Machine Learning, Earnings Forecasting, and Implied Cost of Capital - US and International Evidence

Akash Chattopadhyay
University of Toronto

Bingxu Fang
Singapore Management University

Partha Mohanram*
University of Toronto

October 2022
Preliminary Draft
(Please do not cite or distribute without permission)

Abstract

This study offers the first large-scale evaluation of various earnings forecasting models and implied cost of capital estimates from these models across an international sample. We use machine learning (ML) methods to predict future earnings over multiple horizons and compute implied cost of capital (ICC) estimates from these forecasts. Nonlinear tree-based models outperform extant cross-sectional forecasting models and a naive random walk model over different horizons. ML models produce significantly lower (6-10 %) absolute forecast errors for smaller firms and firms with more volatile earnings, as well as for firms outside the US. ICC estimates based on forecasts from the ML models also exhibit greater alignment with realized returns than estimates from extant models. Our analyses suggest that the nonlinear interactions among predictors drive the outperformance of the ML models.

Keywords: Machine Learning; Forecasting; Valuation; Implied Cost of Capital

JEL: G18, G34, G38, G41, L51, M14, M52

*First draft: March 2022. Chattopadhyay (akash.chattopadhyay@rotman.utoronto.ca) is an Assistant Professor of Accounting at the University of Toronto. Fang (bxfang@smu.edu.sg) is an Assistant Professor of Accounting at the Singapore Management University. Mohanram (partha.mohanram@rotman.utoronto.ca) is the John Watson Chair in Value Investing and Professor of Accounting at the University of Toronto. We acknowledge helpful comments from Richard Crowley, Hongping Tan, and seminar participants at Santa Clara University, McGill University, KAIST, and IIM Bangalore.

1. Introduction

Forecasts of future earnings are a necessary input for valuing a risky asset. Researchers in accounting and finance have primarily relied on two sources of forecasts — sell-side analysts and forecasts from linear cross-sectional models (e.g., Hou, van Dijk, and Zhang, 2012; Li and Mohanram, 2014). However, analyst forecasts are unavailable for large swathes of firms and tend to be optimistically biased. The bias in analyst forecasts also renders them relatively less helpful in estimating the implied cost of capital (ICC), a widely used metric of cost of equity capital in accounting and finance research (e.g., Easton and Monahan, 2005; Mohanram and Gode, 2013). While more available and less biased, forecasts from linear models tend to be inaccurate (Gerakos and Gramacy, 2013).

Moreover, researchers have designed and evaluated existing linear models exclusively focusing on the United States. Empirical findings based on US data might not translate to an international setting. For example, Hou, Karolyi, and Kho (2011) and Fama and French (2017) document in the empirical asset pricing literature that the set of factors which explain the cross-section of returns in the US and outside the US are not necessarily the same. Variations in product and labor market competition and in accounting and regulatory standards across countries are likely to affect firms' underlying economics and the properties of reported profitability (Healy, Serafeim, Srinivasan, and Yu, 2014). The absence of any international evidence on the efficacy of profitability forecasting models is particularly relevant because, as Chattopadhyay, Lyle, and Wang (2022) document, a significant body of research uses the most common forecasting application, Implied Cost of Capital (ICCs), as a variable of interest in an international setting. Moreover, Chattopadhyay et al. (2022) also document that ICCs based on the extant models tend to perform inconsistently outside the US. Therefore, the evaluation of forecasting models outside the US is an important empirical question.

We aim to answer this question by exhaustively reviewing the performance of extant models and new candidate forecasting models within and outside the US. We further validate the choice of model by comparing their accuracy in subsets of firms where forecasting is likely to be more challenging, or model-based forecasts are likely to be more relevant. Finally, we also evaluate models based on the performance of ICCs generated using the outputs of the models. In doing so, we aim to offer researchers guidance on the choice of forecasting models in various settings.

We turn to the machine learning literature for new candidate forecasting models. Machine learning models are ideally suited for earnings forecasting because of their ability to handle high-dimensional data while being robust to overfitting. Moreover, machine learning allows us to relax the assumption of a simple linear relationship between firm characteristics and future profitability. We, therefore, evaluate the efficacy of two classes of ML models — penalized linear models (Lasso and Ridge), which can handle high-dimensional data and a decision-tree-based model (Extreme Gradient Boosting Regression), which relaxes the assumptions of parameterization and linearity. Our inputs include 56 predictors from the cross-section of accounting data and four predictors based on macroeconomic data. We examine the accuracy of earnings forecasts from these models for multiple horizons, ranging from 1-year to 5-years-ahead, and the predictive ability of ICC estimates based on these forecasts.

We begin our analysis with an initial benchmarking exercise on the cross-section of US firms. We show that the extreme gradient-boosting regression model (XGBoost) produces forecasts with the lowest mean absolute error (MAFE) over multiple horizons. The best performing extant model, the RI model developed by Li and Mohanram (2014), generates forecast errors that are about 7% higher at the 1-year horizon and about 1.5% higher at the 5-year horizon.

While the initial benchmarking analysis suggests relatively modest gains from using the ML models in the US, it offers valuable insights. An OLS-based linear model and the Lasso and Ridge models, using the exhaustive set of predictors used in the XGBoost model, produce less accurate forecasts than both XGBoost and the parsimonious RI model. This result suggests that the flexibility of the XGBoost model in accommodating non-linearities, rather than the broader set of characteristics used, is the source of its higher accuracy. We posit that XGBoost's ability to accommodate non-linearities flexibly is likely to be particularly relevant in firms with more complex underlying economics, for example, small firms with immature business models or highly volatile earnings.

Consistent with our intuition, the relative efficacy of the XGBoost model in the US is more pronounced in sub-samples of firms where the prediction problem is likely to be more complicated or essential. We partition firms by characteristics commonly shown to be correlated with a firm's "information environment" — size, presence of analyst forecasts, and earnings volatility. We find that for small firms, firms without analyst coverage, and firms with more volatile earnings, i.e., for firms with "harder-to-predict" earnings, the XGBoost model produces MAFE that are lower in an economically meaningful manner at both the shorter and longer horizons. For these sub-samples, the XGBoost model produces forecasts more accurate than the RI model by 7-8% at the 1-year horizon to about 2-4% at longer horizons.

We next turn to our question of the suitable forecasting model for an international sample. We focus on a global (non-US) sample of firms from 35 countries. The relative efficacy of the ML models outside the US is not obvious ex-ante. The flexible structure of the XGBoost model might better capture the variation in accounting standards and economic forces across countries compared to a simple linear model. On the other hand, the noise in data outside the US might favour parsimony over the extensive data requirements of the ML models.

We document that forecasting is more challenging in a global sample because all models' accuracy declines. However, the relative efficacy of the ML models is significantly greater in this global sample. For the overall global sample, XGBoost produces about 12% and 18% more accurate forecasts than the best-performing extant model, RI, at the one-year and the 5-year horizon, respectively. Consistent with the US evidence, the outperformance of XGBoost is higher in the harder-to-predict sub-samples. However, notably, in the global sample, the outperformance of XGBoost is also economically meaningful in the sub-sample of firms for which the prediction problem is likely to be less severe. For the sample of larger firms, firms with analyst coverage, and firms with less volatile earnings, XGBoost produces forecasts that are more than 10% accurate even at the 5-year horizon.

We next show that the relative efficacy of the ML models, particularly outside the US, translates to an alternative evaluatory framework, the ICC framework. As a first step towards examining the ICCs computed using the forecasting models, we compare the relative unconditional bias of the outputs of the various models. We find that, on average, model bias is significantly higher in the global sample relative to the US.[1]. XGBoost produces the least biased forecast in the global sample across all horizons. Consistent with this evidence, we find that ICCs produced using the XGBoost model perform the best in terms of their correlation with realized returns in the global sample.

We make several contributions to the burgeoning literature on the usefulness of machine learning in accounting and finance. To our knowledge, we are the first to examine the efficacy of ML forecasting models outside the US. We document that a regression-based tree model outperforms existing models in a global sample in an economically meaningful manner. Moreover, we validate our findings in contexts where earnings are likely harder to predict and document further evidence of the relative efficacy of the XGBoost model. Finally, we also document evidence suggesting that the incremental predictive power of the XGBoost model stems from its ability to accommodate non-linearities flexibly rather than feature selection. Our findings complement and build on contemporary work on earnings forecasting using machine learning (e.g., Chen, Cho, Dou, and Lev, 2022; Cao and You, 2020; Hansen and Thimsen, 2020), which show the ability of decision-tree-based models to produce more accurate forecasts but do not focus on specific contexts. Our research also extends this strand of literature by examining the performance of ML models at longer horizons, while existing research has an exclusive focus on one-year ahead forecasts.

To our knowledge, we are also the first to examine the predictive ability of ICC estimates generated using forecasts from ML models. By doing so, we contribute to the emerging literature in asset pricing examining the role of ML in estimating expected returns (e.g., De Silva and Thesmar, 2021;

[1]The magnitude of forecasting bias for the XGBoost model is about 40% lower than the best-performing extant model at the one-year horizon. The superior performance of XGBoost becomes more prominent for longer horizon forecasts, and it is the only model that produces unbiased forecasts for three- to five-year ahead forecasts. In terms of the magnitude, XGBoost is about 80% less biased than the best extant model, RI, at the five-year horizon.

van Binsbergen, Han, and Lopez-Lira, 2020; Gu, Kelly, and Xiu, 2020), as well as methodological work focusing on estimating profitability and expected returns outside the US (Chattopadhyay et al., 2022).

On a more technical note, we also highlight the efficacy of the novel XGBoost model. Prior studies (e.g., Chen et al., 2022; Cao and You, 2020; Gu et al., 2020) use random forests, standard gradient-boosted regression trees, or neural networks to forecast earnings or stock returns. XGBoost requires significantly lower computing resources and is multiple times faster than these models while producing similar or even better forecasting performance in our context.[2] We believe these features of XGBoost significantly enhance the accessibility of ML models for researchers. Overall, based on our exhaustive evidence, we recommend that future research that requires earnings forecasts, whether used as measures of expected earnings or to compute ICCs, use the XGBoost model both for the US and international settings.

We have organized the rest of the paper as follows. Section 2 positions our paper in the related literature on forecasting, the use of machine learning models, implied cost of capital, and differences between the US and international context. Section 3 outlines our research methodology — model description, estimation, and validation. Section 4 presents the results of our estimation and compares the performance of the different models in the US and international settings. Section 5 concludes.

2. Relation to Literature

Forecasting profitability constitutes a rich literature in accounting. This literature has often intersected productively with another body of methodological work in accounting — estimating expected returns using the so-called Implied Cost of Capital (ICC). However, researchers in this area are only beginning to explore innovations in machine learning for forecasting. Further, while

[2] In untabulated tests, we verify that XGBoost is at least as accurate, if not more, compared to random forests or standard gradient-boosting regression trees. However, XGBoost reduces the computing time by nearly 90% in our research setting. See section 3.1..2 for a more detailed discussion of the XGBoost model.

a large body of work uses outputs of forecasting models and ICCs in research in an international context, methodological work remains scarce outside the US context.

