Interpretive Earnings Forecasts via Machine Learning: A High-Dimensional Financial Statement Data Approach

Dieter Hess* Frederik Simon† Sebastian Weibels‡

First Version: November 2023; This Version: May 2024

Abstract

We predict earnings for forecast horizons of up to five years by using the entire set of Compustat financial statement data as input and providing it to state-of-the-art machine learning models capable of approximating arbitrary functional forms. Our approach yields on average 12% more accurate one-year-ahead predictions than the best-performing traditional linear approach. This superior performance is consistent across a variety of evaluation metrics and translates into more profitable investment strategies. Extensive model interpretation reveals that income statement variables, especially different definitions of earnings, are by far the most important predictors. Importantly, however, we find that the relevance of income statement variables consistently declines, whereas the relevance of balance sheet information consistently increases with increasing forecast horizon. Lastly, we show that simple interactions among financial statement variables play virtually no role for earnings predictions.

JEL classification: G11, G12, G17, G31, G32, M40, M41
Keywords: Earnings Forecasts, Cross-Sectional Earnings Models, Machine Learning

*University of Cologne, Department of Business Administration and Corporate Finance and Centre for Financial Research (CFR), hess@wiso.uni-koeln.de

†University of Cologne, Department of Business Administration and Corporate Finance and Centre for Financial Research (CFR), simon.frederik@wiso.uni-koeln.de

‡University of Cologne, Institute of Econometrics and Statistics, weibels@wiso.uni-koeln.de

1 Introduction

Future earnings are of central importance against the background of deriving the intrinsic value of assets (e.g., Monahan, 2018). In particular, analysts use earnings forecasts to derive buy/sell recommendations for stocks (e.g., Schipper, 1991; Brown, 1993). Earnings are also used in corporate decision-making, as they, inter alia, represent one of the primary financial metrics for external stakeholders in general (Graham et al., 2005). Lastly, as first shown by Ball and Brown (1968), earnings are assumed to directly relate to stock returns. One may even derive return expectations directly from predicted earnings in the form of the implied cost of capital (ICC) of a company (e.g., Gordon and Gordon, 1997; Claus and Thomas, 2001; Gebhardt et al., 2001; Easton, 2004; Ohlson and Juettner-Nauroth, 2005).

In general, earnings predictions are usually either retrieved from analysts or from statistical models. Traditionally, statistical approaches have primarily been of simple, linear nature (e.g., Hou et al., 2012; Li and Mohanram, 2014).[1] With the advent of more advanced statistical models, i.e., machine learning approaches in particular, various more flexible approaches have been proposed by researchers (e.g., Cao and You, 2021; Jones et al., 2023).

As outlined by Gu et al. (2020) and Israel et al. (2020), machine learning introduces flexibility in terms of three dimensions: first, it allows for flexibility in terms of functional form. That is, in contrast to the traditional linear approaches that have been used to predict earnings, machine learning allows for complex non-linear functional forms. Second, machine learning allows the use of large conditioning information sets, thereby enabling researchers to search for relations that have been undetected thus far. Third, machine learning entails advanced optimization techniques such as regularization to avoid overfit.

To the best of our knowledge, earnings prediction research has predominantly been focused on the first dimension so far. Put differently, researchers have proposed the use of different types of machine learning algorithms to approximate the possibly complex functional form that relates predictors and future earnings (e.g., Cao and You, 2021). It is conceivable that in general, more complex machine learning models lead to more accurate earnings predictions. In fact, Kelly et al. (2022) prove that increasing complexity, i.e., the ratio of model parameters to data, is always

[1] Hereafter, we refer to these models as traditional models.

beneficial in terms of out-of-sample prediction performance for return prediction. They explicitly recommend to use rich non-linear model specifications rather than simple linear ones. We argue that further research along this dimension thus bears little additional insights. Moreover, extant studies on earnings prediction typically only assess a rather limited set of predictor variables. An important exception is the study by Chen et al. (2022) which exploits the entirety of Extensible Business Reporting Language (XBRL) data. Yet, they restrict their analysis to the prediction of earnings changes and do not predict earnings per se.[2] The fact that machine learning allows for the use of large conditioning information sets as mentioned by Israel et al. (2020) is rather unexploited in this context thus far.

We fill this gap in the research and predict annual earnings per share conditional on a comprehensive set of variables, i.e., the entire set of financial statement variables from Compustat. This allows us to thoroughly analyze how fundamental accounting-based information drives and relates to future earnings. To do so, we use a selection of prominent, flexible machine learning models, namely a random forest model (RF), a gradient boosted tree model (GBT), a gradient boosted tree model with dropout (DART), a feed-forward neural network (NN) and an ensemble of the aforementioned models (ENML).[3] We further show how our approach relates to the most widely used traditional linear approaches, namely a simple model only including earnings as a predictor (L) (Gerakos and Gramacy, 2012), the HVZ-model (Hou et al., 2012), the EP-model (Li and Mohanram, 2014), the RI-model (Li and Mohanram, 2014) and an ensemble of the aforementioned models (ENTD).

Importantly, we also provide extensive model interpretation. The ability to understand the inner workings of prediction models is a fundamental requirement in most asset management applications (Israel et al., 2020). However, due to their complexity, machine learning models are hard to interpret. In the machine learning earnings prediction case, model interpretation has thus far been restricted to metrics of variable importance and partial dependencies. More so, these metrics have usually been applied to a predetermined, restrictive set of predictor variables as mentioned above. Put differently, researchers typically choose or construct a set of predictor

[2] In fact, as an extension to their main analysis, they also predict earnings levels. However, they only use a set of 24 variables for this and not the high-dimensional XBRL data set which they use for their main analysis. Surprisingly, their machine learning approach does worse than a simple random walk model. They conclude that earnings levels are hard to predict and use this as an argument for the fact that they focus on earnings changes in their primary analysis.

[3] The model selection is based on Bali et al. (2023).

variables that they deem important before estimating the model. After estimating the models, they derive the extent to which variables from this predetermined set contribute to the predictions and assess the partial dependencies of predicted earnings in regards to them.

This study aims to broaden the limited scope of model interpretation found in the existing literature on earnings forecasts: first, we derive the relative importance of variables using SHAP (SHapley Additive exPlanations) values, an approach based on cooperative game theory (Lundberg and Lee, 2017).[4] Since we do not select or construct variables beforehand as done in comparable studies (e.g., Hansen and Thimsen, 2020; Cao and You, 2021), we are able to holistically infer which out of all the financial statement variables are important from a statistical perspective. We also derive the relative importance of different groups of financial statement variables, such as cash flow statement (CF/S) variables, income statement (I/S) variables and balance sheet (B/S) variables. Importantly, we conduct this analysis for forecast horizons of up to five years. This allows us to assess how different (groups of) variables might vary in terms of their predictive power, depending on the forecast horizon considered. Second, we analyze non-linearity in the context of earnings prediction. In addition to partial dependencies that extant studies focus on in that context (e.g., Cao and You, 2021; Chen et al., 2022), we infer the degree to which different types of non-linearity play a role. More precisely, we show the degree to which interaction effects across financial statement variables and other forms of non-linearities, i.e., non-linearity of the functional form, are important by means of surrogate modeling. This approach is completely transparent, intuitive and easy to replicate, irrespective of the model or software used. Again, we conduct these analyses for forecast horizons of up to five years.

Our results can be summarized as follows: first, we find that generally, ensembles of models perform significantly better than their component models. This holds true both for traditional linear approaches and the machine learning approaches considered. For one-year-ahead predictions, for example, the machine learning ensemble yields around 2% more accurate predictions than the best performing component machine learning model.

In terms of bias, the traditional and the machine learning approaches yield very similar results.

[4]SHAP values are a way of explaining the results of any machine learning model. They are based on a game-theoretic approach that measures the contribution of each player to the final outcome. In machine learning, each variable is assigned an importance value that represents its contribution to the model's outcome (Lundberg and Lee, 2017).

Both types of models yield mostly unbiased predictions for forecasts horizons of up to three years. For forecasts horizons of four and five years, the models considered begin to systematically overestimate earnings. However, the machine learning models are consistently less biased in terms of levels as compared to their traditional analogues.

Mirroring previous findings on machine learning earnings predictions, we further show that our machine learning approaches constantly outperform traditional linear approaches in terms of accuracy (e.g., Cao and You, 2021). For the one-year forecast horizon, the best performing machine learning model, i.e., the ENML is around 12% more accurate than the best performing traditional model, i.e., the ENTD. Even for long forecast horizons of five years, we find that the ENML beats the ENTD in terms of accuracy by around 7%. The superiority in terms of accuracy holds similarly for both small and large firms. Furthermore, assessing accuracy differences across out-of-sample periods shows that model performance converges in the periods following the financial crisis, and diverges in favor of the machine learning approaches afterwards.

Assessing the degree to which the models are able to explain out-of-sample variation in earnings makes an even more convincing case for the ML approaches. In fact, the average out-of-sample R^2 (OOS R^2) of the ENML is between 14% and 28% higher than that of the ENTD for the forecast horizons considered.

Lastly, we show that the more accurate predictions of the machine learning approach translate into more profitable portfolios based on ICC. Buy-and-hold returns of long-short portfolios sorted on ICC based on the best-performing machine learning model (ENML-ICC) surpass the returns of long-short portfolios sorted on ICC based on the best-performing traditional model (ENTD-ICC). Moreover, while returns stay statistically significant for portfolios based on ENML-ICC for each rebalancing frequency considered, they are only statistically significant for one of three rebalancing frequencies considered in the ENTD-ICC case.

Turning to model interpretation, we find that current I/S variables, especially current earnings, are the most important group of predictors. With increasing forecast horizon, however, variable importance becomes more balanced among financial statement types. For one-year-ahead predictions, I/S variables contribute around 65% while B/S variables contribute around 20% of total importance. Total importance of these two groups of variables consistently converges with increasing forecast horizon. More precisely, for $t + 5$-predictions, I/S variables contribute around

48% while B/S variables contribute around 37% to total importance. CF/S variables consistently contribute around 15% to total importance throughout the forecast horizons considered. Put differently, the longer the forecast horizon, the more important B/S information. Further disentangling the effects of different components of financial statement information reveals that certain pieces of financial statement information dominate others. For example, debt and supplemental information resemble the most important pieces of B/S information. Turning to the CF/S, we find that variables related to the operating cash flow are much more relevant than variables related to either the investing cash flow or the financing cash flow. Lastly, turning to the I/S, we find that especially the EBIT and the net income are important, contributing around 14% and 31% to total importance for one-year-ahead forecasts, respectively. Interestingly, however, the importance of net income consistently declines with increasing forecast horizon to around 16% for five-years-ahead forecasts, whereas the importance of the EBIT stays constant. This suggests that information which is less exposed to accounting manipulation gains relevance for longer-term forecasts.

We further show that for one-year-ahead predictions, a linear surrogate model is able to explain around 80-90% of the earnings predictions across out-of-sample periods. In contrast to Jones et al. (2023), we find that interactions across financial statement variables are irrelevant.[5] We attribute the remaining 10-20% of unexplained variation to other types of non-linearities, which are not captured by interactions, i.e., non-linearity of the functional form. As the forecast horizon increases, the linear surrogate model approximates the relationship between predictions and inputs slightly worse. Interestingly, interaction effects across financial statement variables stay irrelevant for all forecast horizons.

The remainder of this study is structured as follows: In Section 2 we outline the relevant literature that we are contributing to. In Section 3 we describe our empirical approach. We evaluate and compare our approach in Section 4. In Section 5 we provide extensive interpretation. Finally, Section 6 concludes the study.

[5]This comes with a word of caution, as our input variables differ from theirs.

2 Related Literature

Our work relates to three strands of literature in particular. First, we contribute to the literature on machine learning applications in finance. Machine learning methods have become the prevalent way of conducting prediction exercises, primarily due to their superiority in terms of flexibility as compared to traditional econometric methods and their efficacy in regards to large sets of input data (e.g., Israel et al., 2020; Kelly et al., 2022). For example, Gu et al. (2020) show how different machine learning approaches perform in terms of predicting stock returns and Bali et al. (2023) apply machine learning to the task of predicting option returns. Our study is similar, in the sense that we predict another financial variable, i.e., earnings, with machine learning.

Second, we apply state-of-the-art techniques to interpret our machine learning predictions, thereby explicitly responding to the "need for interpretability" of financial machine learning models as formulated by Israel et al. (2020). We hence also contribute to the literature that aims to foster transparency and understanding of machine learning methods for prediction in finance research, such as e.g., Bali et al. (2023) who extensively assess machine learning option return predictions.

