0.0608	-8.23	-8.24%

This table reports the time series average of the mean absolute forecast errors for the 12 individual models and the 4 composite models and compares them with the values of the benchmark RW model. In the four columns on the left, the absolute forecast error is calculated as the absolute value of the difference between the actual one-year ahead earnings and the model-based earnings forecasts scaled by total assets. In the four columns on the right, the absolute forecast error is calculated as the absolute value of the difference between the actual one-year ahead earnings and the model-based earnings forecasts scaled by shares outstanding at the end of the fiscal year (i.e., per share forecast errors). DIFF is the time series average of the difference calculated as the mean absolute forecast error of each model minus that of the benchmark model. A negative DIFF value indicates an improvement in the forecast accuracy of the specific model relative to that of the benchmark model, and vice versa. The Newey–West t-statistic of DIFF is adjusted using three lags and reported accordingly. The percentage difference (%DIFF) is DIFF divided by the time series average of the mean (median) absolute forecast error of the benchmark model.

Table 9: Comparison of forecast accuracy of longer horizon forecasts.

Panel A: Accuracy of two-year ahead earnings forecasts

	Mean absolute forecast errors				Median absolute forecast errors			
	Average	Comparison with RW			Average	Comparison with RW		
		DIFF	t-stat	%DIFF		DIFF	t-stat	%DIFF
Benchmark model								
RW	0.1028				0.0473			
Extant models								
AR	0.1018	-0.0010	-1.21	-0.95%	0.0470	-0.0002	-0.36	-0.48%
HVZ	0.0971	-0.0057	-4.80	-5.50%	0.0462	-0.0011	-1.60	-2.36%
EP	0.0964	-0.0064	-3.94	-6.20%	0.0466	-0.0007	-0.83	-1.52%
RI	0.0956	-0.0071	-4.55	-6.94%	0.0460	-0.0012	-1.53	-2.61%
SO	0.1031	0.0003	0.18	0.32%	0.0491	0.0018	1.86	3.82%
Linear machine learning models								
OLS	0.0954	-0.0074	-4.66	-7.19%	0.0463	-0.0009	-1.07	-2.01%
LASSO	0.0944	-0.0084	-5.49	-8.15%	0.0459	-0.0014	-1.64	-2.96%
Ridge	0.0946	-0.0082	-5.08	-7.94%	0.0460	-0.0012	-1.34	-2.60%
Nonlinear machine learning models								
RF	0.0917	-0.0110	-6.81	-10.75%	0.0448	-0.0024	-2.81	-5.15%
GBR	0.0921	-0.0107	-6.13	-10.39%	0.0449	-0.0024	-2.88	-4.99%
ANN	0.0942	-0.0086	-5.63	-8.35%	0.0463	-0.0009	-1.06	-1.98%
Composite models								
COMP_EXT	0.0954	-0.0074	-5.52	-7.20%	0.0457	-0.0016	-1.99	-3.29%
COMP_LR	0.0947	-0.0081	-5.14	-7.88%	0.0460	-0.0013	-1.43	-2.69%
COMP_NL	0.0911	-0.0116	-7.04	-11.33%	0.0444	-0.0029	-3.49	-6.12%
COMP_ML	0.0915	-0.0112	-6.80	-10.93%	0.0445	-0.0028	-3.22	-5.83%