2.1. Cross-sectional Models in Forecasting

Forecasts of future earnings are essential inputs for valuing a risky asset. In research on valuation, forecasts have been used in various contexts. For example, Frankel and Lee (1998) use forecasts to estimate intrinsic value (V) to identify undervalued and overvalued stocks using a value-to-price ratio. The entire literature on implied cost of capital (e.g., Gebhardt, Lee, and Swaminathan, 2001; Gode and Mohanram, 2003; Easton, 2004; Botosan and Plumlee, 2002) also relies on forecasts. For a long time, the only source of forecasts was analyst forecasts from sources such as I/B/E/S or First Call.

Using analyst forecasts presents researchers with two significant problems. First, analyst forecasts are generally optimistic and not very accurate. Easton and Monahan (2005), Easton and Sommers (2007), and Mohanram and Gode (2013) are among the papers that show that inaccurate and optimistic forecasts are the leading reasons why ICC models do not perform well in predicting future returns. Second, analyst forecasts are unavailable for all firms, as analysts tend to follow larger firms in better information environments. This skewed lack of coverage means that researchers cannot answer important questions regarding information quality and disclosure in the subset of firms where the answers to such questions would be particularly insightful. Time series models also do not offer an appropriate solution in these cases. The lengthy firm-specific time series of data required to estimate such models renders them ineffective in the subset of younger firms with shorter histories.

Cross-sectional forecasting, a technique that has emerged in the last decade, attempts to address these shortcomings. First, it uses the cross-section of data without imposing any firm-specific data limitation - i.e. a firm need not have existed for the entire estimation period. Consequently, we can obtain a forecast for the entire cross-section of firms at any point in time. Secondly, the models are not subject to the behavioural biases that plague analyst forecasts.

The first notable paper in this area, Hou et al. (2012), builds on models in Fama and French (2000, 2006) and regresses future earnings on total assets, dividends, earnings and accruals. However, Gerakos and Gramacy (2013) shows that the HVZ model is less accurate than a naïve random walk model where future earnings are equal to past earnings. The economic magnitude of the forecast errors of the HVZ model is high — on average, implying an error equal to the earnings estimate. More importantly, the HVZ model generates larger errors for firms without analyst coverage, where the need for a forecasting model is crucial.

Li and Mohanram (2014), present two alternative models - the earnings persistence model (EP) and the residual income model (RI), based on the evidence in (e.g., Dechow, 1994) that accrual-based measures like earnings have more persistence and predictability than cash flow based measures, such as the inputs that HVZ use. The EP and RI models outperform the HVZ model as well as the random walk model. Further, they perform even better in the subset of small firms and firms without analyst following, where the utility of these models is the most salient. However, in absolute terms, the average level of forecast error reported in Li and Mohanram (2014) is still relatively high. In this paper, we will attempt to see if machine learning-based approaches can generate forecasts with significantly greater accuracy.

2.2. Application of Cross-sectional Forecasts: Implied Cost of Capital

The cost of equity or expected returns plays a central role in valuation, portfolio selection, and capital budgeting. Estimates of expected returns derived from popular asset pricing models such as the CAPM or factor models perform poorly in their association with realized returns (Chattopadhyay et al., 2022). Realized returns are noisy and, thus, unsuitable ex-post measures of expected returns (Elton, 1999). An important innovation in this area has been the implied cost of capital (ICC) developed in the accounting literature. An ICC is the discount rate that equates the current stock price to the present value of expected future dividends. Researchers have used variations of the present value model, the residual income model, or the abnormal earnings growth model to estimate ICCs.[3] Mohanram and Gode (2013) validate the current practice in the literature of using a composite ICC, based on the popular measurement approaches, by showing that such a metric has lower measurement error than outputs of the individual models.

ICCs are now a widely used measure of the cost of equity in accounting and finance research. Researchers have used ICCs to study how a firm's cost of capital responds to changes in its information environment through voluntary disclosure (Dhaliwal, Li, Tsang, and Yang, 2011), accounting standards (Daske, Hail, Leuz, and Verdi, 2008; Li, 2010), securities regulation (Hail and Leuz, 2006), tax laws (Dhaliwal, Krull, Li, and Moser, 2005), etc.. A significant proportion of this body of work is based in international contexts.

The ICC paradigm is the most direct forecasting application because short-term and long-term earnings forecasts are crucial in estimating ICCs. Early research on ICCs relied on analyst forecasts as inputs, despite the problems of limited coverage and bias in the forecasts. Easton and Monahan (2005) and Easton and Sommers (2007)) demonstrate issues with analyst-forecast-based ICCs. Optimistic analyst forecasts give rise to ICCs that are biased upwards and have poor predictability for future returns. While Mohanram and Gode (2013) show that if researchers adjust analyst forecasts for predictable error and bias, the ICCs generated from these forecasts perform better, currently, most researchers use forecasts from the cross-sectional models (HVZ or EP and RI) to estimate ICCs. Consequently, ICCs are a natural tool to validate an earnings forecasting model — for a model to supersede predecessors, ICCs based on that model should perform better. In this paper, we analyze the performance of the ICCs generated from ML-based forecasting models relative to estimates from extant models.

[3] Gebhardt et al. (2001) and Claus and Thomas (2001) use variants of the residual income model to solve for the discount rate that equates price to the sum of book value and the present value of future abnormal earnings. Gode and Mohanram (2003) and Easton (2004) develop proxies based on the abnormal earnings growth model of Ohlson and Juettner-Nauroth (2005).

2.3. Emergence of Machine Learning in Forecasting

Machine Learning (ML) as a tool to solve the two canonical prediction problems in accounting and finance research — predicting profitability and returns — is receiving considerable attention because of the ability of ML models to handle correlated, high-dimensional data and unspecified non-linearities within the data.

This growing literature has primarily focused on predicting asset returns (e.g., Freyberger, Neuhierl, and Weber, 2020; Gu et al., 2020). A key motivation behind this work has been to solve the dimensionality problem created by the proliferation of return-predictive signals. This literature suggests that machine learning models successfully identify the return predictors with independent information and generate significant improvement over existing models in terms of the quality of predictions. Moreover, this literature suggests that the advantage of machine learning models is realized in methods that allow for nonlinear predictor interactions.

The literature focusing on the utility of ML in forecasting the other key input of equity valuation — profitability — is relatively nascent (e.g., Chen et al., 2022; Cao and You, 2020; van Binsbergen et al., 2020; De Silva and Thesmar, 2021). While the overarching goal of this literature has been to evaluate the efficacy of ML models in forecasting future firm profitability, the key focus of the individual papers differs. van Binsbergen et al. (2020) and De Silva and Thesmar (2021) focus on using ML to create an optimal earnings benchmark to precisely identify expectation errors by analysts. In work related to ours, Cao and You (2020) focus on the entire cross-section of US firms to identify the optimal forecast of future profitability. Consistent with the findings from the literature using ML to predict returns, ML models incorporating non-linearities yield the best results in forecasting profitability. More specifically, Cao and You (2020) find that the nonlinear ML models (Random Forest, Gradient Boosting Regressions, and Artificial Neural Networks) yield predictions that outperform those from a naive benchmark from the RW model. In contrast, those from the extant linear models (HVZ, RI, and EP) do not. Chen et al. (2022) also use random forests and gradient boosting-based models, but they include various financial items collected from XBRL filings as input variables. They show that ML models perform better when predicting the direction of future earnings changes than logistic regressions and random guesses. To our knowledge, we are the first to evaluate ML models' efficacy in accommodating various information environments within and outside the US.

2.4. US vs International Evidence on Forecasting and ICCs

The body of methodological work described above almost exclusively focuses on the US context. There is little evidence on the efficacy of the various linear cross-sectional models in forecasting profitability and serving as inputs to ICC models outside the US. The performance of the extant models outside the US is not obvious ex-ante. First, at a general level, the US and other global markets vary on economic, political, legal, and institutional dimensions (Ball, 2016) which should motivate examination of economic models outside the US. Secondly, empirical evidence suggests that institutional and economic differences across countries also affect the earnings process, the critical outcome of interest in this literature (Healy et al., 2014). Consequently, it is plausible that findings based on linear parameterized earnings forecasting models estimated on US data might not extend internationally.

The question of identifying earnings forecasting models that perform robustly internationally is an important one because a burgeoning body of work in accounting and finance uses ICCs to study the influence of various policies on firms' cost of capital in an international context (Chattopadhyay et al., 2022). As Chattopadhyay et al. (2022) document, the performance of ICCs using forecasts from the extant cross-sectional models is relatively inconsistent outside the US, while the availability of analyst forecasts is sparse. Moreover, as Fang, Hope, Huang, and Moldovan (2020) document, the availability of analyst forecasts in Europe is declining following the enactment of MiFID II.

We fill this gap and build on the literature described above by examining the efficacy of the ML models outside the US context, as well as the validity of the ICC estimates generated using the outputs of these models. The latter analysis differs from the literature on using ML to predict returns in one key way. Instead of fitting a wide set of accounting characteristics to a noisy outcome variable (returns), we use ML to generate optimal forecasts of a less noisy accounting variable (earnings). We map these earnings to returns by using the theoretically motivated present-value approach to estimate ICCs.

3. Research Methodology

In this section, we briefly discuss the conceptual underpinnings of the various forecasting models and ICC measures we consider and detail their estimation processes. We also discuss our empirical framework for evaluating the forecasting models and the ICCs.

3.1. Forecasting Profitability

We evaluate three candidate machine learning models — two from the class of linear penalized models (Lasso and Ridge) and a tree-based model incorporating non-linearities (Extreme Gradient Boosting) — against the extant models described in Section *2.1.*.

3.1..1 Traditional Models

The three traditional cross-sectional forecasting models we consider are the HVZ model developed by Hou et al. (2012) and the RI and EP models are based on the work by Li and Mohanram (2014). We use a naive random-walk model (RW) prediction as a benchmark. All three cross-sectional models produce earnings forecasts by estimating the following general model:

$$\mathbb{E}[E_{i,t+\tau}] \quad = \beta_0 + \beta_1 X_{i,t} + \epsilon_{i,t} \tag{1}$$

where $\mathbb{E}[E_{i,t+\tau}]$ represents expectation of earnings τ periods away and X_i represents firm-level characteristics measured at time t. The Appendix describes each model in terms of the firm characteristics involved. One difference between the HVZ model and the RI and EP models is that HVZ estimates earnings while RI and EP estimate earnings per share. We follow the extant literature in estimating the HVZ, EP,and RI models (Hou et al., 2012; Li and Mohanram, 2014). Specifically, for each year t in our sample, we use the previous 10 years' of observations ($t-1$ to $t-10$) as the training sample to estimate the parameters of each model, and then we use the parameters and the financial information in year t to generate the forecasts for year $t+1$ to year $t+3$.[4] When estimating the models for the international setting, we use a pooled sample with observations across all countries in our sample to increase the size of the training sample. [5]

3.1..2 ML Models: Background and Estimation

We briefly discuss the machine learning models we evaluate and detail our estimation process. Readers should refer to Hastie, Tibshirani, and Friedman (2009) for a significantly more technical description of these models.