Third, we contribute to the literature on model-based earnings forecasts. Traditionally, researchers have suggested predicting earnings using time-series regression models (e.g., Ball and Watts, 1972; Albrecht et al., 1977; Watts and Leftwich, 1977), cross-sectional regression models (e.g., Hou et al., 2012; Li and Mohanram, 2014; Harris and Wang, 2019) or even simple random walk models (e.g., Li and Mohanram, 2014). More recently, however, several machine learning earnings prediction approaches have been implemented in the research (e.g., Hansen and Thimsen, 2020; Cao and You, 2021; Chen et al., 2022; Hendriock, 2022; Campbell et al., 2023; Jones et al., 2023; Van Binsbergen et al., 2023). However, the extant literature differs from our study in several key aspects which are outlined in the following.

Hansen and Thimsen (2020) also estimate a range of machine learning methods. They use a more high-dimensional input vector than other studies on predicting level earnings. However, it is still restrictive in the sense that it is based on prior research and not as high-dimensional as our data. Moreover, in contrast to our study, no model interpretation is provided.

The study of Cao and You (2021) also encompasses different machine learning models. How-

ever, they use a more restrictive set of input variables than us and provide only limited model interpretation. Moreover, Cao and You (2021) validate their models using traditional cross-validation (and a limited hyperparameter space). This, however, destroys the temporal structure of the observations and introduces information leakage (Gu et al., 2020). We preserve the temporal ordering of the observations by using fixed training, validation and test intervals.

Chen et al. (2022) use a single model (gradient boosted trees) as opposed to our multi-model approach. Furthermore, they predict binary earnings changes, while we focus on predicting level earnings.

Hendriock (2022) suggests predicting earnings by predicting the complete conditional density function. However, he restricts his input variable space to the one as defined by traditional linear models. Moreover, he does not provide model interpretation.

Campbell et al. (2023) benchmark an extensive range of machine learning model specifications with the aim of identifying the ones which compare the best to analyst forecasts. Apart from the fact that their model choices differ from ours, they use a different, smaller set of inputs, i.e., the Wharton Research Data Services (WRDS) Financial Suite Ratios extended by some additional variables like e.g., the stock return. Furthermore, they provide limited model interpretation as their primary focus is on the aforementioned horse-race between model specifications.

Jones et al. (2023) use a single model, i.e., a gradient boosted tree model algorithm, as opposed to our ensemble approach. In contrast to our study, their target variable is return on net operating assets and they use a set of six ratios as their predictors. Another crucial difference is that Jones et al. (2023) exclusively forecast earnings (changes) in $t+1$, while we forecast earnings (per share) for horizons $t+1$ to $t+5$. Finally, to the best of our knowledge, as the only other study in this context, they assess the impact of interactions. However, their method of doing so relies on the proprietary software that they use and is hence neither easily replicable nor transparent. Our surrogate modeling approach, in contrast, is easily applicable to any type of model, in any software and also allows us to explicitly determine the effect of interaction effects and non-linearity in parameters. Interestingly, our findings in regards to interactions differ strongly from the ones provided by Jones et al. (2023). They find substantial importance of interactions, while we find that interactions among financial statement variables are irrelevant.

In contrast to our study, Van Binsbergen et al. (2023) use a random forest model to predict

earnings conditional on financial ratios, similar to Campbell et al. (2023). In contrast, our approach involves estimating a spectrum of machine learning models individually and in ensemble configurations. Moreover, we utilize an entirely distinct set of input data, specifically the comprehensive Compustat financial statement dataset. Finally, unlike their study, which primarily assesses analyst biases, our analysis encompasses detailed explanations of model predictions.

Summing up, to the best of our knowledge, we are the first to predict level earnings per share for forecast horizons of up to five years using the entirety of available Compustat financial statement variables. Furthermore, we contribute novel guidance for future research on earnings (per share) predictions by thoroughly interpreting our state-of-the-art machine learning approaches using model agnostic and easily applicable methods, something that has been not done extensively thus far.

3 Empirical Approach

3.1 General Setup

We express earnings E of firm i in period $t + \tau$ as the expectation in period t plus an error ϵ:

$$E_{i,t+\tau} = \mathbb{E}_t[E_{i,t+\tau}] + \epsilon_{i,t+\tau}. \tag{1}$$

Every model that aims to predict earnings can be interpreted as an attempt to derive $\mathbb{E}_t[E_{i,t+\tau}]$, i.e., the expectation. More precisely, we assume that expected earnings in $t + \tau$ are a function of a vector of inputs $\boldsymbol{X}$ of firm i known at time t:

$$\mathbb{E}_t[E_{i,t+\tau}] = f(\boldsymbol{X}_{i,t}). \tag{2}$$

It becomes evident that modelling expected earnings consists of three crucial parts. First, one has to determine which inputs enter the model, that is, how $\boldsymbol{X}$ is defined. Second, one has to decide which functional form $f(\cdot)$ takes on. This corresponds to the decision about which empirical model to choose from. Lastly, one has to decide on how to estimate $f(\cdot)$, i.e., which statistical loss function to minimize. The latter is not the explicit focus of this study and thus we keep it simple. We follow the original implementations of the traditional models and estimate

them using the mean squared error (MSE). In case of the machine learning models, we follow Gu et al. (2020) and estimate the models both using the MSE and the mean absolute error (MAE) and report the predictions based on the loss function that leads to more accurate forecasts according to the price scaled absolute forecast error (PAFE) at the 1-year horizon.[6]

Thus, apart from the decision regarding the loss-function, the two contrasting *extreme* approaches to predicting earnings are: (1) actively making the decision which variables and which functional form to assume ex ante, and (2) letting the data speak by selecting a model that permits flexible functional forms and providing it with the entire dataset available. The former corresponds to the traditional prediction approaches suggested by e.g., Hou et al. (2012). To the best of our knowledge, the latter has not yet been implemented for earnings per se, a noteworthy exception being Chen et al. (2022) who employ this approach for (binary) earnings *changes*. There are approaches that fall in between (1) and (2). In these approaches, either the functional form or the input vector is restricted significantly (e.g., Hendriock, 2022; Van Binsbergen et al., 2023).

Our study focuses on assessing the second, flexible approach (2) and comparing it to traditional approaches (1). We more thoroughly elaborate on the choice of the input vector and of the model in 3.2 and 3.3, respectively.

3.2 Data

US annual financial statement data is obtained from Compustat. Our sample period ranges from 1988 to 2021. This is due to the fact that CF/S-data is only sparsely available prior to 1988. To conduct the ICC portfolio evaluation, we add price and return data from CRSP to the Compustat data used for model estimation. We exclude companies from the financial services industry (as defined by Compustat). Moreover, we drop observations with missing prices, prices smaller than 1$, missing common shares outstanding or missing earnings. This results in a final sample for model estimation that consists of 191,273 observations.

For our machine learning models, we use the Compustat financial statement items as predictors. We drop variables with more than 50% of observations missing or no observations in any of the cross-sections (i.e., estimation years), yielding 192 variables.[7] An overview over these variables

[6]To be precise, Gu et al. (2020) choose either the MSE or the Huber loss, depending on which performs better. We choose either the MSE or the MAE, depending on which performs better.

[7]We also drop variables already scaled by shares. The reason for that is, that we scale all our variables by shares as

is given in Table B.2 in the Appendix. Analogous to Chen et al. (2022), we include lags and first-order differences of these variables, resulting in a set of 576 predictor variables in total. For the traditional models, we construct input variables according to the respective models. An overview over these variables is given in Table A.1 in the Appendix. All our variables, including our target variable, are scaled by common shares outstanding and winsorized at the 1%- and 99%-level, respectively. Finally, since neural networks are sensitive to scale differences across input variables, we standardize each variable.

3.3 Models

We estimate two groups of models. The first group consists of popular simple linear models that have been introduced in the literature thus far. All of these models assume a linear additive relation between earnings and some low-dimensional input vector, i.e.,

$$\mathbb{E}_t[E_{i,t+\tau}] = \boldsymbol{\beta} \boldsymbol{X_{i,t}}, \tag{3}$$

where $\boldsymbol{\beta}$ denotes a vector of coefficients. The models differ in terms of which variables the input vector consists of. A more detailed description is given in Appendix A.1. A difference to be noted is that the models also slightly differ in how they define the output. While the RI and the EP model use earnings per share, the HVZ model uses earnings. In this study, we define earnings as "income before extraordinary items" (Compustat variable: ib). As mentioned above, we consistently scale our output as well as our input variables by common shares outstanding in all of the models. Our forecast horizon, both for this and the following group of models spans from $\tau = 1$ to $\tau = 5$.

The second group of models consists of flexible models that are able to approximate arbitrary complex functional forms. Put differently, we do not assume any specific type of functional form when estimating these models. Analogous to Bali et al. (2023), we estimate a random forest model (RF), a gradient boosted tree model with (DART) and without dropout (GBT) and a neural net (NN). Importantly, we try to restrict our input vector as little as possible. Specifically, we feed the models the 576 variables as outlined above, including the lags and first-level differences. We argue

mentioned below and hence these variables are redundant. However, this only pertains to five variables in our study.

that this input vector corresponds to a proxy for the entirety of (relevant) financial statement input variables. Note that extending the input vector even further implies more computational effort.[8]

Since studies in other realms of financial forecasting have shown that using an ensemble of models may prove superior to using single models (e.g., Bali et al., 2023), we derive the equally weighted average prediction for both groups of models, i.e., we derive two ensemble model predictions:

$$\mathbb{E}_t^{(En)}[E_{i,t+\tau}] = \frac{1}{J}\Sigma_{j=1}^{J}\mathbb{E}_t^{(j)}[E_{i,t+\tau}], \tag{4}$$

where $j \in J$ denotes the respective single model and En denotes the respective ensemble model.

The ensemble of the traditional models is denoted by *ENTD* and the ensemble of the fully flexible machine learning prediction models is denoted by *ENML*.

3.4 Out-of-sample approach

We employ a rolling window strategy to obtain our out-of-sample prediction results. Specifically, for the machine learning models, we divide our data into training, validation and test sets. For each forecast horizon τ, the process for generating forecasts as of t proceeds as follows: we train our models using earnings from $t-11$ to $t-2$ as output and corresponding financial statement data lagged by τ as predictors. Next, we tune the machine learning models using earnings from $t-1$ to t as output and lagged financial statement data by τ as predictors. Tuning involves determining the optimal hyperparameter values for the model, as detailed in Table B.1 in the Appendix. Subsequently, earnings predictions for $t+\tau$ are derived by inputting variables from t into the optimized models. This process is repeated recursively, advancing one year at a time. We follow the approach of Hou et al. (2012) and Li and Mohanram (2014), estimating models at the end of June each year, under the assumption of a reporting lag of three to fourteen months for financial statements.[9]

[8]We could have also included other variables like price, analyst forecasts, etc. (e.g., Campbell et al., 2023; Van Binsbergen et al., 2023). However, the focus of our study is the thorough analysis of the relationship between fundamental accounting-based information and future earnings.

[9]Specifically, data from April of year $t-1$ to March of year t is considered the most recent fiscal year-end data available as of June in year t, capturing the information as of t.

Traditional linear approaches, on the other hand, do not necessitate a tuning window. Hence, we only partition the data into training and test sets when estimating these models. The subsequent steps of the procedure remain unchanged. More precisely, for each forecast horizon τ, models are trained using earnings from $t-11$ to t as output and corresponding lagged financial statement data by τ as predictors. Earnings predictions for $t+\tau$ are then derived by utilizing predictor variables from t.

3.5 Evaluation

We evaluate the predictive performance of the models across a range of evaluation metrics. First, we compute the error metrics that are common in the earnings prediction literature (e.g., Hou et al., 2012). For each forecast horizon $\tau \in [1, 2, 3, 4, 5]$, these include the price scaled forecast error (PFE) or bias:

$$PFE_{i,t+\tau} = \frac{E_{i,t+\tau} - \hat{E}_{i,t+\tau}}{Price_{i,t}}, \tag{5}$$

and the price scaled absolute forecast error (PAFE) or accuracy:

$$PAFE_{i,t+\tau} = \frac{|E_{i,t+\tau} - \hat{E}_{i,t+\tau}|}{Price_{i,t}}, \tag{6}$$

where $E_{i,t+\tau}$ denotes actual earnings for firm i in period $t+\tau$, $\hat{E}_{i,t+\tau}$ denotes the respective forecast and $Price_{i,t}$ is the firm's stock price at the end of June in the respective estimation year.

Second, analogous to Gu et al. (2020), we make pairwise comparisons of the individual as well as the ensemble models using an adjusted version of the Diebold and Mariano (1995) procedure. This procedure allows for a quantitative comparison of the different forecasts. More precisely, we compare the forecast performance of model (1) and (2) using the test statistic $DM = \bar{d}/\hat{\sigma}_{\bar{d}}$, where

$$d_{t+\tau} = \frac{1}{n_{t+\tau}} \Sigma_{i=1}^{n_{t+\tau}} ((PAFE_{i,t+\tau}^{(1)} - PAFE_{i,t+\tau}^{(2)})^2), \tag{7}$$

with $n_{t+\tau}$ being the number of firms in the out-of-sample period $t+\tau$. $\bar{d}$ and $\hat{\sigma}_{\bar{d}}$ then denote the mean and the Newey-West adjusted standard error (Newey and West, 1987) of $d_{t+\tau}$ over the out-of-sample periods.