Panel B: Accuracy of three-year ahead earnings forecasts

	Mean absolute forecast errors				Median absolute forecast errors			
	Average	Comparison with RW			Average	Comparison with RW		
		DIFF	t-stat	%DIFF		DIFF	t-stat	%DIFF
Benchmark model								
RW	0.1225				0.0592			
Extant models								
AR	0.1227	0.0002	0.12	0.13%	0.0593	0.0001	0.06	0.11%
HVZ	0.1142	-0.0083	-4.46	-6.81%	0.0573	-0.0020	-1.75	-3.32%
EP	0.1138	-0.0087	-3.59	-7.09%	0.0579	-0.0014	-1.13	-2.29%
RI	0.1121	-0.0104	-4.51	-8.45%	0.0570	-0.0022	-1.77	-3.73%
SO	0.1203	-0.0022	-1.13	-1.82%	0.0611	0.0018	1.27	3.11%
Linear machine learning models								
OLS	0.1134	-0.0091	-3.83	-7.39%	0.0580	-0.0012	-0.86	-2.10%
LASSO	0.1125	-0.0100	-4.38	-8.20%	0.0575	-0.0018	-1.25	-2.96%
Ridge	0.1127	-0.0098	-4.25	-8.04%	0.0576	-0.0017	-1.19	-2.82%
Nonlinear machine learning models								
RF	0.1102	-0.0123	-5.19	-10.06%	0.0568	-0.0025	-1.68	-4.19%
GBR	0.1104	-0.0121	-4.85	-9.90%	0.0570	-0.0023	-1.59	-3.85%
ANN	0.1135	-0.0090	-4.16	-7.38%	0.0582	-0.0010	-0.63	-1.67%
Composite models								
COMP_EXT	0.1123	-0.0102	-5.02	-8.32%	0.0568	-0.0024	-2.02	-4.09%
COMP_LR	0.1127	-0.0098	-4.21	-7.98%	0.0577	-0.0016	-1.13	-2.67%
COMP_NL	0.1092	-0.0133	-5.51	-10.84%	0.0560	-0.0033	-2.22	-5.53%
COMP_ML	0.1093	-0.0132	-5.49	-10.76%	0.0558	-0.0035	-2.43	-5.85%

This table reports the time series average of the mean and median absolute forecast errors for the 12 individual models and the 4 composite models and compares them with the corresponding values of the benchmark RW model. In Panel A, the absolute forecast error is calculated as the absolute difference between the actual two-year ahead earnings and the corresponding model-based earnings forecasts scaled by market equity at the end of three months after the end of the last fiscal year. In Panel B, the absolute forecast error is calculated as the absolute difference between the actual three-year ahead earnings and the corresponding model-based earnings forecasts scaled by market equity at the end of three months after the end of the last fiscal year. DIFF is the time series average of the difference calculated as the mean (median) absolute forecast error of each model minus that of the benchmark model. A negative DIFF value represents an improvement in the forecast accuracy of the specific model relative to that of the benchmark model, and vice versa. The Newey–West t-statistic of DIFF is adjusted using three lags and reported accordingly. The percentage difference (%DIFF) is DIFF divided by the time series average of the mean (median) absolute forecast error of the benchmark model.

Table 10: Regression of analyst forecast errors on the new information uncovered using the machine learning models.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	Avg. R^2(%)
Linear machine learning models								
OLS	-0.053	0.242	0.008	-0.028	0.030	0.014	-0.031	13.20
	(-6.03)	(4.00)	(5.68)	(-4.91)	(3.76)	(1.47)	(-2.98)	
LASSO	-0.053	0.272	0.008	-0.028	0.029	0.014	-0.032	13.25
	(-6.03)	(3.84)	(5.68)	(-4.89)	(3.77)	(1.46)	(-2.94)	
Ridge	-0.053	0.258	0.008	-0.028	0.030	0.014	-0.032	13.19
	(-6.04)	(4.16)	(5.68)	(-4.90)	(3.76)	(1.47)	(-2.95)	
Nonlinear machine learning models								
RF	-0.053	0.248	0.008	-0.028	0.029	0.018	-0.030	12.93
	(-5.96)	(3.34)	(5.69)	(-4.86)	(3.84)	(1.79)	(-3.03)	
GBR	-0.053	0.184	0.008	-0.028	0.030	0.017	-0.030	12.91
	(-5.97)	(3.11)	(5.67)	(-4.83)	(3.86)	(1.72)	(-3.12)	
ANN	-0.053	0.204	0.008	-0.028	0.030	0.017	-0.029	13.12
	(-6.03)	(3.44)	(5.71)	(-4.84)	(3.83)	(1.76)	(-2.87)	
Composite models								
COMP_LR	-0.053	0.259	0.008	-0.028	0.030	0.014	-0.031	13.21
	(-6.03)	(4.01)	(5.68)	(-4.90)	(3.77)	(1.47)	(-2.95)	
COMP_NL	-0.053	0.251	0.008	-0.028	0.029	0.018	-0.029	13.04
	(-5.99)	(3.27)	(5.68)	(-4.85)	(3.86)	(1.81)	(-2.97)	
COMP_ML	-0.053	0.282	0.008	-0.028	0.029	0.017	-0.030	13.15
	(-6.01)	(3.55)	(5.68)	(-4.88)	(3.82)	(1.71)	(-2.92)	