The first class of models we evaluate is the so-called penalized class of models. The advantage of these models over linear regression is their lesser susceptibility to overfitting, even as the number of parameters to be estimated increases. OLS estimates parameters to minimize a standard least squares objective function :

$$\beta^{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{2}$$

with p predictors and N observations. With an increase in the number of parameters to be

[4]To mitigate look-ahead bias, we follow Li and Mohanram (2014) and assume that firms with fiscal year ending in April to June do not have their financial information available by end of June. We only include firms with fiscal year ending in April of year t-1 to March of year t when estimating the models for year t.

[5]There is a trade-off in the international setting between the size and the relevance of the training sample. Using a pooled sample across countries increases the size of the training sample, while countries with heterogeneous business environments are designed to have the same model parameters. On the other hand, estimating the models by country will allow each country to have different parameters, but it decreases the training sample sizes. In untabulated analyses, we estimate the models by country in the international setting and our inferences regarding the superior performance of machine learning models remain unchanged. However, the performance of each model generally becomes worse because of the smaller training samples. Similarly, estimating the models in subsamples based on firm size, analyst coverage, volatility, and industry also leads to worse performance due to the smaller training sample size.

estimated, OLS is prone to overfitting the model in-sample, leading to poor predictive performance out-of-sample. Moreover, coefficient estimates in linear regressions can be poorly determined and harder to interpret with many correlated predictors. Penalized models, also referred to as *Shrinkage* methods, are constrained to place the greatest weight on the subset of predictors with the highest predictive content. Penalized models thus allow for bias in the parameter estimates to minimize expected prediction error. By shrinking coefficients, penalized models also avoid the problem of overfitting for high-dimensional models. We examine two popular candidates from this class of estimators, Ridge Regression and Lasso. Ridge Regression minimizes a penalized sum of squares:

$$\beta^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3}$$

where λ is the shrinkage parameter which scales coefficient values lower. Lasso minimizes the following penalized sum of squares:

$$\beta^{Lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4}$$

where λ is the corresponding shrinkage parameter for Lasso. We can describe the key difference between Ridge and Lasso in a simplified manner by saying that Ridge shrinks all coefficients proportionally, and Lasso shifts coefficients by a constant, truncating at zero. Thus, in Ridge, all predictors will technically get a non-zero coefficient, while Lasso will altogether discard some predictors. Moreover, while the Ridge estimate has a closed-form solution, Lasso requires numerical estimation.

We estimate all machine learning models with 60 predictors — 56 are 28 financial statement line items obtained from Compustat, and their changes relative to the previous year. The other four predictors are macroeconomic variables: GDP growth rate, unemployment rate, industrial production growth rate, and consumption growth rate .[6] Similar to the estimation process for the traditional models, we use the previous ten years of data when estimating the ML models. For the international sample, we pool data across all countries to increase the training sample size. Following extant literature, (Hastie et al., 2009), we use five-fold cross-validation to identify the optimal hyperparameter values for each year. [7]

Finally, we consider a non-parametric decision-tree-based model — Extreme Gradient Boosting (XGBoost). XGBoost is a recent implementation of Gradient Boosting Machines (GBM) by Chen and Guestrin (2016). Like Random Forest (RF) and GBM, XGBoost is an ensemble learning model consisting of a set of decision trees. The fundamental model in each ensemble is a regression tree, which partitions the data into a set of regions where the predicted value in each region is a constant that minimizes a loss function. So, for a dataset partitioned into M regions R_1, R_2, ..., R_m, a decision tree can be represented as:

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{1}(x \in R_m) \tag{5}$$

The predicted value in each region is simply the average outcome variable for that region. A decision tree is modelled using a greedy algorithm which starts with the entire data and splits it

[6]Detailed definitions of the predictors are available in the Appendix. We use 58 predictors for the international sample because XAD is unavailable in Compustat Global. Similar to Cao and You (2020), we set the following variables to zero if they are missing: accounts payable, advertising expenses(only for the US setting), current assets, current liabilities, dividends per share, income taxes payable, intangible assets, interest and related expenses, investments and advances-other, SGA expenses, short-term investments, research and development expense, and special items.

[7]The K-fold cross-validation process is a popular way to reduce potential over-fitting of the estimated models. The process starts with randomly dividing the training sample into K groups without replacement and then using one group as testing data and the other K-1 groups as training data. This process repeats K times, so each group will be used as test data once. We use each group as test data and calculate the average performance of the model for a particular parameter value. After iterating over candidate values for the parameter (ranging from 0.0001 to 0.1 in steps of a thousandth of the interval), we use the average model performance to determine the optimal shrinkage parameter for the Lasso and Ridge models.

using a predictor j and a split point s into regions R_1 and R_2 that solves:

$$\min_{j,s}\left[\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{6}$$

where c_1 and c_2 are the predictions in each region. For any j and s, the average of the outcome variable in each region solves the minimization problem. Thus, a splitting point s can be found for each predictor. As the first split, the algorithm chooses the predictor which produces the lowest value of the sum of squared errors as described in Equation 6. The algorithm repeats the process for each subsequent resulting region. Since a few iterations of this process are likely to result in a very complex tree, one of the critical parameters for designing a regression tree is the depth of the tree. The tree depth is usually chosen by pruning a large tree. The pruned tree is selected by minimizing a loss function which includes a penalty for the number of terminal nodes in the smaller tree. The critical advantage of regression trees is their conceptual simplicity and flexibility to accommodate non-linearities and interactions in the data. However, the high variance of an individual tree undermines its flexibility. A small change in the data can lead to a tree of a very different structure. Therefore, an ensemble approach is preferred to produce more robust models.

XGBoost is an implementation approach that offers two distinct advantages over traditional GBM models used in prior finance or accounting studies (e.g., Chen et al. 2022). Like traditional GBM models, XGBoost starts from a simple tree that produces errors that are not significantly lower than the sample average as the predictor. Then the model fits a regression tree to the residuals from the first tree in the next iteration. It keeps repeating this process for a set number of iterations, with the final output being an addition of the individual regression trees. However, XGBoost introduces a new regularized learning objective, which penalizes the complexity of the model and better deals with data overfitting problems. Secondly, XGBoost better accommodates parallel and out-of-core computing, significantly reducing the time required to train the model.[8]

[8]Our tests suggest that XGBoost runs at least ten times as fast as traditional GBM models in our research setting. XGBoost takes about one to two minutes for each round of training, while RF and GBM take about 20 to

Our estimation process of the XGBoost model is similar to that for the Lasso and Ridge models. We estimate the model each year using the same set of 60 predictors and the previous ten years' data. We also use the five-fold cross-validation process to search for the optimal choice of hyperparameters for the model for each year.[9].

3.2. Proxying for the Information Environment

One of the motivations of this paper is to improve forecasting accuracy for firms lacking quality forecasts or when the forecasting exercise is more complex, i.e., for firms with arguably weaker information environments. We consider three proxies for information environment. Prior research has long used firm size as a proxy for the quality of the information environment (e.g., Brown, Richardson, and Schwager, 1987; Wiedman, 1996) because small firms have lower disclosure quality, auditing quality, media coverage, institutional investment, and analyst following. Our first proxy is hence firm size, measured by market capitalization. We classify firms into *Small Firms* and *Large Firms* based on whether a firm's market capitalization is above or below the cross-sectional median.

Our second proxy is a binary variable indicating whether the firm has any analyst following or not. Earlier researchers would ignore such firms in their analysis. Cross-sectional forecasting allows researchers to generate estimates of future earnings for these firms. Given the lack of other alternatives, forecast accuracy is essential for this subgroup.

Our third proxy is the volatility of earnings. Volatile earnings make forecasting more difficult. Prior research (e.g., Bhushan, 1989; O'Brien and Bhushan, 1990) shows that analysts often eschew firms with volatile earnings. In addition, the volatility also makes cross-sectional forecasting potentially less effective, as extrapolating from the past to the present is more complicated. We measure

30 minutes. For our forecasting exercise — one-year to five-year ahead forecasts for around 50 years of data for the US setting (and 25 years for the international setting), XGBoost requires nearly one week less computing time. In untabulated analyses, we find that this lower computation time does not involve any trade-offs in performance — XGBoost produces lower errors than RF or GBM.

[9]We set the number of trees in the forest to be 500 and the learning rate to be 0.01. The hyperparameter to fine-tune is the tree's maximum depth, which ranges from 1 to 7 with a step of 2. We use the Huber loss function to make the model more robust to potential outliers.

earnings volatility as the standard deviation of the firm's quarterly return on assets (ROA) in the previous 5 years.

3.3. Estimating Implied Cost of Capital (ICC)

We evaluate the various earnings forecasting models by also examining the predictive content of ICCs calculated using their ouputs. We compute four ICC variants commonly used in the literature. Two are based on the residual income valuation model from Ohlson (1995) and Feltham and Ohlson (1995) - the GLS model from Gebhardt et al. (2001) and the CT model from Claus and Thomas (2001). Two are based on the abnormal earnings growth model from Ohlson and Juettner-Nauroth (2005) - the OJ model from Gode and Mohanram (2003) and a simplifed PEG model from Easton (2004). We estimate annual ICCs at the end of June of each year and winsorize each ICC estimate at the 1% and 99% levels for each cross-section. To further mitigate the effect of outliers, following Mohanram and Gode (2013), we calculate a composite ICC as the average of the ICCs from the four approaches mentioned above. If one or more of the four individual ICCs are unavailable, we follow Hou et al. (2012) and compute the composite ICC as the average of those available.

3.4. Validating Forecasts and ICCs

Following prior research (Li and Mohanram, 2014), we evaluate the earnings forecasting models by examining the mean absolute forecast error (MAFE) — the absolute value of the difference between the estimated earnings per share (estimated Net Income in dollars for HVZ) and the actual realized earnings per share (realized Net Income for HVZ), scaled by price per share (market capitalization for HVZ) — of the output of each model. More specifically, we compute the cross-sectional average of the MAFE from each model and then examine the time series of these averages. We perform this exercise for forecasts up to five years ahead for both the US and our international sample. We also compare the relative efficacy of the ML models over the traditional models by examining the difference in the cross-sectional averages of various models.

We also follow prior research (e.g., Chattopadhyay et al., 2022; Li and Mohanram, 2014; Lewellen, 2015) in evaluating the ICCs computed using the outputs of the various forecasting models. We assess the ICCs by examining their association with future realized returns. We use both a regression-based approach and a non-parametric approach using portfolio sorts to validate that the cross-sectional differences in ICCs are directionally consistent with the differences in realized returns. For the regression-based method, we estimate Fama-Macbeth (FM) regressions of one-year-ahead realized returns on the various composite ICC estimates:

$$R_{i,t+1} \quad = \beta_0 + \beta_1 ICC_{i,t+1} + \epsilon_{i,t+1} \tag{7}$$

A positive and significant β_1 would validate the predictive ability of an ICC estimate. Ideally, we would expect β_1 to be 1 for an accurate measure of expected returns. Consequently, we also evaluate ICCs on whether we can statistically distinguish β_1 from 1. As Chattopadhyay et al. (2022) discuss, this is a minimally sufficient criterion for evaluating a proxy of expected returns. We also examine the equal-weight predicted and realized returns of decile portfolios based on each ICC measure.