Third, we assess the out-of-sample R^2 ($OOSR^2$) of each individual as well as the ensemble forecasts, i.e., for every out-of-sample period we calculate

$$OOS\, R^2_{t+\tau} = 1 - \frac{\Sigma_{i=1}^{n_{t+\tau}} (E_{i,t+\tau} - \hat{E}_{i,t+\tau})^2}{\Sigma_{i=1}^{n_{t+\tau}} (E_{i,t+\tau} - \bar{E}_{t+\tau})^2}, \tag{8}$$

for each model. Here, $\bar{E}_{i,t+\tau}$ denotes average earnings of firms in period $t + \tau$. Albeit not commonly used in the earnings prediction literature (e.g., Hendriock (2022) being an exception), this evaluation metric is of particular importance for the typical use case of earnings predictions, i.e., long-short ICC portfolios. In this context, predicting cross-sectional variation is much more important than accurately predicting earnings per se, since an investor goes long (short) the stocks which's ICC is high (low) in cross-sectional comparison.

Lastly, we derive the ICC based on the two ensemble forecasts. We follow the literature and calculate ICC following the methods of Gordon and Gordon (1997), Claus and Thomas (2001), Gebhardt et al. (2001), Easton (2004) and Ohlson and Juettner-Nauroth (2005).[10] More precisely, our ICC estimates are derived as the average of the five aforementioned methods using both the traditional ensemble and the machine learning ensemble earnings predictions. We then construct long-short zero investment portfolios based on ICC and assess their average performance across the out-of-sample periods.

3.6 Interpretation

A primary contribution of this study is the comprehensive interpretation of the machine learning approach, addressing a key issue in machine learning applications in finance and accounting (Israel et al., 2020). Our study fills a gap in the existing literature, which either lacks model interpretation entirely or provides only limited insights, as discussed above. More precisely, we derive the variable importance and the degree to which different types of non-linearity play a role for our best performing machine learning model, i.e., the machine learning ensemble.

Variable Importance

We determine the importance of the different variables with respect to the earnings prediction.

[10] A description of the models is provided in Appendix C.

To do so, we compute their SHAP values, a state-of-the-art approach for assessing the importance of input variables which is based on cooperative game theory (Lundberg and Lee, 2017). In essence, SHAP values approximate how a model's prediction changes when knowing the value of a respective input variable. The approach is model-agnostic and allows us to evaluate the importance of input variables irrespective of the model used. We conduct these analyses at both the individual variable level and the grouped-variable level. Specifically, we determine the relative importance of predictors grouped into balance sheet, cash flow statement, and income statement data as well as predictors grouped into current, lagged, and difference variables. Furthermore, we provide an in-depth accounting perspective on which specific types of financial statement information are important by breaking the financial statements down into schematic components. Importantly, we conduct these analyses per forecast horizon.

Non-linearity

In addition to deriving the variable importance, we also evaluate the extent to which non-linearity plays a role in our machine learning model.

First, we assess the degree to which non-linearity in terms of variables and non-linearity in terms of functional form play a role. In a first step, we regress predicted earnings on the 50 most important input variables using a linear Lasso-penalized regression model.[11] We add a Lasso-term to control for the overfit that would otherwise be induced by the large input vector. In a second step, we add all possible two-way interactions to the surrogate regression model from the prior step. Assuming there is some degree of non-linearity present in the fully flexible model, this allows us to disentangle the degree to which non-linearity in terms of variables (i.e., interactions across inputs) and non-linearity in terms of functional form play a role. More specifically, we assess the adjusted in-sample R^2s of the two surrogate models. The R^2 of the first-step surrogate model indicates the degree to which the predictions are linear. The difference between the R^2 of the first-step surrogate model and the second-step surrogate model indicates the degree to which non-linearity in terms of variables, i.e., (two-way) interactions across financial statement variables, plays a role. We attribute the portion that remains unexplained by the second-step surrogate

[11] We only use the 50 most important variables, because otherwise, including all possible two-way interactions in the second step requires an excessive amount of computing power.

model to non-linearity in functional form.[12]

Second, we assess the partial dependence of earnings with respect to the most important input variables. As typically done in the literature, we evaluate the partial dependencies graphically via so-called partial dependence plots. To do so, we fit a non-parametric lowess model (locally weighted linear regression) to the SHAP values of a predictor value of interest (the output) and the associated predictor values (the input) and plot the result. This allows us to approximate the effect of the respective predictor variable on future earnings.

Again, we conduct the surrogate modeling and the partial dependence analyses per forecast horizon.

4 Evaluation

4.1 Accuracy and Bias

Chen et al. (2022) report that their machine learning approach does worse than a simple random walk type model when predicting level earnings. They explicitly state that level earnings are hard to predict and thus resort to the prediction of earnings changes in their main analysis. In contrast, mirroring findings by e.g., Cao and You (2021) our flexible machine learning models for earnings (per share) level prediction outperform the traditional linear models by a significant margin.

Price Scaled Forecast Error

Table 1 shows the time-series averages of the median PFE for the four traditional, the four machine learning and the two ensemble models. The PFE provides insight into whether the estimated earnings are systematically over- or underestimated (biased) relative to actual earnings.

For forecast horizons $t+1$-$t+3$ neither the traditional ensemble nor the machine learning ensemble yield statistically significant PFEs. Put differently, neither of the two ensemble approaches systematically over- or underestimates earnings for these forecast horizons. This is in line with extant studies on earnings prediction, which report that earnings predictions by statistical models

[12]Theoretically, the unexplained portion also includes effects of interaction terms of order three and higher. However, we assume that these can be neglected and find evidence for this assumption in undocumented analyses.

do not exhibit biases as opposed to those by analysts (Hou et al., 2012). Nevertheless, the ENML model has a lower bias compared to the ENTD (0.0003 vs 0.0019 for $t+1$ predictions). The traditional ensemble systematically overestimates earnings in $t+4$ and $t+5$. The same is true for the machine learning ensemble, but the bias is again 41% lower for $t+4$ predictions (-0.0082 vs -0.0138) and 37% lower for $t+5$ predictions (-0.0146 vs -0.0230).

The results further show that at the non-ensemble level across horizons $t+1$ to $t+4$, some models yield small but statistically significant PFEs, with the most biased machine learning model for $t+1$ being the RF model (0.0045) and the most biased traditional model for $t+1$ being the RI model (0.0047). For $t+5$, all non-ensemble models yield PFEs that are statistically significantly different from zero. However, all non-ensemble machine learning models score lower PFEs than their traditional counterparts.

Overall, the machine learning models generally exhibit lower bias. Moreover, earnings prediction models appear to systematically overestimate earnings as the forecast horizon increases. This effect holds, irrespective of whether we predict earnings using traditional linear or machine learning methods.

[TABLE 1 ABOUT HERE]

As a robustness check, we stratify our predictions by size and report the resulting PFEs in Table 2. More precisely, at every prediction date t, we split the sample into two equally sized groups based on their market capitalization as of t. This allows us to separately assess PFEs for small and large firms, respectively.

For large firms, we find that the majority of the traditional models significantly underestimate the actual earnings for forecast horizons $t+1$-$t+2$, while the biases of the machine learning models are smaller and, with the exception of the RF model, not significantly different from zero for these horizons. For forecast horizons $t+3$-$t+4$, no model is significantly biased. For predictions in $t+5$, some of the non-ensemble models significantly overestimate earnings. Similar to the full sample case, both ensemble models somewhat equally overestimate actual earnings in $t+5$ (-0.0096 vs -0.0079). However, both PFEs are not significantly different from zero.

Turning to small firms, the traditional ensemble model overestimates actual earnings with

increasing bias (-0.0045 to -0.0449 from $t+1$ to $t+5$). The bias is statistically significant for horizons $t+2$ to $t+5$. The machine learning ensemble also overestimates actual earnings, but the biases are much smaller (-0.0019 to -0.0255 from $t+1$ to $t+5$). The biases are statistically significant for horizons $t+3$ to $t+5$. We conclude that the machine learning ensemble performs much better than its linear analogue for small firms specifically.

[TABLE 2 ABOUT HERE]

Price Scaled Absolute Forecast Error

Turning to the next evaluation metric, Table 3 reports the time-series averages of the median PAFEs for the four traditional, the four machine learning and the two ensemble models. The PAFE is a measure for the accuracy of a model, with values closer to zero indicating higher accuracy.

First, we observe a positive effect of model stacking. Overall, among the traditional models, the best-performing one is the traditional ensemble, while among the machine learning models, the best-performing one is the machine learning ensemble. Specifically, the machine learning ensemble outperforms each of its individual component models for every forecast horizon. The same is true for the traditional ensemble for the horizons $t+1$ to $t+2$. Especially in the traditional case, this result is surprising, since the models differ very little in terms of the predictor variables. For forecast horizons $t+3$ to $t+5$, the traditional ensemble performs slightly worse than the best performing traditional component model (0.0406 vs 0.0403 in $t+3$, 0.0467 vs 0.0453 in $t+4$ and 0.0536 vs 0.0517 in $t+5$). Second, we find that in most cases, all machine learning models, including the ensemble, outperform all traditional models, including the ensemble, for all prediction horizons. However, for forecast horizons $t+3$ to $t+5$, the RI model is more accurate than some of the machine learning component models. This stresses the benefit of model averaging for the non-linear machine learning models specifically. The difference in accuracy between the machine learning and the linear ensemble is at a statistically significant level between -0.0022 and -0.0038. This translates into a relative difference of 11.70% for earnings in $t+1$ to 7.09% for earnings in $t+5$. The ENML thus provides not only statistically significant, but also

economically meaningful gains in accuracy over the traditional models.

[TABLE 3 ABOUT HERE]

Figure 1 illustrates the median PAFE of the ENML and ENTD for each out-of-sample year and for forecast horizons of $t+1$ and $t+5$, respectively. The plots illustrate that the overall PAFE levels strongly increase with increasing forecast horizon. Furthermore, they reveal that the machine learning ensemble is more accurate in all years and for both forecast horizons, except for 2009, in which the ENTD yields slightly more accurate $t+5$ forecasts than the ENML. In general, the difference between the ENML-PAFE and the ENTD-PAFE varies across out-of-sample periods.

[FIGURE 1 ABOUT HERE]

As a robustness check, we again stratify our predictions into those for small firms and those for large firms and report the results in Table 4. Consistent with the existing literature, we find that accuracy is generally much higher for large firms than for small firms (Li and Mohanram, 2014). This holds true for all models considered. Furthermore, the accuracy superiority of the machine learning ensemble over its traditional analogue is more pronounced for the small firm sample. For small firms, the accuracy difference ranges from 14.38% for $t+1$ predictions to 11.88% for $t+5$ predictions. Nonetheless, for large firms, the difference still ranges from 13.16% for $t+1$ predictions to 2.65% for $t+5$ predictions. Moreover, except for large firms in $t+5$, the difference in accuracy is highly significant for both firm samples and for all forecast horizons. Lastly, similar to the full sample case, we again observe that the traditional ensemble performs slightly worse than some of its component models for large forecast horizons.

[TABLE 4 ABOUT HERE]

Diebold and Mariano Forecast Comparison

Table 5 shows the pairwise comparisons of the models using the aforementioned modified

Diebold and Mariano (1995) test statistic. We restrict this analysis to predictions for earnings in $t+1$ and exclusively use the PAFE. This is because pairwise comparisons using the PFE do not yield interpretable results, as a higher (lower) PFE of a model compared with another model may indicate both better or worse performance of the respective model. A positive statistic indicates the column model outperforms the respective row model.

The results confirm the finding that the machine learning approaches outperform their traditional counterparts. Moreover, the ENML is again the best performing model, beating every other one except for the DART model.[13]

[TABLE 5 ABOUT HERE]

4.2 Out-of-sample R^2

The next metric considered is the out-of-sample R^2 ($OOS\ R^2$). The results are reported in Table 6. The $OOS\ R^2$ allows us to assess how the models perform in terms of explaining out-of-sample variation in future earnings.

As expected, the $OOS\ R^2$ decreases with increasing forecast horizon. Moreover, the results generally confirm the positive effect of model stacking. Our analysis demonstrates that the machine learning ensemble consistently outperforms its individual components across all forecast horizons considered. In contrast, there are instances in which the traditional ensemble exhibits slightly lower performance than the RI model for specific forecast horizons, i.e., $t+1$ and $t+2$. Nevertheless, the traditional ensemble consistently outperforms all of its component models for the remaining forecast horizons.