This table reports the Fama–MacBeth regression results that regress the analyst forecast error (FERR) on the new information uncovered using the machine learning models (ML_RESD) with firm-specific controls: $FERR = \beta_0 + \beta_1 ML\_RESD + \beta_2 SIZE + \beta_3 BM + \beta_4 MOM + \beta_5 ACC + \beta_6 LTG + \varepsilon$. FERR is estimated as the realized difference between earnings per share (EPS) as reported in IBES and the consensus EPS forecast made fourth months after the end of the fiscal year, scaled by the stock price on the day of generation of the consensus forecast. ML_RESD is estimated as the residual by regressing the machine-learning-based forecasts on that of the RW model and the five extant models each year. The definitions of the control variables can be found in Appendix 1. All independent variables are winsorized at 1% and 99% each year. The table presents the average coefficients and the Newey–West t-statistics (in brackets) with three lags and the average adjusted R-square. The subscripts are omitted for brevity.

Table 11: Improving the extant models with insights from the nonlinear machine learning models.

Panel A: Effect of augmenting the extant models with TXT and ΔTXT

	Original model	Augmented model	Difference	t-stat	% Difference
RW	0.0764				
AR	0.0755	0.0719	-0.0037	-6.47	-4.90%
HVZ	0.0743	0.0719	-0.0024	-5.37	-3.23%
EP	0.0742	0.0723	-0.0019	-4.00	-2.56%
RI	0.0741	0.0720	-0.0020	-4.59	-2.70%
SO	0.0870	0.0848	-0.0022	-6.32	-2.53%

Panel B: Linear models using the top five most important features

	Mean absolute forecast errors				Median absolute forecast errors			
	Average	Comparison with RW			Average	Comparison with RW		
		DIFF	t-stat	%DIFF		DIFF	t-stat	%DIFF
Benchmark model								
RW	0.0764				0.0309			
Updated models								
OLS (Top 5)	0.0713	-0.0052	-6.69	-6.81%	0.0302	-0.0006	-2.02	-1.94%
LASSO (Top 5)	0.0714	-0.0051	-6.49	-6.68%	0.0303	-0.0006	-2.20	-1.94%
Ridge (Top 5)	0.0713	-0.0051	-6.44	-6.68%	0.0303	-0.0006	-1.98	-1.94%

Panel A reports the time series average of the mean absolute forecast errors of the five extant models and their counterparts augmented with TXT and change of TXT. Difference is calculated as the time series average of the mean absolute forecast error difference between the augmented model and its corresponding original model. The Newey–West t-statistic of Difference is adjusted using three lags and reported accordingly. %Difference is Difference divided by the time series average of the mean absolute forecast error of the original model. Panel B reports the time series average of the mean and median absolute forecast errors of the updated models and their comparisons with the benchmark model (i.e., the RW model). DIFF is the time series average of the difference calculated as the mean (median) absolute forecast error of each model minus that of the benchmark model. The Newey–West t-statistic of DIFF is adjusted using three lags and reported accordingly. %DIFF is DIFF divided by the time series average of the mean (median) absolute forecast error of the benchmark model. In both panels, the absolute forecast error is defined as the absolute difference between the actual one-year ahead earnings and the corresponding model-based earnings forecasts scaled by market equity at the end of three months after the end of the last fiscal year.