3.5. Data Sources

We obtain annual financial information for US firms from Compustat and stock returns from CRSP, and we collect financial and stock price information for the international sample from Compustat Global. We obtain analyst coverage information from I/B/E/S. Our macroeconomic data comes from the Federal Reserve Bank of St. Louis for the US sample and the OECD for international firms. Our US sample covers the period from 1969 to 2020. We require each firm to have common shares listed on the NYSE, AMEX, and NASDAQ, and the stock price at the end of June to be higher than 1 USD. We exclude financial and utility firms from our sample. After further dropping observations with missing values for our predictors, we have 106,459 firm-year observations in the final sample. We impose similar data requirements for our international sample and convert all variables to US dollars before applying the filters. We further require each country in the sample to have at least 100 observations, which leaves us with a final sample of 132,160 firm-year observations from 35 countries.

4. Empirical Findings

4.1. Model Performance in the US: Absolute Forecast Error

4.1..1 Overall Sample

We begin our analyses by analyzing the forecast errors of our candidate models in the sample of US firms. In addition to the models described in **??**, we consider two benchmark models — RW and Linear. RW is a naive random walk model, while Linear is a simple linear model with the same 60 predictors used by the ML models. We present the results of this analysis in Table 1.

The overall results in Table 1 document that XGBoost produces the lowest MAFE across short and long horizons. At the one-year-ahead level, reported in column 1, XGBoost's MAFE of 0.056 is 7% lower than the MAFE of 0.060 of the best-performing cross-sectional models, EP and RI. As the last three rows of Table 1 document, these differences are statistically significant at the 1% level. The MAFEs of the penalized ML models, Lasso and Ridge, at 0.059 each, are comparable to that of EP and RI, suggesting that simply including many predictors without incorporating non-linearities does not improve model performance. At the one-year horizon, HVZ produces the highest MAFE at 0.081.

The results in columns 2–5 suggest that the general patterns observed in the one-year-ahead results extend to longer horizons. Although the MAFE for all models, including XGBoost, goes up as the horizon lengthens, XGBoost continues to produce the lowest MAFE across all models. However, the relative outperformance of XGBoost diminishes as the horizon lengthens, culminating in a statistically significant 1.5% improvement over RI at the 5-year horizon. Interestingly, RI outperforms Lasso and Ridge at longer horizons, suggesting that adding predictors without incorporating non-linearities is detrimental. Consistent with this inference, the OLS-based model using the complete set of predictors, Linear, produces MAFEs that are about 10% higher than RI.

4.1..2 Partitions based on the Information Environment

In our following analyses, we examine whether the non-linear and flexible nature of the XGBoost model lends it more advantages relative to the parsimonious and linear models in samples of firms where forecasting is likely to be more challenging. We consider the following three partitions, discussed in section xxx - firm size, analyst following and earnings volatility.

Panel A of Table xxx suggests that XGBoosts outperformance is more persistently and economically significant in smaller firms compared to the overall sample. We split our sample with forecasts each year into two groups (Small Firms and Large Firms) based on the median market capitalization. Columns (1) – (5) present the time-series averages of the average cross-sectional MAFE of the individual models over a one-year to a five-year horizon. Columns (6) – (10) present the corresponding results for firms with above-median market capitalization. Consistent with the results in Table 1, XGBoost and RI are the two standout models for the sample of smaller firms. Between them, XGBoost produces errors lower than RI's by about 7.2% at the one-year horizon. The magnitude of the outperformance dissipates marginally at longer horizons, ranging from 6.3

The results in columns (6) – (10) of Panel A suggest that, on average, all models are significantly more accurate for larger firms (MAFEs are more than 50% lower for larger firms) and that there is little to choose between XGBoost and RI for this sample. While XGBoost is generally the best-performing model and outperforms EP and RI by about 5% in year 1, the outperformance is not statistically significant for three of the five years in our forecasting horizon. Overall, the results in Panel A are consistent with our hypothesis that predicting earnings is inherently more difficult for smaller firms and that non-linearities and flexibility in the prediction model are essential for such firms.

Panel B examines the forecast accuracy of the models in samples with and without analyst coverage and documents that XGBoost produces economically and statistically significant outperformance in both samples, including the crucial subsample of firms lacking analyst forecasts. Columns (1) — (5) present results for this sub-sample. Like previous analyses, XGBoost produces the lowest MAFE at all horizons, with RI being the next best-performing model at most horizons. Between the two, XGBoost produces MAFE that are about 7% lower at the 1-year horizon and 4% lower at a longer horizon. Columns (6) — (10) report results for firms with analyst coverage. As in Panel A, the overall magnitude of errors for all models is significantly lower in the subsample we earmark as having the better information environment. Model errors are more than 30% lower for all models for the subsample of firms with analyst coverage. XGBoost continues to be the best-performing model, with forecast errors about 8% to 2% lower than the RI model.

In our final subsample analysis, we document that the flexible nature of the XGBoost model continues to outperform in firms where the forecasting exercise is likely to be complicated by the volatility of underlying performance, i.e., firms with high earnings volatility. We compute earnings volatility as the standard deviation of the firm's quarterly ROA for the previous five years. We use the median to partition each cross-section into high and low-volatility firms. The results of this subsample analysis are comparable to that in Panel A. Columns (1) — (5) document about an 8% to 5% lower MAFE for the XGBoost model, relative to RI, at shorter and longer horizons, respectively. Consistent with the analysis of large firms, overall forecast errors diminish significantly (by about 50% for all models) for firms with lower earnings volatility. This overall improvement also diminishes the differences between the performance of the models.

Overall, our analysis of US firms suggests that XGBoost yields economically meaningful improvement in forecast accuracy over extant models, particularly in samples of firms with "harder-to-predict" earnings where forecast errors tend to be larger. Moreover, the choice of characteristics matters less than the choice of model in improving forecast accuracy. Finally, the relative differences between the models diminish in firms with a likely better information environment.

4.2. Model Performance Internationally: Absolute Forecast Error

We next focus on our question of identifying forecasting models for international samples. To our knowledge, we are the first to evaluate the efficacy of forecasting models in an international setting. Whether our findings related to the relative performance of the forecasting models in the US will extend to our global sample is an empirical question. A significant body of research suggests that the information environment weakens as we move outside the US. For instance, a vast literature on cross-listing suggests that the information environment of international firms improves after they cross-list in the US (e.g, Lang, Lins, and Miller, 2003; Bailey, Karolyi, and Salva, 2006). Consequently, our finding that ML models are particularly efficacious for firms with weaker information environments in the US should point to their utility for international firms. However, it is also plausible that the noise in international data would mean that the parsimonious cross-sectional models, which are less susceptible to overfitting, prove to be more helpful outside the US. We examine this question by pooling observations across 35 countries resulting in about 132,000 observations.

4.2..1 Overall Sample

Table XXX documents significant gains from using XGB and a relative decline in the performance of the traditional forecasting models for the international sample. The first insight from the overall results is that model errors increase across the board, plausibly due to the noisier international data. Secondly, consistent with the idea that the performance of models developed and tested with a focus on the US does not necessarily translate outside the US, we find that all the traditional forecasting models underperform relative to a naive random-walk model (RW). For example, the best-performing traditional model, RI, produces MAFE that are about 3% higher at the 1-year horizon and 10% higher at the 5-year horizon relative to RW. Finally, consistent with our US results, XGB produces the lowest MAFE at all horizons, with the improvement relative to RW ranging from about 9-13%. The relative underperformance of the linear models with the complete set of characteristics (Linear, Lasso, and Ridge) again suggests that increasing the number of predictors without allowing interactions leads to noisier estimates.

4.2..2 Partitions based on the Information Environment

We next focus on subsamples of firms where forecasting is likely to be more challenging or relevant to provide further insights into the relative efficacy of the XGBoost model in the international context. We consider identical partitioning variables as in our analysis for the US sample — size, analyst following, and earnings volatility.

Table XXX documents that XGB continues to produce the lowest MAFE for both shorter and longer horizons across sub-samples. In contrast, the linear models (both the traditional and ML models) continue to underperform relative to the naive RW model. Panel A presents the analysis of model forecast errors for samples partitioned by size (market capitalization). Columns (1) – (5) present the results for small firms. The forecast errors for all models are significantly greater for smaller firms, with all the linear models, including the ML-based ones, producing MAFEs that are worse than the RW model. The only exception here is that at the 1-year horizon, the RI model produces a MAFE of 0.134, which is only marginally better than that of the RW model (0.135). However, XGBoost continues to outperform the RW model by about 12% at the 1-year and 15% at the 5-year horizon. Columns (6) – (10) of Panel A present the results for larger firms. The overall pattern of results stays the same, with XGBoost outperforming the traditional models at all horizons. However, the relative difference between forecast errors for XGBoost and RW is not statistically significant for larger firms.

Panel B presents the results of forecast analyses for samples partitioned by the absence or presence of analyst coverage. Columns (1) — (5) focus on the subgroup of firms without analyst following, where the performance of forecasting models is critical. Consistent with the international results thus far, none of the linear forecasting models are able to improve on the naive RW benchmark at any horizon. XGBoost outperforms all models across all horizons. XGBoost produces

MAFE lower than RW by about 8% at the 1-year horizon and 13% at the 5-year horizon. Columns (6) — (10) present the results of the analyses of the sample of firms with analyst coverage. Partly reflecting the likely variation in forecasting complexity of this broad sample, XGBoost continues to outperform the RW model, unlike the sub-sample of larger firms. The magnitude of outperformance also remains significant, with MAFEs of XGBoost being lower than RW's by about 9-10% at all horizons.

Finally, Panel C presents the results of forecast analyses of samples partitioned by lagged earnings volatility which largely mirrors the results in Panel B. Columns (1) — (5) present results for the subsample of firms with higher earnings volatility. For this subsample, only XGBoost delivers earnings forecasts with significantly lower errors than the RW benchmark. All other cross-sectional models fare worse across all five horizons. The improvement in forecast errors of XGBoost relative to RW range from about 10% at the 1-year horizon to For one-year-ahead forecasts, XGBoost generates MAFE of 0.093, which is around 10% better than the MAFE of 0.103 for the RW model. We see similar strong reduction in MAFE for XGBoost over the two-year to five-year forecast horizons. We obtain similar results for the sample of firms with lower earnings volatility, reported in Columns (6) – (10). XGB is the only model to outperform relative to the RW benchmark at any horizon. XGB's outperformance is lower in this sample of firms with plausibly easier-to-forecast earnings, ranging from 4.5% at shorter horizons to about 6% at longer horizons.

Overall, our partition analysis in the international setting again validates the usefulness of the XGBoost model in contexts where the forecasting problem is likely to be more complex (small firms or volatile firms) or forecasts are critical (firms without analyst coverage).