Furthermore, our findings validate the notion that machine learning approaches surpass traditional linear models in terms of predictive performance. To be more specific, when we compare the $OOS\ R^2$ of the machine learning models, including the ensemble, against that of the traditional models, including the ensemble, we find that the machine learning models outperform the traditional ones in almost every case. In fact, just assessing the best-performing models, i.e., the ensembles, we find that the machine learning ensemble beats its traditional counterpart for

[13]Note that we adjust our p-values very conservatively via a multiple comparisons Bonferroni correction. This significantly increases the hurdles for reaching significance, hence possibly explains why the test statistic is not statistically significant.

every forecast horizon. The difference in *OOS* R^2 between the ensemble models is statistically significant at the 1% level for forecast horizons $t+1$ to $t+4$. Further, while the relative PAFE difference between the ensembles decreases with increasing forecast horizon, the difference in *OOS* R^2 increases from 14.03% for $t+1$ predictions to 18.72% for $t+5$ predictions, in relative terms.

[TABLE 6 ABOUT HERE]

Figure 2 plots the *OOS* R^2 for the ENML and ENTD for each year and for $t+1$ and $t+5$, respectively. Again, it is evident that the machine learning ensemble outperforms the traditional ensemble in the majority of years and for both forecast horizons. Remarkably, the *OOS* R^2 of both ensemble models exhibits a noticeable dip in 2009, particularly pronounced in forecasts for $t+5$. Notably, for $t+5$ forecasts, it takes some years for the *OOS* R^2 to rebound. This result underscores the delayed integration of new information, such as the financial crisis in this case, into longer-term forecasts. Such delayed adaptation is inherent in the rolling window approach we employ.

[FIGURE 2 ABOUT HERE]

Again, for robustness, we stratify our predictions into those for small firms and large firms, respectively and report the results in Table 7.[14] Consistent with our prior findings, we observe notable disparities in the *OOS* R^2 performance between large and small firms, with the former exhibiting higher predictive accuracy. Furthermore, our analysis demonstrates that the ENML model consistently outperforms the ENTD model across all forecast horizons and for both subsamples. In line with the overarching patterns evident in our prior results, we find that the differential in performance between the ENML and ENTD models is more pronounced for small firms. Specifically, our findings indicate that the ENML model exhibits a 15.16% higher *OOS* R^2 for large firms compared to the 23.37% higher *OOS* R^2 for small firms for $t+1$ predictions. This pronounced discrepancy reaffirms the consistency of our earlier observations and substantiates the

[14]Note that the *OOS* R^2s of both subsamples are lower than the total *OOS* R^2. This is because the *OOS* R^2 is a non-linear function.

argument that a traditional linear model is inadequate for capturing the intricate nuances in future earnings, particularly for smaller firms. Interestingly, the relative outperformance of the ENML in terms of *OOS* R^2 for small firms strictly increases with increasing forecast horizon to 35.00% for $t+5$. In the large firm sample, the relative outperformance of the ENML as compared to the ENTD also increases to 45.21% for $t+5$-predictions. Across horizons $t+1$-$t+4$, the difference between the ENML and the ENTD is statistically significant for both subsamples.

[TABLE 7 ABOUT HERE]

4.3 Implied Cost of Capital

A common application of earnings forecasts is the derivation of the implied cost of capital (ICC), for which earnings predictions serve as a crucial input. We restrict this analysis to the best performing traditional earnings forecast model, i.e., the traditional ensemble, as well as the best performing machine learning earnings forecast model, i.e., the machine learning ensemble.

As typically done in the literature, we evaluate the return expectations in form of ICC by evaluating the performance of long-short portfolios sorted on ICC (e.g., Hou et al., 2012; Li and Mohanram, 2014). More specifically, we sort the stocks into deciles according to ICC in t and assess the realized annual geometric average return of a buy-and-hold portfolio that goes long the highest ICC decile and short the lowest ICC decile in t. We consider three rebalancing scenarios, i.e., annually, every two years, and every three years.

Table 8 shows that long-short ICC portfolios based on the ENML predictions outperform those based on the ENTD predictions for all rebalancing frequencies considered. A buy-and-hold long-short ICC portfolio based on ENML predictions yields (geometric) average annual returns of 8.45%, 9.14% and 6.80% when rebalancing every year, every second year and every third year, respectively. The linear analogues yields corresponding (geometric) average returns of 7.15%, 7.75% and 5.89%. Moreover, while the portfolio returns based on the ENML model are statistically significant for every rebalancing frequency, the portfolio returns based on the ENTD model are only statistically significant when rebalancing the portfolio every three years.

We conclude that the improved accuracy of machine learning predictions translates into more

profitable investment strategies, thereby stressing the practical importance of earnings prediction accuracy.

[TABLE 8 ABOUT HERE]

5 Interpretation

5.1 Variable importance

As outlined above, we assess the degree to which financial statement variables matter in the machine learning ensemble by computing their SHAP values.[15] More precisely, we compute SHAP values per out-of-sample period and derive their respective averages for each variable.

Figure 3 shows the mean absolute SHAP values of the ENML, averaged over all out-of-sample periods and scaled so that variable importance per forecast horizon sums to one. We show the twenty most important variables for predicting earnings in $t+1$ and sort them according to their importance.[16] The higher the SHAP value, the more important the variable. For $t+1$ predictions, *ib*, i.e., current earnings, is the most important variable by far.[17] This comes as no surprise, considering that a simple model including only current earnings as a predictor performs comparably well in predicting future earnings.

A striking finding is that the remaining 19 most important variables are primarily different definitions of earnings. For example, the second most important variable *oiadp* resembles "operating income after depreciation" and the third most important variable *ibcom* resembles "earnings before extraordinary items - available for common". The only top-twenty variables that do not originate from the income statement are *oancf* and *seq*. However, they resemble comparably low importance.

In general, six to seven variables dominate across forecast horizons $t+1$ to $t+5$. Another finding regarding the different forecast horizons is that the significance of *ib* gradually diminishes with increasing horizon. Instead, *oiadp*, another earnings variable, emerges as the most important

[15] We focus on the ensemble model as it is the best performing machine learning approach.

[16] The variable definitions are provided in B.2

[17] Note that this is the earnings definition that we use as our target variable. Further note that all of our variables are scaled by common shares outstanding. Thus, strictly speaking, we refer to earnings per share when talking about *ib*.

variable. Moreover, the top-twenty variables not stemming from the income statement, i.e., *oancf* and *seq*, become increasingly important with increasing forecast horizon.

Lastly, the results suggest that current data is more important than lagged data or first order differences. We revisit this claim below.

[FIGURE 3 ABOUT HERE]

We now explore whether the most important predictor variables act as substitutes or complements. This assessment is conducted by examining the absolute Pearson correlation coefficients of the top-twenty variables reported in Figure 4. If the variables are substitutes, one would expect high coefficient values. In general, we find mixed results. *ibcom*, *pi*, *ibadj*, *ni* and *niadj* are correlated quite strongly with *ib* and each other. All of these are variations of earnings definitions which do not differ strongly from each other. This observation leads us to consider these variables as substitutes for *ib*, indicating that they do not necessarily possess significant independent predictive power. Other variables, that are not as closely related to *ib*, either because they explicitly exclude major income statement items, such as *oiadp*, or because they are not income statement items at all, such as *oancf*, do not correlate as strongly with earnings. We interpret this as evidence that these items are complements to *ib* and hence possess stand-alone predictive power.

[FIGURE 4 ABOUT HERE]

5.2 Group importance

Table 9 reports variable importance per group. More precisely, we group the variables according to the financial statement they originate from (Panel A), whether they are current, lagged or change information (Panel B), and according to the two aforementioned categories (Panel C). Grouping the variables according to the financial statement they originate from reveals that income statement (I/S) variables are the most important variables. On average, for one-year-ahead predictions, I/S variables contribute approximately 65% to the total importance, while balance sheet (B/S) variables and cash flow statement (CF/S) variables contribute around 20% and 15%, respectively.

This finding aligns with the analysis of the most important variables, indicating that I/S variables significantly outweigh others in importance for predicting earnings. However, we also find that I/S variables become less important with increasing forecast horizon. In fact, the importance of I/S variables decreases to around 47% for $t+5$ forecasts. In contrast, B/S variables become more important with increasing forecast horizon (around 37% for $t+5$ predictions) while CF/S variables stay at a constant level.

[TABLE 9 ABOUT HERE]

Grouping variables according to whether they are current, lagged or difference variables in Panel B reveals that current data is by far the most important group out of these categories, contributing around 71% of total importance for $t+1$ predictions. Lagged and difference data each contribute around $14-15\%$ to total importance for $t+1$ predictions. Again, importance becomes more evenly distributed among the groups with increasing forecast horizon. More precisely, current variables become less important while lagged variables become more important. This might be the case because short-term forecasts are heavily influenced by current information due to their sensitivity to recent developments. Longer-term forecasts benefit from a combination of current and lagged information to capture the interplay of short-term dynamics and longer-term trends.

Further breaking down the groups according to the two aforementioned categories stresses the findings above. Overall, current I/S variables contribute around 53% to total importance for $t+1$ forecasts and hence represent the most important group of variables by a significant margin. This is intuitive and supports the finding that simple earnings forecasts models only considering current earnings items, like the L model or the EP model, perform comparably well in predicting future earnings.

We conclude our variable importance assessment by more thoroughly analyzing the variable importance per financial statement type in Table 10. More precisely, we assess the importance of each financial statement type per schematic financial statement component, such as e.g., current assets, fixed assets or equity, in the B/S case. This analysis provides an intuitive accounting perspective on which components of financial statements are important and how the importance

might change across forecast horizons.

[TABLE 10 ABOUT HERE]

The table provides several key insights: first, assessing the B/S, we find that the debt and supplemental items are the most important pieces of B/S information, with both contributing around 5-6% to total importance for $t+1$ forecasts. Moreover, all pieces of B/S information consistently increase in importance with increasing forecast horizon.

Second, variables associated with the operating cash flow resemble the most important category of CF/S variables. This comes as no surprise, since the operating cash flow closely relates to earnings. The relevance of the investing cash flow slightly increases with increasing forecast horizon. However, overall, the differences are minor.

Third, different definitions of earnings, i.e., EBITDA, EBIT, EBT and net income, resemble the most important I/S categories. Out of these categories, EBIT and net income are the most important categories with around 14% and 31% share in total importance, respectively. Interestingly, the importance of EBIT stays somewhat constant throughout forecast horizons, whereas the importance of the net income consistently declines with increasing forecast horizon to around 16% for $t+5$ forecasts. This dynamic might be attributable to the fact that net income is more strongly exposed to accounting manipulation than EBIT and hence less reliable in the long-term. Moreover, we find that while sales contribute very little overall, they consistently increase in importance with increasing forecast horizon. This supports the notion that items which are less exposed to discretionary accounting gain predictive value when considering longer forecast horizons. Revisiting the aforementioned finding that operating cash flow variables maintain consistent importance across forecast horizons further reinforces this notion. Unlike earnings, cash flows include no discretionary accural items and are hence not exposed to earnings management (e.g., Jones, 1991). Consequently, their predictive value does not decrease for longer-term forecasts. In summary, these findings suggest that the variations in importance across forecast horizons are primarily driven by the presence of earnings management. Future research endeavors could offer additional insights into these dynamics.

5.3 Non-linearity

We now shed light on the degree to which non-linearity of the functional form and non-linearity of variables, i.e., interactions among financial statement variables considered, play a role in predicting earnings.

Surrogate model

We find that for our flexible ENML approach, around 89% of the variation in predicted earnings can be explained by a linear surrogate model for $t+1$ predictions on average. This does not change when including two-way interactions, suggesting that interaction effects play virtually no role in predicting earnings using raw financial input data. Interestingly, this stands in stark contrast to the results by Jones et al. (2023) who find that interactions among the ratios which they use as predictors contribute substantially to the prediction. Apart from the fact that our target variables differ, this might be attributable to the fact that we use a significantly larger set of inputs, which might directly capture interactions among variables of a more limited set of variables. The remaining unpredictable portion of the ENML can be attributed to non-linearity of the functional form.[18] We depict this graphically in Figure 5. The figure shows the in-sample R^2 per out-of-sample period, derived by regressing predicted earnings on a linear surrogate model and a linear surrogate model including two-way interactions. The figure also includes the surrogate models for predictions for $t+5$. In fact, assessing the surrogate models for $t+5$ reveals that even for longer forecast horizons, interaction effects across financial statement variables do not play a role. However, with increasing horizon, a slightly larger portion of the earnings-predictor relation can be attributed to non-linear functional form. Nonetheless, the linear surrogate model still explains around 87% of the predictions in $t+5$ on average.

[FIGURE 5 ABOUT HERE]

Partial dependence plots

We now turn to how the aforementioned degree of non-linearity is expressed at the variable-

[18] Theoretically, it can also be attributed to higher-order interactions. However, in undocumented results, we find that this is not the case.

level. Figure 6 shows the partial dependence plots for *ib*, *oiadp*, *ibcom* and *oancf* for all forecast horizons.[19] The partial dependence measures the sensitivity of the predicted earnings to the individual financial statement variables.

The upper-left panel shows that the *ib* effect is the strongest, i.e., the steepest. Remarkably, the sensitivity appears to be linear for both positive and negative values of *ib*. However, there is a distinction in the slope of the line for positive and negative values, suggesting varying sensitivities of future earnings to current earnings for profit and loss firms, respectively. This may explain why the EP model and the RI model by Li and Mohanram (2014) yield comparably good forecasting results, especially for short forecast horizons. The two models include a dummy for negative earnings and the interaction between earnings and the negative earnings dummy, which essentially allows for different slopes of *ib* for profit and loss firms.

For *ibcom* in the lower-left panel we find a similar trend as for *ib*. In contrast, *oiadp* and *oancf* are essentially linear across all forecast horizons (with the exception of *oancf* for $t+1$ predictions).

[FIGURE 6 ABOUT HERE]

6 Conclusion

We show that earnings per share predictions based on state-of-the-art machine learning approaches using high-dimensional financial statement data are more accurate than those based on traditional linear approaches. These improvements hold across all evaluation metrics assessed, i.e., commonly used error metrics, the *OOS* R^2 as well as the performance of long-short ICC portfolios based on the predictions.

Importantly, we provide an intuitive breakdown of how important the different pieces of fundamental accounting information are for predicting earnings. We find that current I/S variables, especially current earnings, are the most important predictors. However, with increasing forecast horizon, variable importance becomes more balanced. More precisely, B/S information becomes much more important whereas I/S information becomes less important with increasing forecast horizon. Thoroughly disentangling the different financial statements suggests that this dynamic

[19] We plot the partial dependence of the three most important variables and the most important variable not stemming from the income statement.

may be attributable to earnings management.

As the first study to thoroughly decompose the effects of non-linearity in the earnings prediction context we find that especially for short term-horizons, the relationship as approximated by the best performing machine learning model, i.e., the machine learning ensemble, can still be described by a linear surrogate model to a large extent. More precisely, we find that on average around 84-89% of the variance in predictions can be explained by a linear model, depending on the forecast horizon. Interactions between financial statement variables play virtually no role, implying that non-linearity of the functional form resembles the non-linear part of the model. As the forecast horizon increases, the linear surrogate R^2 decreases slightly. Nevertheless, interactions across financial statement variables remain irrelevant, thus implying that non-linearity of functional form becomes more important with increasing forecast horizon.

Our findings provide important guidance for future research. First, we show that machine learning approaches are an excellent tool for earnings predictions. We hence argue that research which uses (model) earnings predictions in some way or another should resort to machine learning methods, if high accuracy is desired. Second, we show which financial statement variables and groups thereof are important. Future research may build upon that when building models and deciding which variables to include. Importantly, this includes the differences in terms of variable importance across forecast horizons. For example, if one is interested in an earnings prediction model including only a small number of variables for computation- or interpretation-related reasons, employing distinct (small) sets of variables for different forecast horizons might be beneficial.

Lastly and importantly we show that interactions among financial statement variables bear no predictive power. Again, future research may build upon this finding when deciding on an earnings prediction model.

References

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

Bali, T. G., H. Beckmeyer, M. Moerke, and F. Weigert (2023). Option return predictability with machine learning and big data. *Working Paper*.

Ball, R. and P. Brown (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research 6*(2), 159–178.

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

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

Campbell, J., H. Ham, Z. G. Lu, and K. Wood (2023). Expectations matter: When (not) to use machine learning earnings forecasts. *Working Paper*.

Cao, K. and H. You (2021). Fundamental analysis via machine learning. *Working Paper*.

Chen, X., Y. H. T. 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. *The Journal of Finance 56*(5), 1629–1666.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–63.

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.

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. J. and R. B. Gramacy (2012). Regression-based earnings forecasts. *Chicago Booth Research Paper No. 12-26*.

Gordon, J. R. and M. J. Gordon (1997). The finite horizon expected return model. *Financial Analysts Journal 53*(3), 52–61.

Graham, J. R., C. R. Harvey, and S. Rajgopal (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics 40*(1), 3–73.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

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

Harris, R. D. F. and P. Wang (2019). Model-based earnings forecasts vs. financial analysts' earnings forecasts. *British Accounting Review 51*(4), 424–437.

Hendriock, M. (2022). Forecasting Earnings with Predicted, Conditional Probability Distribution Density Functions. Working Paper.

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

Israel, R., B. Kelly, and T. Moskowitz (2020). Can machines "learn" finance? *Journal of Investment Management 18*(2).

Jones, J. J. (1991). Earnings management during import relief investigations. *Journal of Accounting Research 29*(2), 193–228.

Jones, S., W. J. Moser, and M. M. Wieland (2023). Machine learning and the prediction of changes in profitability. *Contemporary Accounting Research n/a*(n/a).

Kelly, B., S. Malamud, and K. Zhou (2022). The virtue of complexity in return prediction. *Working Paper*.

Li, K. K. and P. Mohanram (2014). Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies 13*(3), 1152–1185.

Liaw, R., E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica (2018). Tune: A research platform for distributed model selection and training. *Working Paper*.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774.

Monahan, S. J. (2018). Financial statement analysis and earnings forecasting. *Foundation and Trends in Accounting 12*, 105–215.

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

Ohlson, J. A. and B. E. Juettner-Nauroth (2005). Expected eps and eps growth as determinantsof value. *Review of Accounting Studies 10*, 349–365.

Schipper, K. (1991). Analysts' forecasts. *Accounting Horizons 5*(4), 105–121.

Van Binsbergen, J. H., X. Han, and A. Lopez-Lira (2023). Man vs. machine learning: The term structure of earnings expectations and conditional biases. *Review of Financial Studies 36*(6), 2361–2396.

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

Figures

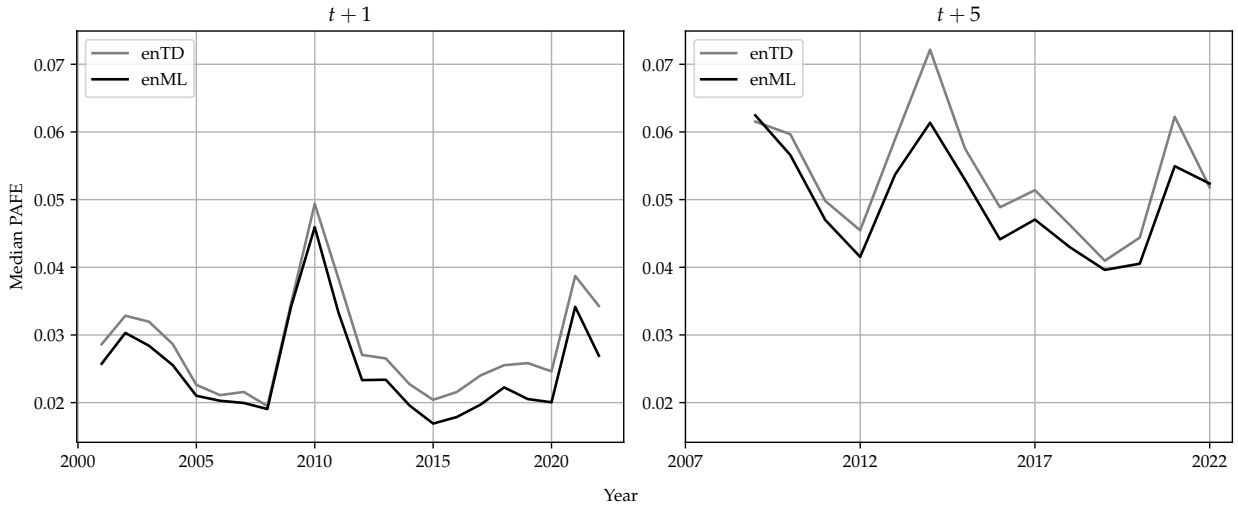

Figure 1: PAFE across out-of-sample periods
This figure shows the median price scaled absolute forecast errors (PAFEs) of the machine-learning ensemble (ENML) and the traditional ensemble (ENTD) per out-of-sample period for forecast horizons $t+1$ and $t+5$.

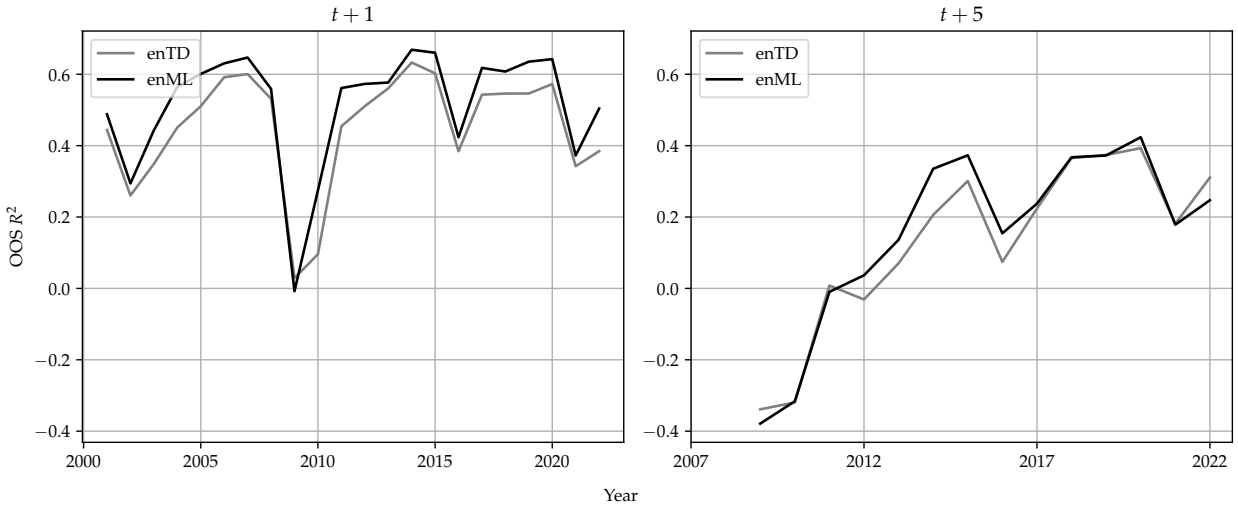

Figure 2: R^2 across out-of-sample periods
This figure shows the OOS R^2 of the machine-learning ensemble (ENML) and the traditional ensemble (ENTD) per out-of-sample period for forecast horizons $t+1$ and $t+5$.

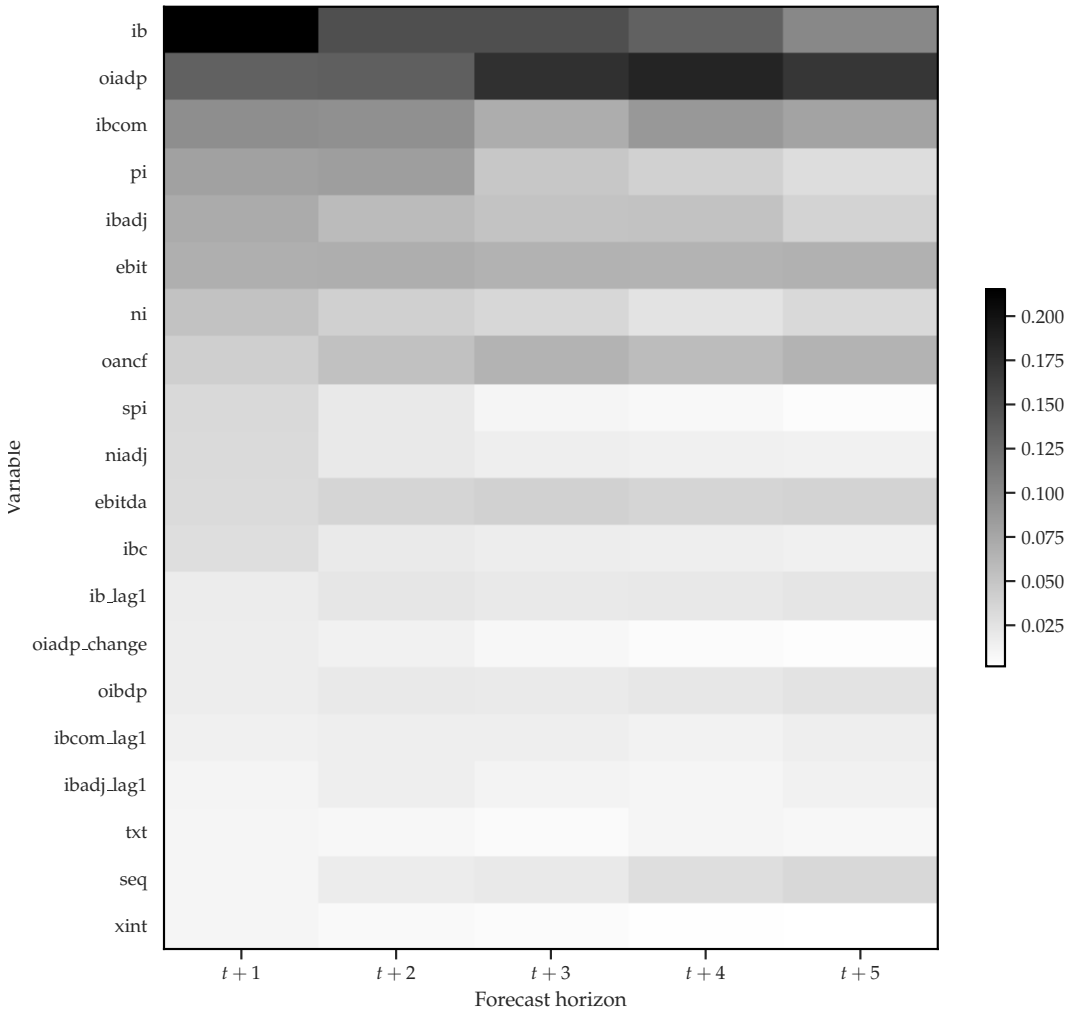

Figure 3: Variable importance for the machine learning forecast ensemble
This figure depicts the absolute SHAP values of the 20 most important variables for the machine learning ensemble, averaged over out-of-sample periods and scaled so they sum up to one within each forecast horizon. In this context, importance is defined as the ranking of the respective variable according to the aforementioned metric for forecast horizon $t + 1$.