4.3. Model Bias

Our tests thus far have focused on forecasting accuracy measured by unsigned forecast error (MAFE). We now turn our attention to forecast bias. Bias and forecast accuracy are fundamentally different concepts - bias is a measure of signed error, while forecast accuracy is a measure of unsigned error. It is possible for a model with higher MAFE to have lower bias, if the errors "cancel out". Conversely, a model can have a high degree of accuracy and yet be biased. Why might researchers care about bias? Bias may not be that important if the focus is on the firm-level. However, if the focus is on the aggregate, e.g. estimating the aggregate market premium as in Claus and Thomas (2001) or estimating aggregate implied cost of capital as in Li, Ng, and Swaminathan (2013), then it is important to have unbiased forecasts. Easton and Sommers (2007)) show that ICC estimates derived from analyst forecasts are systematically biased upwards because the forecasts used to generate them are optimistically biased.

In our next set of tests, we examine the bias of the forecasts generated by the cross-sectional as well as the ML models. We define bias as the difference between the predicted earnings forecasts minus actual earnings , scaled by either price per share (for models that forecast EPS) or market capitalization (for models that forecast unscaled total earnings). A positive bias indicates that the forecast is higher than the actual,i.e. optimistic forecasts. Conversely, a negative bias on the other hand indicates that forecasts are pessimistic. The results are presented in Table 5.

Panel A presents the results for the international sample. Unsurprisingly, the RW model performs poorly, especially as the horizon gets longer. The mean bias increases in magnitude from -0.016 for one-year-ahead to -0.083 for three-year-ahead forecasts. This is because the static RW model is inherently pessimistic, as it does not incorporate any growth into its forecasts. Among the cross-sectional forecasts, the HVZ shows a high level of optimistic bias - i.e. the actual earnings are considerably less than the forecasted earnings. The EP model also produces optimistically biased forecasts, but the bias is far less than that of the HVZ model. The cross-sectional model that performs the best is the RI model for which the mean bias across the three-years are insignificantly different from zero. The linear model has a high level of bias compared with RI. Among the ML models, the bias of Lasso and Ridge models is generally comparable to that of the RI model. XGBoost produce slightly biased forecasts, but the level of bias is relatively low, especially for one-year and two-year-ahead forecasts. The mean bias for XGBoost is 0.002, -0.002, -0.008, -0.019, and -0.034 for the five horizons, respectively. What the bias results indicate is that researchers may have a trade-off to make with model selection in the US context. The models that produce the least biased forecasts are not the same as the models that produce the most accurate forecasts. Among the cross-sectional models, the RI model produces forecasts that are reasonably accurate (Tables 1 and 2) as well as unbiased. Among the ML models, the XGBoost model produces forecasts that are the most accurate (Tables 1 and 2) and have, what might be considered, an acceptable level of bias.

Panel B presents the results for the international sample. Here too, the static RW model performs poorly, especially as the horizon gets longer. The mean bias increases from -0.013 to -0.054 for one- and five-year-ahead forecasts respectively. Among the cross-sectional forecasts, the RI model dominates with the least biased forecasts for all three horizons. As for the linear model, the level bias is also higher than the RI model in the international setting. Among the ML models, the bias results are very different compared to the US sample. The Lasso and Ridge models produce extremely biased (optimistic) forecasts. The XGBoost model performs the best, with unbiased forecasts for all horizons, except for one-year-ahead forecasts. Combining these bias results with the earlier results for forecast accuracy produces a clear winner in the international context. The XGBoost model produces the most accurate forecasts (Tables 3 and 4) which are also mostly unbiased. Among the cross-sectional models, the RI model produces unbiased forecasts, but the forecast error can often be higher than that of the naive RW model.

4.4. Performance of Model-Based ICCs

Our final set of tests uses the Implied cost of capital (ICC) paradigm to test the quality of the forecasts. As Mohanram and Gode (2013) show, the poor performance of ICC metrics can be largely attributed to the poor quality of the forecasts. When forecasts are accurate, the ICCs generated from them "perform well" - i.e. they are a reliable measure of expected returns. Li and Mohanram (2014) test their proposed cross-sectional models (EP and RI) and show that these models generate better ICCs than those from the HVZ model.

Using forecasts from each of the cross-sectional (HVZ, EP and RI) as well as ML models (Lasso, Ridge, and XGBoost), we generate measures of ICC, which is defined as the average of the ICC from the GLS,CT, PEG and OJ models. Note that the naive RW model cannot be used to compute ICC, as it does not provide any estimate of earnings growth which is a prerequisite for two of the ICC models (PEG and OJ). We test the performance of the ICCs generated from the different forecasting models using two sets of tests. First, we run univariate regressions of realized returns on the measure of ICC and test how close the coefficient on the ICC is to the theoretical benchmark of "1". Second, we create portfolios based on the level of ICC and examine the pattern of returns to see if there is a monotonic relationship between ICC and future returns.

4.4..1 US results

Table 6 presents the results for the sample of US firms. Panel A presents the results of the univariate regression of ICC on realized returns. Each of the seven models produces ICCs that are significantly positively correlated with future returns. Among all models, HVZ performs the worst with the smallest coefficient of 0.493, while the regression coefficients are close to the benchmark of one for all other models,

Panel B of Table 6 presents the portfolio results for the US sample across deciles of ICC generated using each of the seven measures. In this table, a model can be deemed to perform well if we find a significant spread in returns across extreme ICC deciles, and if the spread in realized returns is comparable to the spread in the ICC. For all seven models, we find significant return spreads between the lowest and highest ICC deciles. The return spread for HVZ (11.08%) is far lower than the spread in ICC (28.47%). The performance for all other models are similar and no model stands out from the crowd. The XGBoost model performs well with a return spread of 14.1% and a spread in ICC of 10.40%.

Overall, the results from Table 6 highlight the importance of forecast quality for the estimation of ICCs. It is not surprising that the models that performs reasonably well in ICC tests (e.g., RI and XGBoost) also perform well in forecasting errors and bias (Tables 2 and 5).

4.4..2 International results

Table 7 presents the results for the international sample. Panel A presents the results of the univariate regression of ICC on realized returns. XGBoost is the only model that has a coefficient indistinguishable from one. RI also has good performance, with the coefficent being only margianlly different from one. All other models perform poorly. Lasso and Ridge actually show statistically insignificant correlation with realized returns, and HVZ, EP, and Linear produce coefficients are relatively far from one.

Panel B of Table 7 presents the portfolio results for the international sample across deciles of ICC generated using each of the seven measures. These results also mirror the regression results. HVZ, Linear, Lasso, and Ridge produce return spreads that are much smaller than the spreads in ICCs. EP and RI models perform moderately with return spreads of 12.21% and 11.02%, though still smaller than the spreads in ICCs of 16.69% and 19.26%, respectively. XGBoost performs the best with a return spread of 10.27%, which is comparable to the spread in ICC of 1225%.

The results from Table 7 highlight an important contribution of this paper. One particular model, the XGBoost model, dominates all other models in its ability to generate the most accurate and least biased forecasts, in the international setting where information environment is relatively poor. Unsurprisingly, it also generates ICCs that perform the best.

5. Conclusion

In this paper, we test whether recently developed machine learning (ML) techniques can help researchers seeking to generate accurate and unbiased forecasts of future earnings, and whether these forecasts can lead to better estimates of implied cost of capital (ICC). We examine these questions, not just in US firms like most prior research, but also in an international sample. We consider three ML models - Lasso regression, Ridge regression, and Extreme Gradient Boosting (XGboost). We benchmark the performance of these models against both a naive random walk (RW) model as well as extant cross-sectional models of forecasting, specifically the HVZ model from Hou et al. (2012) and the earnings persistence (EP) and residual income (RI) models from Li and Mohanram (2014). We also benchmark the machine-learning models against a simple linear model with augmented set of 60 predictors

Within the US sample, we find that the XGBoost Model performs well, generating forecasts with the greatest ex-post accuracy. The EP and RI models also perform reasonably well, while the HVZ, Lasso and Ridge models perform poorly. The improvements generated by the XGBoost model, while not dramatic, are concentrated in the important subgroups of small firms and firms with volatile earnings. However, it is in the international sample where one ML model, the XGBoost model, really shines in its ability to generate forecasts with dramatically better forecasting accuracy. The results from ICC tests mirror the forecast accuracy tests - with the XGBoost model performing the strongest, especially for international firms.

The results of this paper have important methodological contributions for researchers in finance and accounting, striving to generate accurate earnings forecasts and reliable measures of expected risk. We recommend that future research use the XGBoost model to generate estimates of future earnings as well as ICC. This recommendation is particularly important in the subset of international earnings for two reasons. First, our results show that the cross-sectional models that perform moderately well in the US sample, do not fare as well internationally. Second, the problem of scarce coverage and volatile earnings is likely to more severe in international settings, and these are some of the subsamples in which the XGBoost model does extremely well. Moreover, our paper extends prior studies that use machine-learning based techniques to forecast earnings. Prior research mostly employs models such as random forest or gradient boosting models, which takes significant amount of time and computing power to execute. Our paper builds on the recent innovation in the machine learning field and show that the XGBoost model is able to achieve superior forecasting performance with much lower demand for computing resources.

We must mention that ours is only a first attempt at showing that ML models can add a lot of value in both forecasting space as well as the estimation of ICC. In fact, one can view our results as a lower bound of what ML models can do. We have used a simple and static (though reasonably exhaustive) set of potential explanatory variables in our estimation models. Using a wider set of variables, including non-financial variables as well as market-based signals, might also increase the accuracy of the forecasts and the performance of the ICCs from these forecasts. We leave this question for future research to examine.

References

Bailey, W., G. A. Karolyi, and C. Salva (2006). The economic consequences of increased disclosure: Evidence from international cross-listings. *Journal of Financial Economics 81*, 175–213.

Ball, R. (2016). Why we do international accounting research. *Journal of International Accounting Research 15*(2), 1–6.

Bhushan, R. (1989). Firms characteristics and analyst following. *Journal of Accounting and Economics 11*, 255–274.

Botosan, C. A. and M. A. Plumlee (2002). A Re-examination of Disclosure Level and the Expected Cost of Equity Capital. *Journal of Accounting Research 40*(1), 21–40.

Brown, L. D., G. D. Richardson, and S. J. Schwager (1987). An information interpretation of financial analyst superiority in forecasting earnings. *Journal of Accounting Research*, 49–67.

Cao, K. and H. You (2020). Fundamental analysis via machine learning. *HKUST Business School Research Paper* (2020-009).

Chattopadhyay, A., M. R. Lyle, and C. C. Wang (2022). Expected stock returns worldwide: A log-linear present-value approach. *The Accounting Review 97*(2), 107–133.

Chen, X., Y. H. Cho, Y. Dou, and B. Lev (2022). Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research 60*(2), 467–515.

Claus, J. and J. Thomas (2001). Equity premia as low as three percent? Evidence from analysts' earnings forecasts for domestic and international stock markets. *Journal of Finance 56*(5), 1629–1666.

Daske, H., L. Hail, C. Leuz, and R. Verdi (2008). Mandatory IFRS Reporting around the World: Early Evidence on the Economic Consequences. *Journal of Accounting Research 46*(5), 1085–1142.

De Silva, T. and D. Thesmar (2021). Noise in expectations: Evidence from analyst forecasts. Technical report, National Bureau of Economic Research.

Dechow, P. M. (1994). Accounting & Economics The role of accounting accruals. *Journal of Accounting and Economics 18*, 3–42.