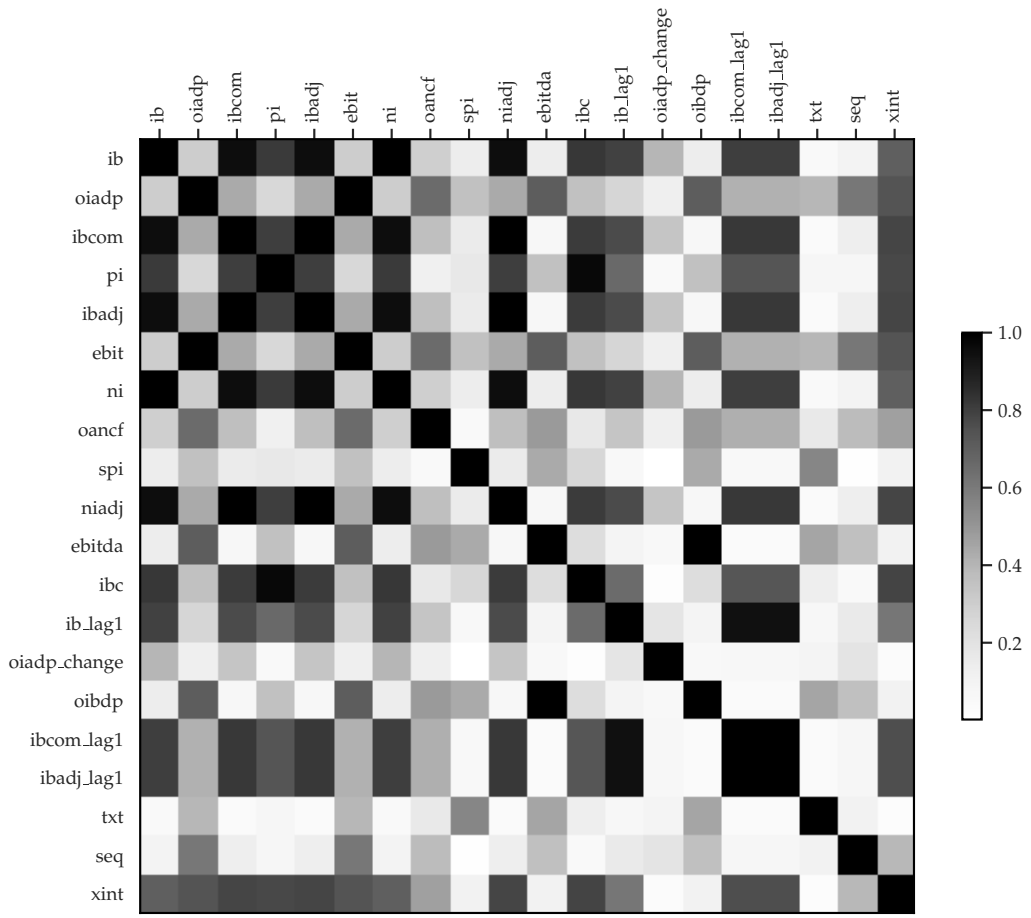

Figure 4: Correlation heatmap for the most important variables
This figure shows the absolute Pearson correlation coefficients for the 20 most important variables for the machine learning ensemble.

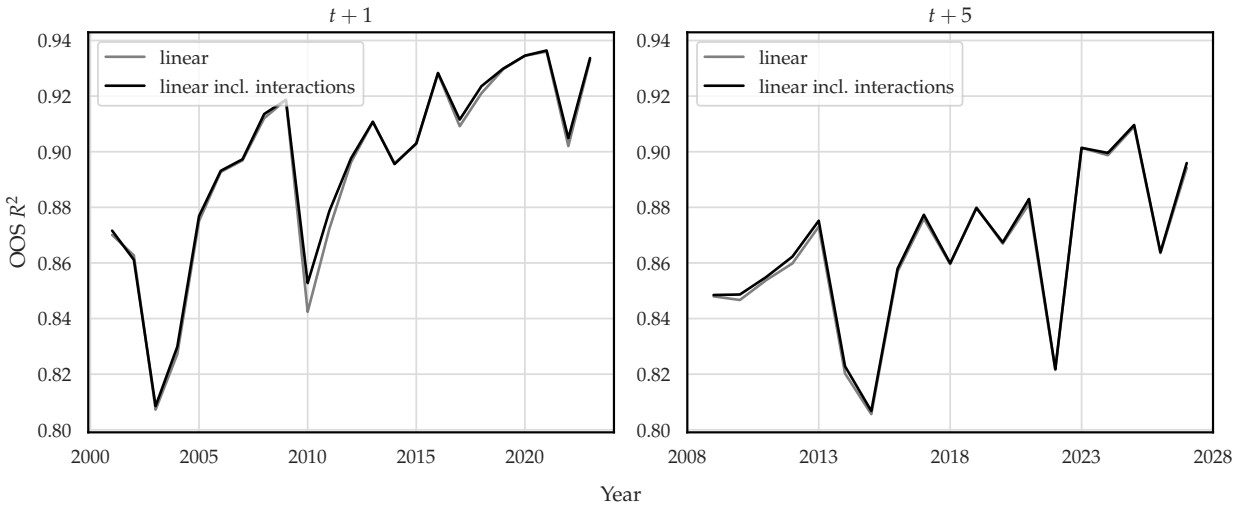

Figure 5: Surrogate models
This figure shows the adjusted R^2 of the surrogate models that we fit to our machine learning ensemble predictions for forecast horizons $t+1$ and $t+5$. The linear model (linear) is a simple linear model in which we regress the respective predictions on the 50 most important predictor variables according to their average absolute SHAP values for $t+1$ forecasts. The linear model including interactions (linear incl. interactions) is a linear model in which we use the same set of predictors as well as all possible two-way interactions.

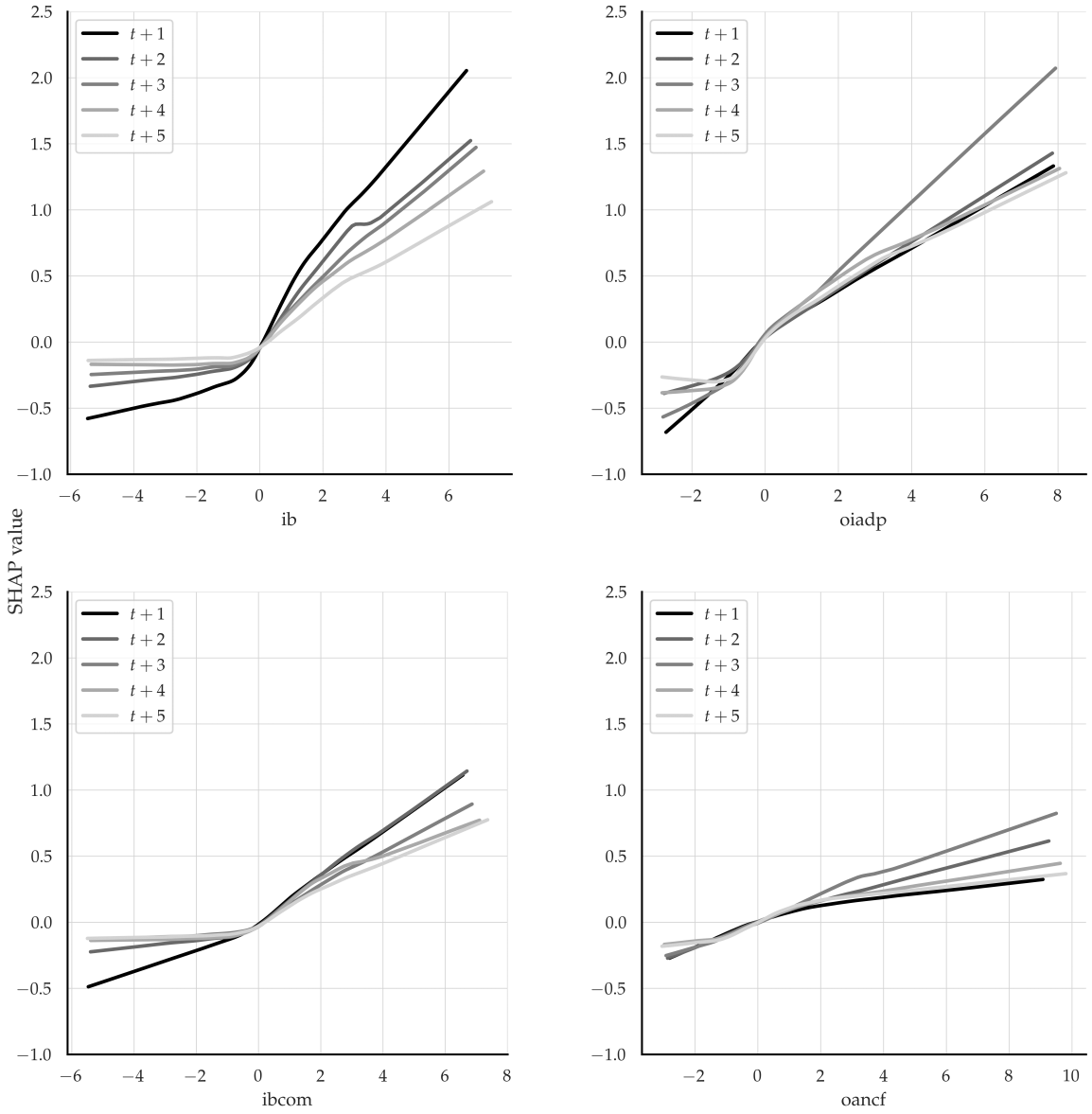

Figure 6: Partial dependence plots
The panels show the sensitivity of expected future earnings to the respective variable for all forecast horizons. More specifically, we fit a nonparametric lowess model (locally weighted linear regression) to the SHAP values of the respective variable.