Dhaliwal, D., L. Krull, O. Z. Li, and W. Moser (2005). Dividend Taxes and Implied Cost of Equity Capital. *Journal of Accounting Research 43*(5), 675–708.

Dhaliwal, D. S., O. Z. Li, A. Tsang, and Y. G. Yang (2011). Voluntary Nonfinancial Disclosure and the Cost of Equity Capital: The Initiation of Corporate Social Responsibility Reporting. *The Accounting Review 86*(1), 59–100.

Easton, P. and S. Monahan (2005). An evaluation of accounting-based measures of expected returns. *The Accounting Review 80*(2), 501–538.

Easton, P. D. (2004). PE Ratios, PEG Ratios, and Estimating the Implied Expected Rate of Return on Equity Capital. *The Accounting Review 79*(1), 73–95.

Easton, P. D. and G. A. Sommers (2007). Effect of Analysts' Optimism on Estimates of the Expected Rate of Return Implied by Earnings Forecasts. *Journal of Accounting Research 45*(5), 983–1015.

Elton, E. J. (1999). Expected Return, Realized Return, and Asset Pricing Tests. *Journal of Finance 54*(4), 1199–1220.

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

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

Fama, E. F. and K. R. French (2017). International tests of a five-factor asset pricing model. *Journal of Financial Economics 123*(3), 441–463.

Fang, B., O.-K. Hope, Z. Huang, and R. Moldovan (2020). The effects of mifid ii on sell-side analysts, buy-side analysts, and firms. *Review of Accounting Studies 25*(3), 855–902.

Feltham, G. a. and J. a. Ohlson (1995). Valuation and Clean Surplus Accounting for Operating and Financial Activities. *Contemporary Accounting Research 11*(2), 689–731.

Frankel, R. M. and C. M. C. Lee (1998). Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics 25*, 283–319.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Gebhardt, W. R., C. M. C. Lee, and B. Swaminathan (2001). Toward an Implied Cost of Capital. *Journal of Accounting Research 39*(1), 135–176.

Gerakos, J. and R. B. Gramacy (2013). Regression-based earnings forecasts.

Gode, D. and P. Mohanram (2003). Inferring the cost of capital using the Ohlson–Juettner Model. *Review of Accounting Studies 8*, 399–431.

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

Hail, L. and C. Leuz (2006). International Differences in the Cost of Equity Capital: Do Legal Institutions and Securities Regulation Matter? *Journal of Accounting Research 44*(3), 485–531.

Hansen, J. W. and C. Thimsen (2020). Forecasting corporate earnings with machine learning. *Working Paper*.

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

Healy, P., G. Serafeim, S. Srinivasan, and G. Yu (2014). Market competition, earnings management, and persistence in accounting profitability around the world. *Review of Accounting Studies 19*, 1281–1308.

Hou, K., G. A. Karolyi, and B. C. Kho (2011). What Factors Drive Global Stock Returns? *Review of Financial Studies 24*(8), 2527–2574.

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

Lang, M. H., K. V. Lins, and D. P. Miller (2003). ADRs, analysts, and accuracy: Does cross listing in the US improve a firm's information environment and increase market value. *Journal of Accounting Research 41*, 317–345.

Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review 4*(1), 1–44.

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.

Li, S. (2010). Does Mandatory Adoption of International Financial Reporting Standards in the European Union Reduce the Cost of Equity Capital ? *The Accounting Review 85*(2), 607–636.

Li, Y., D. T. Ng, and B. Swaminathan (2013). Predicting market returns using aggregate implied cost of capital. *Journal of Financial Economics 110*(2), 419–436.

Mohanram, P. and D. Gode (2013). Removing predictable analyst forecast errors to improve implied cost of equity estimates. *Review of Accounting Studies 18*, 443–478.

O'Brien, P. and R. Bhushan (1990). Analyst following and institutional ownership. *Journal of Accounting Research 28*, 55–82.

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*, 349–365.

van Binsbergen, J. H., X. Han, and A. Lopez-Lira (2020). Man vs. machine learning: The term structure of earnings expectations and conditional biases. Technical report, National Bureau of Economic Research.

Wiedman, C. I. (1996). The relevance of characteristics of the information environment in the selection of a proxy for the market's expectations for earnings: An extension of brown, richardson, and schwager [1987]. *Journal of Accounting Research 34*(2), 313–324.

Appendix: Description of Models and Variables

This table describes the models used for generating earnings forecasts and computations of ICCS, and defines the variables used in earnings forecast models, including traditional models (e.g., HVZ, EP, and RI) and machine learning models. All variables are obtained from Compustat North America or Compustat Global; non-U.S. fundamental data are converted to U.S. dollars using Compustat's exchange-rate file.

Variable	Description	Computation
A1. Earnings Forecasting Models		
	Models Estimated	
HVZ	Generate earnings forecasts using the model of Hou et al. (2012)	$\mathbb{E}[E_{i,t+\tau}] = \beta_0 + \beta_1 A_{i,t} + \beta_2 D_{i,t} + \beta_3 DD_{i,t} + \beta_4 E_{i,t} + \beta_5 Neg\_E_{i,t} + \beta_6 ACCRUAL_{i,t} + \epsilon_{i,t}$
EP	Generate earnings forecasts using the model of Li and Mohanram (2014)	$\mathbb{E}[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 Neg\_E_{i,t} + \beta_3 Neg\_E_{i,t} \times E_{i,t} + \epsilon_{i,t}$
RI	Generate earnings forecasts using the model of Li and Mohanram (2014)	$\mathbb{E}[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 Neg\_E_{i,t} + \beta_3 Neg\_E_{i,t} \times E_{i,t} + \beta_4 B_{i,t} + \beta_5 TACC_{i,t} + \epsilon_{i,t}$
A2. Computation of ICCS		
	Valuation Models	
ICC_{GLS}	R_e computed using the model in Gebhardt et al. (2001)	$P_{i,t} = B_{i,t} + \sum_{\tau=1}^{11} \frac{\mathbb{E}_t[E_{i,t+\tau}] - (R_e - 1) \times \mathbb{E}_t[B_{i,t+\tau-1}]}{(R_e)^\tau} + \frac{\mathbb{E}_t[E_{i,t+12}] - (R_e - 1) \times \mathbb{E}_t[B_{i,t+11}]}{(R_e - 1)(R_e)^{11}}$
ICC_{CT}	R_e computed using the model in Claus and Thomas (2001)	$P_{i,t} = B_{i,t} + \sum_{\tau=1}^{3} \frac{\mathbb{E}_t[E_{i,t+\tau}] - (R_e - 1) \times \mathbb{E}_t[B_{i,t+\tau-1}]}{(R_e)^\tau} + \frac{\mathbb{E}_t[E_{i,t+3}] - (R_e - 1) \times \mathbb{E}_t[B_{i,t+2}]}{((R_e - 1) - g)(R_e)^3}(1+g)$
ICC_{PEG}	R_e computed using the "PEG" model in Easton (2004)	$R_e = 1 + \sqrt{\frac{E_{i,t+2} - E_{i,t+1}}{P_{i,t}}}$
ICC_{OJ}	R_e computed using the model in Ohlson and Juettner-Nauroth (2005)	$P_t = \frac{E_{t+1}}{(R_e - 1)} + \frac{E_{t+1}(E_{t+2} + (R_e - 1)D_{t+1} - (R_e)E_{t+1})}{(R_e - 1)(R_e - 1) - \frac{E_{t+3} + (R_e - 1)D_{t+2} - R_e I_{t+2}}{E_{t+2} + (R_e - 1)D_{t+1} - R_e E_{t+1}})}$
A3. Definitions of Variables in the HVZ Model		
$E_{i,t+\tau}$	Earnings in year t+τ	ib-spi
A_t	Total assets in year t	at

Variable	Description	Computation
D_t	Dividend payment in year t	dvc
DD_t	Dividend Payer Indicator	An indicator variable that equals 1 if dividend is higher than 0
Neg_E_t	Negative earnings indicator	An indicator variable that equals 1 for firms with negative earnings
$Accruals$	Accruals	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)
A4. Definitions of Variables in *EP* and *RI* Models		
$E_{i,t+\tau}$	Earnings per share in year t+τ	((ib-spi)/csho
Neg_E_t	Negative earnings indicator	An indicator variable that equals 1 for firms with negative earnings
B	Book value of equity per share	ceq/csho
$TACC$	Total accruals	Sum of the change in WC ((act - che) - (lct - dlc)), change in NCO ((at - act - ivao) - (lt - lct - dltt)), and change in FIN ((ivst + ivao) - (dltt + dlc + pstk))
A5. Definitions of Variables in Machine Learning Models		
$Sale$	Total sales	sale/csho
$COGS$	Cost of goods sold	cogs/csho
$XSGA$	Selling, general, and administrative expenses	xsga/csho
XAD	Advertising expense	xad/csho
XRD	Research and development expense	xrd/csho
DP	Depreciation and amortization	dp/csho
$XINT$	Interest and related expense	xint/csho
$NOPIO$	Non-operating income *expense*	nopio/csho
TXT	Income taxes	txt/csho
$XIDO$	Extraordinary items and discontinued operations	xido/csho
EPS	Earnings	(ib - spi)/csho
DVC	Common dividend	dvc/csho
CHE	Cash and short-term investments	che/csho
$INVT$	Inventories	invt/csho
$RECT$	Receivables	rect/csho
ACT	Total current assets	act/csho

Variable	Description	Computation
$PPENT$	Property, plant, and equipment (Net)	ppent/csho
$IVAO$	Investments and advances	ivao/csho
$INTAN$	Intangible assets	intan/csho
AT	Total assets	at/csho
AP	Accounts payable	ap/csho
DLC	Debt in current liabilities	dlc/csho
TXP	Income taxes payable	txp/csho
LCT	Total current liabilities	lct/csho
$DLTT$	Long-term debt	dltt/csho
LT	Total liabilities	lt/csho
CEQ	CommonOrdinary equity	ceq/csho
CFO	Cash flow from operating activities	(oancf - xidoc)/csho
$GDP\ Growth$	GDP Growth Rate	
$Unemployment$	Unemployment Rate	
$IPT\ Growth$	Growth in total industrial production	
$Consumption\ Growth$	Consumption Growth Rate	

Table 1. Forecasting Performance for the US Sample

This table presents the time-series average of the mean absolute forecasting errors (MAFE) using the US sample for both traditional models and machine learning models. Forecasting error for the HVZ model is calculated as the absolute value of the difference between forecast earnings and actual earnings, scaled by market value of equity at the fiscal year end. Forecasting error for all other models (RW, EP, RI, Linear, Lasso, Ridge, and XGBoost) is calculated as the absolute value of the difference between forecast earnings per share and actual earnings per share, scaled by the stock price at the prior fiscal year end. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively

	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)
Model	Mean	Mean	Mean	Mean	Mean
HVZ	0.081***	0.127***	0.168***	0.208***	0.255***
RW	0.065***	0.093***	0.116***	0.139***	0.163***
EP	0.060***	0.082***	0.101***	0.120***	0.140***
RI	0.060***	0.080***	0.097***	0.115***	0.135***
Linear	0.063***	0.092***	0.113***	0.128***	0.152***
Lasso	0.059***	0.082***	0.102***	0.121***	0.141***
Ridge	0.059***	0.081***	0.101***	0.120***	0.140***
XGB	0.056***	0.076***	0.094***	0.112***	0.133***
Comparison					
XGB - RW	-0.009***	-0.017***	-0.022***	-0.027***	-0.030***
XGB - EP	-0.004***	-0.006***	-0.007***	-0.008***	-0.008***
XGB - RI	-0.004***	-0.004***	-0.003***	-0.003***	-0.002*

Table 2. Forecasting Performance in the US: Sub-sample Analyses

This table presents the time-series average of the mean absolute forecasting errors of the traditional and machine learning models for sub-samples within the US. Panels A, B, and C reports results of the sample partitioned by firm size, analyst coverage, and earnings volatility, respectively. Forecasting error for the HVZ model is calculated as the absolute value of the difference between forecast earnings and actual earnings, scaled by market value of equity at the fiscal year end. Forecasting error for all other models (RW, EP, RI, Lasso, Ridge, and XGBoost) is calculated as the absolute value of the difference between forecast earnings per share and actual earnings per share, scaled by the stock price at the fiscal year end. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively

Panel A: Partition Analyses by Firm Size

	Small Firms					Large Firms				
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Model	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
HVZ	0.125***	0.202***	0.271***	0.342***	0.424***	0.038***	0.057***	0.072***	0.087***	0.106***
RW	0.092***	0.131***	0.163***	0.196***	0.228***	0.039***	0.058***	0.073***	0.089***	0.105***
EP	0.084***	0.114***	0.141***	0.169***	0.198***	0.037***	0.052***	0.064***	0.077***	0.090***
RI	0.083***	0.110***	0.133***	0.158***	0.186***	0.038***	0.052***	0.064***	0.077***	0.091***
Linear	0.089***	0.132***	0.162***	0.181***	0.212***	0.038***	0.055***	0.068***	0.080***	0.099***
Lasso	0.081***	0.112***	0.140***	0.167***	0.194***	0.037***	0.054***	0.067***	0.080***	0.094***
Ridge	0.081***	0.111***	0.137***	0.164***	0.193***	0.037***	0.054***	0.067***	0.080***	0.095***
XGB	0.077***	0.103***	0.127***	0.151***	0.178***	0.036***	0.052***	0.064***	0.077***	0.092***
Comparison										
XGB - RW	-0.015***	-0.028***	-0.036***	-0.045***	-0.050***	-0.003***	-0.006***	-0.010***	-0.012***	-0.013***
XGB - EP	-0.007***	-0.011***	-0.014***	-0.018***	-0.019***	-0.001***	-0.001	-0.000	0.001	0.002*
XGB - RI	-0.006***	-0.007***	-0.006***	-0.007***	-0.007***	-0.002***	-0.001*	-0.001	0.000	0.002*

Table 2. (Continued)

Panel B: Partition Analyses by Analyst Coverage

	No Coverage					With Coverage				
	$t+1$ (1)	$t+2$ (2)	$t+3$ (3)	$t+4$ (4)	$t+5$ (5)	$t+1$ (6)	$t+2$ (7)	$t+3$ (8)	$t+4$ (9)	$t+5$ (10)
Model	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
HVZ	0.111***	0.181***	0.242***	0.300***	0.366***	0.055***	0.084***	0.103***	0.126***	0.155***
RW	0.079***	0.112***	0.139***	0.166***	0.193***	0.054***	0.078***	0.098***	0.112***	0.134***
EP	0.075***	0.102***	0.126***	0.150***	0.175***	0.051***	0.065***	0.080***	0.097***	0.104***
RI	0.074***	0.098***	0.119***	0.140***	0.165***	0.051***	0.063***	0.077***	0.093***	0.101***
Linear	0.079***	0.117***	0.143***	0.158***	0.184***	0.050***	0.068***	0.083***	0.097***	0.105***
Lasso	0.072***	0.100***	0.125***	0.149***	0.173***	0.048***	0.062***	0.078***	0.094***	0.104***
Ridge	0.072***	0.099***	0.123***	0.147***	0.172***	0.048***	0.063***	0.077***	0.093***	0.104***
XGB	0.069***	0.092***	0.113***	0.134***	0.159***	0.047***	0.060***	0.073***	0.088***	0.098***
Comparison										
XGB - RW	-0.010***	-0.020***	-0.026***	-0.032***	-0.035***	-0.007***	-0.019***	-0.025***	-0.024***	-0.036**
XGB - EP	-0.006***	-0.010***	-0.013***	-0.016***	-0.016***	-0.004***	-0.006***	-0.008**	-0.009**	-0.006**
XGB - RI	-0.005***	-0.006***	-0.006***	-0.006***	-0.006***	-0.004***	-0.004***	-0.004***	-0.005*	-0.002**

Table 2. (Continued)

Panel C: Partition Analyses by Earnings Volatility

	High Volatility					Low Volatility				
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Model	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
HVZ	0.102***	0.160***	0.207***	0.248***	0.295***	0.055***	0.085***	0.113***	0.139***	0.170***
RW	0.083***	0.117***	0.142***	0.168***	0.192***	0.046***	0.068***	0.089***	0.109***	0.130***
EP	0.081***	0.108***	0.129***	0.151***	0.171***	0.042***	0.058***	0.073***	0.088***	0.104***
RI	0.080***	0.104***	0.122***	0.142***	0.161***	0.042***	0.058***	0.073***	0.088***	0.103***
Linear	0.081***	0.118***	0.144***	0.158***	0.172***	0.043***	0.063***	0.078***	0.092***	0.108***
Lasso	0.076***	0.103***	0.126***	0.148***	0.168***	0.041***	0.059***	0.075***	0.091***	0.107***
Ridge	0.075***	0.102***	0.124***	0.146***	0.166***	0.041***	0.059***	0.075***	0.091***	0.107***
XGB	0.074***	0.097***	0.114***	0.135***	0.159***	0.040***	0.058***	0.072***	0.088***	0.105***
Comparison										
XGB - RW	-0.009***	-0.020***	-0.028***	-0.033***	-0.034***	-0.006***	-0.011***	-0.017***	-0.021***	-0.024***
XGB - EP	-0.006***	-0.011***	-0.015***	-0.016***	-0.012***	-0.002***	-0.001	-0.001**	-0.000	0.001
XGB - RI	-0.006***	-0.007***	-0.008***	-0.007***	-0.002	-0.002***	-0.001	-0.001***	-0.000	0.002*

Table 3. Forecasting Performance for the International Sample

This table presents the time-series average of the MAFE in our international sample for both traditional models and machine learning models. Forecasting error for the HVZ model is calculated as the absolute value of the difference between forecast earnings and actual earnings, scaled by market value of equity at the fiscal year end. Forecasting error for all other models (RW, EP, RI, Lasso, Ridge, and XGBoost) is calculated as the absolute value of the difference between forecast earnings per share and actual earnings per share, scaled by the stock price at the fiscal year end. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively

	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)
Model	Mean	Mean	Mean	Mean	Mean
HVZ	0.141***	0.200***	0.260***	0.320***	0.370***
RW	0.090***	0.109***	0.118***	0.126***	0.138***
EP	0.100***	0.127***	0.139***	0.152***	0.168***
RI	0.093***	0.119***	0.127***	0.138***	0.152***
Linear	0.127***	0.171***	0.192***	0.201***	0.215***
Lasso	0.144***	0.186***	0.216***	0.226***	0.221***
Ridge	0.152***	0.204***	0.246***	0.271***	0.344***
XGB	0.082***	0.095***	0.104***	0.111***	0.124***
Comparison					
XGB - RW	-0.008***	-0.014***	-0.014***	-0.015***	-0.014***
XGB - EP	-0.019***	-0.032***	-0.033***	-0.036***	-0.044***
XGB - RI	-0.011***	-0.024***	-0.021***	-0.023***	-0.028***

Table 4. Forecasting Performance Internationally: Sub-sample Analyses

This table presents the time-series average of the MAFE of the traditional and machine learning models for sub-samples within the international data. Panels A, B, and C reports results of the sample partitioned by firm size, analyst coverage, and earnings volatility, respectively. Forecasting error for the HVZ model is calculated as the absolute value of the difference between forecast earnings and actual earnings, scaled by market value of equity at the fiscal year end. Forecasting error for all other models (RW, EP, RI, Lasso, Ridge, and XGBoost) is calculated as the absolute value of the difference between forecast earnings per share and actual earnings per share, scaled by the stock price at the fiscal year end. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively

Panel A: Partition Analyses by Firm Size

	Small Firms					Large Firms				
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Model	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
HVZ	0.239***	0.351***	0.459***	0.592***	0.730***	0.045***	0.057***	0.069***	0.074***	0.080***
RW	0.135***	0.163***	0.174***	0.196***	0.214***	0.045***	0.058***	0.066***	0.070***	0.079***
EP	0.144***	0.179***	0.197***	0.219***	0.255***	0.058***	0.079***	0.083***	0.091***	0.099***
RI	0.134***	0.166***	0.179***	0.198***	0.230***	0.053***	0.074***	0.078***	0.084***	0.091***
Linear	0.179***	0.238***	0.265***	0.287***	0.305***	0.076***	0.109***	0.122***	0.132***	0.145***
Lasso	0.204***	0.260***	0.303***	0.325***	0.317***	0.089***	0.119***	0.140***	0.153***	0.147***
Ridge	0.212***	0.283***	0.341***	0.386***	0.494***	0.096***	0.133***	0.162***	0.186***	0.235***
XGB	0.120***	0.136***	0.146***	0.164***	0.183***	0.045***	0.057***	0.065***	0.069***	0.078***
Comparison										
XGB - RW	-0.016***	-0.028***	-0.028***	-0.032***	-0.031***	-0.000	-0.002	-0.002	-0.001	-0.001
XGB - EP	-0.024***	-0.043***	-0.050***	-0.055***	-0.072***	-0.013***	-0.022***	-0.017***	-0.019***	-0.021***
XGB - RI	-0.014***	-0.031***	-0.031***	-0.034***	-0.047***	-0.009***	-0.017***	-0.011***	-0.012***	-0.013***

Table 4. (Continued)