Tables

Table 1: Median PFE

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	−0.0031	−0.0065***	−0.0105***	−0.0173***	−0.0254***
HVZ^{MSE}	0.0009	−0.0023	−0.0072	−0.0153***	−0.0256***
EP^{MSE}	0.0041***	0.0019	−0.0023	−0.0096***	−0.0185***
RI^{MSE}	0.0047***	0.0025	−0.0018	−0.0096***	−0.0191***
ENTD	0.0019	−0.0013	−0.0058	−0.0138***	−0.0230***
RF^{MSE}	0.0045***	0.0034	−0.0007	−0.0078***	−0.0163***
GBT^{MAE}	−0.0020	−0.0035	−0.0049	−0.0088***	−0.0129***
$DART^{MAE}$	−0.0013	−0.0029	−0.0045	−0.0075	−0.0133***
NN^{MAE}	−0.0015	−0.0040	−0.0044	−0.0093***	−0.0140***
ENML	0.0003	−0.0017	−0.0035	−0.0082***	−0.0146***

This table reports the time-series averages of the median price scaled forecasting errors (PFEs) for all models. E_{t+1} to E_{t+5} denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). The PFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

Table 2: Median PFE by firm size

	Large firms				
	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.0054***	0.0048***	0.0027	−0.0017	−0.0075
HVZ^{MSE}	0.0041***	0.0029	0.0006	−0.0041	−0.0104***
EP^{MSE}	0.0058***	0.0054***	0.0030	−0.0017	−0.0082
RI^{MSE}	0.0048***	0.0038	0.0009	−0.0047	−0.0116***
ENTD	0.0050***	0.0043	0.0018	−0.0031	−0.0096
RF^{MSE}	0.0047***	0.0043***	0.0017	−0.0030	−0.0090***
GBT^{MAE}	−0.0004	−0.0010	−0.0016	−0.0033	−0.0063
$DART^{MAE}$	0.0002	−0.0010	−0.0020	−0.0039	−0.0094***
NN^{MAE}	0.0001	−0.0001	0.0008	−0.0026	−0.0052
ENML	0.0012	0.0006	−0.0003	−0.0031	−0.0079
	Small firms				
	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	−0.0239***	−0.0315***	−0.0378***	−0.0491***	−0.0621***
HVZ^{MSE}	−0.0049	−0.0116***	−0.0212***	−0.0349***	−0.0508***
EP^{MSE}	0.0003	−0.0048	−0.0127***	−0.0241***	−0.0366***
RI^{MSE}	0.0043	0.0006	−0.0064	−0.0181***	−0.0310***
ENTD	−0.0045	−0.0111***	−0.0195***	−0.0320***	−0.0449***
RF^{MSE}	0.0041	0.0018	−0.0049	−0.0167***	−0.0293***
GBT^{MAE}	−0.0053	−0.0076	−0.0112***	−0.0188***	−0.0241***
$DART^{MAE}$	−0.0044	−0.0058	−0.0086	−0.0132	−0.0200***
NN^{MAE}	−0.0046	−0.0126***	−0.0140***	−0.0208***	−0.0302***
ENML	−0.0019	−0.0056	−0.0090***	−0.0169***	−0.0255***

This table reports the time-series averages of the median price scaled forecasting errors (PFEs) for all models, stratified by firm size. E_{t+1} to E_{t+5} denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). The PFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. Per out-of-sample period, we classify firms as either small or large, depending on whether their market capitalization is below or above the median as of the prediction date. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

Table 3: Median PAFE

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.0317***	0.0409***	0.0463***	0.0531***	0.0606***
HVZ^{MSE}	0.0293***	0.0376***	0.0429***	0.0499***	0.0578***
EP^{MSE}	0.0297***	0.0375***	0.0416***	0.0471***	0.0535***
RI^{MSE}	0.0293***	0.0365***	0.0403***	0.0453***	0.0517***
ENTD	0.0282***	0.0360***	0.0406***	0.0467***	0.0536***
RF^{MSE}	0.0273***	0.0349***	0.0402***	0.0463***	0.0523***
GBT^{MAE}	0.0274***	0.0357***	0.0407***	0.0460***	0.0515***
$DART^{MAE}$	0.0253***	0.0342***	0.0399***	0.0465***	0.0531***
NN^{MAE}	0.0257***	0.0361***	0.0410***	0.0451***	0.0524***
ENML	0.0249***	0.0335***	0.0384***	0.0441***	0.0498***
ENML - ENTD	−0.0033***	−0.0026***	−0.0022***	−0.0026***	−0.0038***

This table reports the time-series averages of the median price scaled absolute forecasting errors (PAFEs) for all models. E_{t+1} to E_{t+5} denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). The PAFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

Table 4: Median PAFE by firm size

	Large firms				
	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.0201***	0.0266***	0.0300***	0.0346***	0.0390***
HVZ^{MSE}	0.0200***	0.0266***	0.0304***	0.0350***	0.0399***
EP^{MSE}	0.0193***	0.0258***	0.0292***	0.0336***	0.0381***
RI^{MSE}	0.0190***	0.0253***	0.0287***	0.0331***	0.0381***
ENTD	0.0190***	0.0252***	0.0287***	0.033***	0.0377***
RF^{MSE}	0.0182***	0.0245***	0.0285***	0.0329***	0.0373***
GBT^{MAE}	0.0180***	0.0246***	0.0291***	0.0332***	0.0377***
$DART^{MAE}$	0.0168***	0.0239***	0.0285***	0.0337***	0.0390***
NN^{MAE}	0.0169***	0.0250***	0.0300***	0.0330***	0.0389***
ENML	0.0165***	0.0231***	0.0277***	0.0320***	0.0367***
ENML - ENTD	−0.0025***	−0.0021***	−0.0009	−0.0010***	−0.0010
	Small firms				
	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.0583***	0.0724***	0.0802***	0.0898***	0.1010***
HVZ^{MSE}	0.0474***	0.0575***	0.0645***	0.0742***	0.0854***
EP^{MSE}	0.0513***	0.0593***	0.0631***	0.0695***	0.0780***
RI^{MSE}	0.0506***	0.0580***	0.0602***	0.0649***	0.0734***
ENTD	0.0466***	0.0554***	0.0613***	0.0693***	0.0791***
RF^{MSE}	0.0443***	0.0527***	0.0594***	0.0677***	0.0764***
GBT^{MAE}	0.0444***	0.0545***	0.0598***	0.0656***	0.0728***
$DART^{MAE}$	0.0408***	0.0512***	0.0588***	0.0662***	0.0747***
NN^{MAE}	0.0416***	0.0555***	0.0590***	0.0641***	0.0733***
ENML	0.0399***	0.0503***	0.0559***	0.0626***	0.0697***
ENML - ENTD	−0.0067***	−0.0051***	−0.0054***	−0.0067***	−0.0094***

This table reports the time-series averages of the median price scaled absolute forecasting errors (PAFEs) for all models, stratified by firm size. E_{t+1} to E_{t+5} denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). The PAFE is calculated as the difference between actual and forecasted earnings per share, scaled by price at the end of June of the respective estimation year. Per out-of-sample period, we classify firms as either small or large, depending on whether their market capitalization is below or above the median as of the prediction date. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

Table 5: Pairwise Diebold-Mariano test statistics

	L^{MSE}	HVZ^{MSE}	EP^{MSE}	RI^{MSE}	ENTD	RF^{MSE}	GBT ^{MAE}	$DART^{MAE}$	NN^{MAE}	ENML
L^{MSE}		0.0107***	0.0098***	0.0104***	0.0123***	0.0168***	0.0149***	0.0182***	0.017***	0.0192***
HVZ^{MSE}			−0.0009	−0.0003	0.0016	0.0061***	0.0042***	0.0075***	0.0062***	0.0085***
EP^{MSE}				0.0006	0.0024**	0.0070***	0.0051***	0.0084***	0.0071***	0.0094***
RI^{MSE}					0.0018**	0.0064***	0.0045***	0.0078***	0.0065***	0.0088***
ENTD						0.0045***	0.0027**	0.006***	0.0047***	0.0069***
RF^{MSE}							−0.0019**	0.0014*	0.0002	0.0024***
GBT^{MAE}								0.0033***	0.0020***	0.0043***
$DART^{MAE}$									−0.0013*	0.0010
NN^{MAE}										0.0023***
ENML										

This table reports the Diebold-Mariano (Diebold and Mariano, 1995) test statistics for each model comparison. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). Positive numbers indicate the column model outperforms the respective row model. ***, **, and * denote the Bonferroni-adjusted significance levels at 10%, 5% and 1%, respectively. Standard errors used to derive statistical significance are adjusted following Newey and West (1987) assuming a lag length of three. P-values are adjusted by applying the Bonferroni procedure.

Table 6: Average out-of-sample R^2

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.3959	0.2511	0.1951	0.1251	0.0535
HVZ^{MSE}	0.4495	0.3295	0.2662	0.1789	0.0899
EP^{MSE}	0.4450	0.3219	0.2581	0.1650	0.0805
RI^{MSE}	0.4589	0.3439	0.2813	0.1883	0.0966
ENTD	0.4519	0.3382	0.2856	0.2056	0.1298
RF^{MSE}	0.4971	0.3755	0.3165	0.2237	0.1390
GBT^{MAE}	0.4858	0.3498	0.2774	0.2089	0.1135
$DART^{MAE}$	0.5091	0.3778	0.3057	0.2085	0.1144
NN^{MAE}	0.4992	0.2075	0.2631	0.2045	0.0736
ENML	0.5153	0.3785	0.3200	0.2390	0.1541
ENML - ENTD	0.0634***	0.0403**	0.0345***	0.0332***	0.0243

This table reports the time-series averages of the out-of-sample R^2s ($OOS\ R^2$s) for all models. E_{t+1} to E_{t+5} denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). ***, **, and * denote statistical significance at the 10%, the 5% and the 1% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years. We only test for statistical significance of the difference between the ensemble models (ENML - ENTD).

Table 7: Average out-of-sample R^2 by firm size

	Large firms				
	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.3808	0.2247	0.1694	0.0999	0.0203
HVZ^{MSE}	0.4360	0.3032	0.2379	0.1495	0.0521
EP^{MSE}	0.4285	0.2885	0.2144	0.1090	0.0050
RI^{MSE}	0.4455	0.3131	0.2394	0.1325	0.0225
ENTD	0.4366	0.3073	0.2484	0.1615	0.0712
RF^{MSE}	0.4781	0.3412	0.2822	0.1877	0.0968
GBT^{MAE}	0.4707	0.3167	0.2379	0.1631	0.0575
$DART^{MAE}$	0.4972	0.3468	0.2628	0.1559	0.0524
NN^{MAE}	0.4872	0.1570	0.2127	0.1558	0.0074
ENML	0.5028	0.3474	0.2825	0.1962	0.1034
ENML - ENTD	0.0662***	0.0400	0.0341***	0.0347***	0.0322
	Small firms				
	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
L^{MSE}	0.2585	0.0967	0.0224	−0.0514	−0.1111
HVZ^{MSE}	0.3255	0.2042	0.1250	0.0320	−0.0439
EP^{MSE}	0.3243	0.2110	0.1500	0.0728	0.0246
RI^{MSE}	0.3372	0.2362	0.1779	0.1042	0.0449
ENTD	0.3322	0.2261	0.1710	0.0991	0.0516
RF^{MSE}	0.3983	0.2831	0.2044	0.1047	0.0311
GBT^{MAE}	0.3761	0.2503	0.1654	0.1097	0.0306
$DART^{MAE}$	0.4011	0.2808	0.2099	0.1227	0.0447
NN^{MAE}	0.3861	0.1182	0.1703	0.1071	−0.0028
ENML	0.4098	0.2832	0.2169	0.1405	0.0697
ENML - ENTD	0.0776***	0.0571***	0.0459***	0.0414***	0.0181

This table reports the time-series averages of the out-of-sample R^2s ($OOS\ R^2$s) for all models, stratified by firm size. E_{t+1} to E_{t+5} denote one- to five-year ahead earnings. L is a model with only current earnings as a predictor, HVZ is the model by Hou et al. (2012), EP and RI are the models by Li and Mohanram (2014), ENTD is an equally weighted ensemble of L, HVZ, EP, and RI, RF is a random forest model, GBT and DART are gradient boosted tree models without and with dropout, NN is a neural net, and ENML is an equally weighted ensemble of RF, GBT, DART, and NN. The superscript MAE (MSE) indicates that the respective model is estimated using the mean absolute error (mean squared error) as its loss function. We follow the literature and estimate the traditional models using the MSE. In untabulated results we find that our results are robust to estimating the traditional models using the MAE. We decide on which loss function to report for the ML models depending on which one yields more accurate predictions (as indicated by the price scaled absolute forecast error (PAFE) for the $t+1$ horizon). Per out-of-sample period, we classify firms as either small or large, depending on whether their market capitalization is below or above the median as of the prediction date. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years. We only test for statistical significance of the difference between the ensemble models (ENML - ENTD).

Table 8: Long-short ICC portfolio performance

		ENTD		ENML	
	Decile	ICC	Realized	ICC	Realized
	1	0.0301	0.0809	0.0249	0.0718
$\hat{r}_{\to t+1}$	10	0.2609	0.1524	0.2587	0.1632
	10-1	0.2308	0.0715	0.2328	0.0914*
	1	0.0301	0.0899	0.0249	0.0845
$\hat{r}_{\to t+2}$	10	0.2609	0.1673	0.2587	0.1780
	10-1	0.2308	0.0775	0.2338	0.0935**
	1	0.0301	0.0990	0.0249	0.0938
$\hat{r}_{\to t+3}$	10	0.2609	0.1583	0.2587	0.1635
	10-1	0.2308	0.0589*	0.2338	0.0691**

This table reports the composite implied cost of capital (ICC) estimates as well as the annual geometric average returns of buy-and-hold portfolios sorted conditionally on them, denoted by $\hat{r}_{\to t+\kappa}$, for rebalancing frequencies of every $\kappa \in [1, 2, 3]$ years. The 10-1 rows denote the long-short portfolios, in which an investor goes short the lowest ICC decile and long the highest ICC decile. We test for significance of the realized buy-and-hold return of this portfolio. ***, **, and * denote statistical significance at the 1%, the 5% and the 10% level, respectively. Standard errors used for deriving statistical significance are adjusted following Newey and West (1987) assuming a lag length of three years.