Panel B: Partition Analyses by Analyst Coverage

	No Coverage					With Coverage				
	$t+1$ (1)	$t+2$ (2)	$t+3$ (3)	$t+4$ (4)	$t+5$ (5)	$t+1$ (6)	$t+2$ (7)	$t+3$ (8)	$t+4$ (9)	$t+5$ (10)
Model	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
HVZ	0.224***	0.332***	0.425***	0.555***	0.656***	0.098***	0.133***	0.163***	0.193***	0.244***
RW	0.116***	0.143***	0.155***	0.175***	0.195***	0.076***	0.092***	0.098***	0.104***	0.116***
EP	0.127***	0.157***	0.176***	0.198***	0.226***	0.087***	0.113***	0.117***	0.128***	0.144***
RI	0.118***	0.147***	0.161***	0.181***	0.206***	0.081***	0.105***	0.108***	0.116***	0.130***
Linear	0.164***	0.211***	0.240***	0.262***	0.268***	0.107***	0.151***	0.161***	0.172***	0.192***
Lasso	0.179***	0.226***	0.270***	0.289***	0.277***	0.128***	0.167***	0.182***	0.200***	0.197***
Ridge	0.190***	0.250***	0.300***	0.337***	0.411***	0.133***	0.182***	0.211***	0.245***	0.314***
XGB	0.107***	0.125***	0.138***	0.157***	0.169***	0.069***	0.080***	0.084***	0.090***	0.104***
Comparison										
XGB - RW	-0.009***	-0.018***	-0.017***	-0.018***	-0.026***	-0.007***	-0.012***	-0.014***	-0.014***	-0.012**
XGB - EP	-0.020***	-0.032***	-0.037***	-0.040***	-0.057***	-0.018***	-0.033***	-0.030***	-0.034***	-0.040***
XGB - RI	-0.011***	-0.022***	-0.022***	-0.023***	-0.037***	-0.012***	-0.025***	-0.020***	-0.022***	-0.026***

Table 4. (Continued)

Panel C: Partition Analyses by Earnings Volatility

	High Volatility					Low Volatility				
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Model	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
HVZ	0.182***	0.264***	0.329***	0.392***	0.507***	0.090***	0.127***	0.159***	0.198***	0.240***
RW	0.103***	0.124***	0.131***	0.137***	0.157***	0.066***	0.083***	0.094***	0.105***	0.113***
EP	0.116***	0.144***	0.154***	0.165***	0.201***	0.079***	0.104***	0.109***	0.124***	0.136***
RI	0.106***	0.133***	0.140***	0.149***	0.181***	0.074***	0.097***	0.101***	0.114***	0.124***
Linear	0.151***	0.194***	0.204***	0.218***	0.257***	0.104***	0.146***	0.158***	0.178***	0.187***
Lasso	0.160***	0.204***	0.221***	0.238***	0.251***	0.117***	0.156***	0.175***	0.206***	0.191***
Ridge	0.175***	0.229***	0.256***	0.299***	0.395***	0.123***	0.171***	0.206***	0.245***	0.309***
XGB	0.093***	0.105***	0.110***	0.115***	0.143***	0.063***	0.076***	0.086***	0.097***	0.106***
Comparison										
XGB - RW	-0.010***	-0.019***	-0.022***	-0.022***	-0.014	-0.003*	-0.007**	-0.008***	-0.008**	-0.007*
XGB - EP	-0.023***	-0.039***	-0.040***	-0.046***	-0.058***	-0.016***	-0.028***	-0.024***	-0.027***	-0.031***
XGB - RI	-0.013***	-0.028***	-0.026***	-0.029***	-0.039***	-0.011***	-0.021***	-0.016***	-0.017***	-0.019***

Table 5. Forecasting Bias

This table presents the time-series average of the mean forecasting bias for both the traditional and machine learning models. Panel A reports results for the US sample while Panel B reports results for the international sample. Forecasting bias for the HVZ model is calculated as the difference between forecast and actual earnings, scaled by market value of equity at the fiscal year end. Forecasting bias for all other models (RW, EP, RI, Lasso, Ridge, and XGBoost) is calculated as the difference between forecast earnings per share and actual earnings per share, scaled by the stock price at the fiscal year end. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively

Panel A: Forecasting Bias for the US Sample

	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)
Model	Mean	Mean	Mean	Mean	Mean
HVZ	0.034***	0.066***	0.092***	0.115***	0.146***
RW	-0.016***	-0.032***	-0.048***	-0.065***	-0.083***
EP	0.010***	0.013***	0.013***	0.011**	0.006
RI	0.004**	0.003	-0.002	-0.009	-0.017**
Linear	0.013***	0.027***	0.025***	0.011*	-0.024**
Lasso	0.004*	0.002	0.000	-0.005	-0.017**
Ridge	0.005***	0.002	-0.002	-0.009	-0.020***
XGB	0.002	-0.002	-0.008**	-0.019***	-0.034***

Panel B: Forecasting Bias for the International Sample

	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
	(1)	(2)	(3)	(4)	(5)
Model	Mean	Mean	Mean	Mean	Mean
HVZ	0.082***	0.138***	0.198***	0.253***	0.304***
RW	-0.013***	-0.027***	-0.036***	-0.042***	-0.054***
EP	0.039***	0.050***	0.051***	0.061***	0.070***
RI	0.021***	0.026***	0.025**	0.032***	0.040***
Linear	0.047***	0.088***	0.089***	0.068***	0.064**
Lasso	0.040***	0.033	0.064***	0.075***	0.080***
Ridge	0.033**	0.041*	0.093***	0.117***	0.144***
XGB	0.013***	0.006	0.003	-0.003	-0.007

Table 6. Performance of Model-Based ICCs for the US Sample

This table presents the performance of model-based ICCs for the US sample. The implied cost of capital is computed as the average value based on four models, GLS, CT, PEG, and OJ. Panel A presents the univariate Fama-MacBeth regression results, with one-year-ahead realized return as the dependent variable and the anuual model-based ICC as the independent variable. Panel B presents the results of firms sorted into deciles by the annual model-based ICCs. The odd columns in Panel B report the equal-weighted mean anuual ICC of the portfolios, while the even columns report the equal-weighted mean realized annual returns of the portfolios. The last rows of Panel B report results of the spread between the highest and lowest decile of firms. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively.

Panel A: Regression Analyses

	(1)	(2)	(3)	(4)	(5)	(6)
Model	Slope Coeff	t-stat	Intercept	t-stat	R^2	F-test for Slope = 1
HVZ	0.493***	[5.17]	0.028	[1.19]	0.019	28.12***
EP	1.007***	[4.66]	0.005	[0.23]	0.020	0.00
RI	1.054***	[5.19]	0.009	[0.41]	0.018	0.07
Lasso	1.125***	[4.57]	0.013	[0.65]	0.022	0.26
Linear	1.118***	[4.72]	0.013	[0.65]	0.021	0.25
Ridge	1.135***	[4.47]	0.011	[0.57]	0.022	0.28
XGB	1.153***	[4.77]	0.009	[0.43]	0.022	0.40

Table 6. (Continued)

Panel B: Portfolio Analyses

	HVZ		*RI*		*EP*		*Linear*		*Lasso*		*Ridge*		*XGB*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Decile	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns
1	4.57	6.96	4.64	6.76	4.01	5.67	4.05	5.20	4.09	5.13	4.06	4.91	3.97	4.22
2	6.86	6.14	5.75	7.06	5.32	6.26	5.29	8.41	5.18	6.88	5.12	7.25	5.11	7.36
3	8.24	6.73	6.40	7.01	5.98	7.39	5.97	8.38	5.79	8.25	5.74	7.86	5.74	7.66
4	9.45	7.40	6.95	8.43	6.57	8.57	6.60	7.32	6.38	7.91	6.30	8.82	6.34	9.02
5	10.62	8.57	7.53	8.27	7.18	8.37	7.28	10.04	6.98	8.99	6.89	8.81	6.93	8.61
6	12.04	8.84	8.21	9.95	7.87	10.38	8.03	9.77	7.68	12.06	7.58	11.82	7.61	10.01
7	13.79	9.97	9.03	10.77	8.71	12.68	9.00	11.77	8.59	11.28	8.46	11.64	8.39	10.48
8	16.42	12.18	10.13	11.03	9.77	10.79	10.30	12.40	9.83	11.69	9.68	11.95	9.42	11.60
9	20.74	16.13	11.88	12.67	11.36	12.58	12.23	12.89	11.67	14.07	11.53	13.99	10.94	14.05
10	33.03	18.04	16.53	19.01	15.10	18.62	17.01	20.60	15.98	20.50	15.80	19.73	14.36	18.32
Spread	28.47***	11.08***	11.89***	12.25***	11.09***	12.95***	12.96***	15.40***	11.89***	15.37***	11.74***	14.82***	10.40***	14.10***

Table 7. Performance of Model-Based ICCs for the International Sample

Panel A: Regression Analyses

This table presents the performance of model-based ICCs for the international sample. The implied cost of capital is computed as the average value based on four models, GLS, CT, PEG, and OJ. Panel A presents the univariate Fama-MacBeth regression results, with one-year-ahead realized return as the dependent variable and the anuual model-based ICC as the independent variable. Panel B presents the results of firms sorted into deciles by the annual model-based ICCs. The odd columns in Panel B report the equal-weighted mean anuual ICC of the portfolios, while the even columns report the equal-weighted mean realized annual returns of the portfolios. The last rows of Panel B report results of the spread between the highest and lowest decile of firms. The t-statistics are reported in the parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Model	Slope Coeff	t-stat	Intercept	t-stat	R^2	F-test for Slope = 1
HVZ	0.103**	[2.28]	0.058	[1.53]	0.004	394.75***
EP	0.498***	[4.01]	0.026	[0.59]	0.008	16.25***
RI	0.742***	[5.76]	0.016	[0.38]	0.009	4.00*
Lasso	0.186	[1.40]	0.055	[1.09]	0.008	37.53***
Linear	0.170*	[1.73]	0.059	[1.44]	0.006	71.36***
Ridge	0.098	[1.22]	0.058	[1.25]	0.005	124.96***
XGB	0.722***	[3.72]	0.022	[0.49]	0.009	2.05

Table 7. (Continued)

Panel B: Portfolio Analyses

	HVZ		*RI*		*EP*		*Linear*		*Lasso*		*Ridge*		*XGB*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Decile	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns	ICC	Realized Returns
1	2.91	3.46	2.22	1.40	1.83	0.83	1.94	6.20	2.03	4.88	2.10	4.60	2.09	1.65
2	5.09	5.10	4.37	3.98	3.58	3.63	3.24	7.67	3.34	7.73	3.63	6.58	3.33	4.10
3	6.69	5.44	5.73	3.69	4.73	3.80	4.44	6.87	4.65	6.92	5.13	6.26	4.19	6.23
4	8.26	6.43	6.83	6.83	5.71	5.06	5.79	6.25	6.17	6.35	6.85	6.54	5.01	6.45
5	10.00	7.63	7.90	6.23	6.68	6.59	7.36	6.52	7.87	6.31	8.68	7.69	5.84	6.88
6	12.08	7.61	9.06	8.53	7.72	7.39	9.24	5.55	9.77	6.15	10.78	6.98	6.71	7.71
7	14.70	8.63	10.39	8.17	8.94	9.42	11.65	6.72	12.00	7.27	13.32	7.76	7.68	8.44
8	18.55	9.99	12.11	10.42	10.52	10.01	14.89	7.51	14.83	6.90	16.80	7.80	8.85	8.21
9	25.62	9.09	14.68	10.75	12.80	12.67	19.57	7.58	18.91	8.16	21.94	7.52	10.47	9.78
10	46.85	9.03	21.48	12.42	18.52	13.04	29.73	10.48	27.89	10.17	33.50	9.12	14.34	11.92
Spread	43.94***	5.57**	19.26***	11.02***	16.69***	12.21***	27.79***	4.29	25.86***	5.29	31.40***	4.52*	12.25***	10.27***