Table 9: Variable importance by groups

Panel A: Financial statement type

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
B/S	0.2023	0.2497	0.3010	0.3437	0.3733
CF/S	0.1517	0.1620	0.1619	0.1557	0.1545
I/S	0.6460	0.5883	0.5371	0.5006	0.4723

Panel B: Variable type

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
Current	0.7123	0.6701	0.6506	0.6462	0.6228
Lagged	0.1422	0.1933	0.2109	0.2164	0.2472
Change	0.1455	0.1366	0.1385	0.1374	0.1299

Panel C: Financial statement type $\times$ variable type

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
B/S current	0.0923	0.1196	0.1493	0.1767	0.2034
B/S lagged	0.0443	0.0654	0.0848	0.0973	0.1027
B/S change	0.0657	0.0648	0.0669	0.0697	0.0671
CF/S current	0.0942	0.0958	0.0928	0.0856	0.0828
CF/S lagged	0.0293	0.0392	0.0408	0.0401	0.0446
CF/S change	0.0282	0.0269	0.0284	0.0300	0.0271
I/S current	0.5258	0.4547	0.4086	0.3840	0.3366
I/S lagged	0.0687	0.0887	0.0854	0.0790	0.0999
I/S change	0.0516	0.0449	0.0432	0.0377	0.0357

Panel A reports the relative variable importance per financial statement group. B/S , I/S and CF/S denote balance sheet, income statement and cash flow statement, respectively. The variables are grouped according to Table B.2 in the Appendix. Panel B reports the relative variable importance per variable type. Panel C reports the relative variable importance per financial statement type $\times$ variable type group. Importance per group in each Panel is defined as the fraction that the respective group contributes to total importance, measured as the sum of absolute SHAP values.

Table 10: Variable importance: financial statements

Panel A: Balance sheet

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
Current assets	0.0302	0.0372	0.0445	0.0542	0.0589
Fixed assets	0.0214	0.0243	0.0259	0.0300	0.0326
Total assets	0.0046	0.0076	0.0105	0.0118	0.0177
Debt	0.0513	0.0604	0.0709	0.0828	0.0925
Equity	0.0367	0.0474	0.0558	0.0613	0.0610
Total debt & equity	0.0028	0.0034	0.0052	0.0074	0.0077
Supplemental	0.0553	0.0695	0.0881	0.0963	0.1028
Sum B/S importance	0.2023	0.2497	0.3010	0.3437	0.3733

Panel B: Cash flow statement

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
Operating cash flow	0.1083	0.1087	0.1049	0.0999	0.1005
Investing cash flow	0.0146	0.0188	0.0204	0.0190	0.0214
Financing cash flow	0.0245	0.0297	0.0309	0.0317	0.0284
Total cash flow	0.0042	0.0047	0.0057	0.0051	0.0043
Sum CF/S importance	0.1517	0.1620	0.1619	0.1557	0.1545

Panel C: Income statement

	E_{t+1}	E_{t+2}	E_{t+3}	E_{t+4}	E_{t+5}
Sales	0.0061	0.0102	0.0135	0.0140	0.0167
Operating expenses	0.0121	0.0163	0.0199	0.0245	0.0261
EBITDA	0.0335	0.0361	0.0350	0.0322	0.0347
Depr. & Amort.	0.0078	0.0066	0.0070	0.0089	0.0123
EBIT	0.1355	0.1355	0.1414	0.1326	0.1279
Interest expenses	0.0600	0.0516	0.0453	0.0393	0.0377
EBT	0.0517	0.0514	0.0308	0.0237	0.0178
Tax expenses	0.0224	0.0259	0.0262	0.0297	0.0306
Net income	0.3063	0.2403	0.2006	0.1812	0.1555
Dividends	0.0106	0.0145	0.0174	0.0146	0.0128
Sum I/S importance	0.6460	0.5883	0.5371	0.5006	0.4723

This table reports the relative variable importance per financial statement group. B/S , I/S and CF/S denote balance sheet, income statement and cash flow statement, respectively. EBITDA denotes earnings before interest, taxes and depreciation and amortization. EBIT denotes earnings before interest and taxes. EBT denotes earnings before taxes. The variables are grouped according to Table B.2 in the Appendix. Importance per financial statement component is defined as the fraction that the respective component contributes to total importance, measured as the sum of absolute SHAP values.

Appendix A Traditional Earnings Prediction Models

Table A.1: Traditional earnings prediction models

Panel A: Traditional model specifications

Name	Model	Source
L	$\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t}$	Gerakos and Gramacy (2012)
HVZ	$\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 A_{i,t} + \beta_3 D_{i,t} + \beta_4 DD_{i,t} + \beta_5 NegE_{i,t} + \beta_6 ACC_{i,t}$	Hou et al. (2012)
EP	$\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 NegE_{i,t} + \beta_3 NegE_{i,t} E_{i,t}$	Li and Mohanram (2014)
RI	$\mathbb{E}_t[E_{i,t+\tau}] = \beta_0 + \beta_1 E_{i,t} + \beta_2 NegE_{i,t} + \beta_3 NegE_{i,t} E_{i,t} + \beta_4 B_{i,t} + \beta_5 ACC_{i,t}$	Li and Mohanram (2014)

Panel B: Variable definitions

Variable	Definition
E	Income before extraordinary items (ib) / Common shares outstanding (csho)
A	Total assets (at) / csho
D	Dividends total (dvt) / csho
DD	1 if dvt > 0; 0 else
$NegE$	1 if ib < 0; 0 else
ACC	(Income before extraordinary items (ib) - Operating activities - net cash flow (oancf)) / csho
B	Common/Ordinary equity- total (ceq) / csho

Panel A reports the traditional earnings models estimated. $\mathbb{E}_t[E_{i,t+\tau}]$ denotes the expectation for earnings E of firm i in period $t+\tau$ as of t. β_0-β_5 are the model coefficients. Panel B reports the variable definitions for the traditional models. Compustat variable names are provided in parentheses. Note the slight changes as opposed to the original papers. More precisely, we scale all variables by common shares outstanding and use a consistent earnings as well as accruals definition.

Appendix B Machine Learning Earnings Prediction Models

Table B.1: Hyperparameters for the machine learning models

	RF, GBT & DART	
	Maximum number of trees	512
	Learning rate	$\in [0.001, 0.01, 0.1, 1]$
	Maximum depth	$U^{int}(2, 10)$
	Maximum number of leaves	$U^{int}(2, 512)$
	L1-regularization	$U(0, 0.1)$
	L2-regularization	$U(0, 0.1)$
	Feature fraction	$U(0.25, 1)$
	Bagging fraction	$U(0.25, 1)$
	Bagging frequency	$\in (1, 10, 50)$
DART	Dropout rate	$\in (0.05, 0.1, 0.15)$
DART	Probability of skipping dropout	$\in (0.25, 0.5)$
	NN	
	Learning rate	$\in [0.001, 0.01, 0.1, 1]$
	L1-regularization	$U(0, 0.1)$
	Dropout	$U(0, 0.5)$
	Number of hidden layers	$\in [1, 2, 3, 4, 5]$
	First layer size	$\in [32, 64, 128]$
	Batch size	$\in [2^{11}, 2^{12}, 2^{13}, 2^{14}]$

This table gives the hyperparameters that we tune and their respective boundaries. U (U^{int}) means drawing from a uniform (integer-wise uniform) distribution. Our choice of hyperparameters and their respective boundaries is based on Bali et al. (2023). We use the *Ray* Python framework to efficiently optimize the hyperparameters (Liaw et al., 2018).

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
1	aco	Current assets - other - total	Balance sheet	Current assets
2	acox	Current assets - other - sundry	Balance sheet	Current assets
3	act	Current assets - total	Balance sheet	Current assets
4	am	Amortization of intangibles	Income statement	Depreciation and amortization
5	ao	Assets - other	Balance sheet	Fixed assets
6	aoloch	Assets and liabilities - other - net change	Cash flow statement	Operating cash flow
7	aox	Assets - other - sundry	Balance sheet	Fixed assets
8	ap	Accounts payable - trade	Balance sheet	Liabilities
9	apalch	Accounts payable and accrued liabilities - increase/(decrease)	Cash flow statement	Operating cash flow
10	aqc	Acquisitions	Cash flow statement	Investing cash flow
11	aqi	Acquisitions - income contribution	Income statement	Interest and other
12	aqs	Acquisitions - sales contribution	Income statement	Sales
13	at	Assets - total	Balance sheet	Total assets
14	caps	Capital surplus/share premium reserve	Balance sheet	Equity
15	capx	Capital expenditures	Cash flow statement	Investing cash flow
16	capxv	Capital expend property, plant and equipment schd v	Cash flow statement	Investing cash flow
17	ceq	Common/ordinary equity - total	Balance sheet	Equity
18	ceql	Common equity - liquidation value	Balance sheet	Supplemental
19	ceqt	Common equity - tangible	Balance sheet	Supplemental
20	ch	Cash	Balance sheet	Current assets
21	che	Cash and short-term investments	Balance sheet	Current assets
22	chech	Cash and cash equivalents - increase/(decrease)	Cash flow statement	Total cash flow

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
23	cld2	Capitalized leases - due in 2nd year	Balance sheet	Supplemental
24	cld3	Capitalized leases - due in 3rd year	Balance sheet	Supplemental
25	cld4	Capitalized leases - due in 4th year	Balance sheet	Supplemental
26	cld5	Capitalized leases - due in 5th year	Balance sheet	Supplemental
27	cogs	Cost of goods sold	Income statement	Operating expenses
28	cstk	Common/ordinary stock (capital)	Balance sheet	Equity
29	cstkcv	Common stock-carrying value	Balance sheet	Supplemental
30	cstke	Common stock equivalents - dollar savings	Income statement	Interest and other
31	dc	Deferred charges	Balance sheet	Fixed assets
32	dclo	Debt - capitalized lease obligations	Balance sheet	Liabilities
33	dcpstk	Convertible debt and preferred stock	Balance sheet	Supplemental
34	dcvsr	Debt - senior convertible	Balance sheet	Liabilities
35	dcvsub	Debt - subordinated convertible	Balance sheet	Liabilities
36	dcvt	Debt - convertible	Balance sheet	Liabilities
37	dd	Debt - debentures	Balance sheet	Liabilities
38	dd1	Long-term debt due in one year	Balance sheet	Liabilities
39	dd2	Debt - due in 2nd year	Balance sheet	Liabilities
40	dd3	Debt - due in 3rd year	Balance sheet	Liabilities
41	dd4	Debt - due in 4th year	Balance sheet	Liabilities
42	dd5	Debt - due in 5th year	Balance sheet	Liabilities
43	dlc	Debt in current liabilities - total	Balance sheet	Liabilities
44	dltis	Long-term debt - issuance	Cash flow statement	Financing cash flow

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
45	dlto	Other long-term debt	Balance sheet	Liabilities
46	dltp	Long-term debt - tied to prime	Balance sheet	Liabilities
47	dltr	Long-term debt - reduction	Cash flow statement	Financing cash flow
48	dltt	Long-term debt - total	Balance sheet	Liabilities
49	dm	Debt - mortgages and other secured	Balance sheet	Liabilities
50	dn	Debt - notes	Balance sheet	Liabilities
51	do	Discontinued operations	Income statement	Interest and other
52	dp	Depreciation and amortization	Income statement	Depreciation and amortization
53	dpact	Depreciation, depletion and amortization (accumulated)	Balance sheet	Fixed assets
54	dpc	Depreciation and amortization (cash flow)	Cash flow statement	Operating cash flow
55	dpvieb	Depreciation (accumulated) - ending balance (schedule vi)	Balance sheet	Supplemental
56	ds	Debt-subordinated	Balance sheet	Liabilities
57	dudd	Debt - unamortized debt discount and other	Balance sheet	Liabilities
58	dv	Cash dividends (cash flow)	Cash flow statement	Financing cash flow
59	dvc	Dividends common/ordinary	Income statement	Dividends
60	dvp	Dividends - preferred/preference	Income statement	Dividends
61	dvpa	Preferred dividends in arrears	Balance sheet	Supplemental
62	dvt	Dividends - total	Income statement	Dividends
63	dxd2	Debt (excl capitalized leases) - due in 2nd year	Balance sheet	Supplemental
64	dxd3	Debt (excl capitalized leases) - due in 3rd year	Balance sheet	Supplemental
65	dxd4	Debt (excl capitalized leases) - due in 4th year	Balance sheet	Supplemental
66	dxd5	Debt (excl capitalized leases) - due in 5th year	Balance sheet	Supplemental

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
67	ebit	Earnings before interest and taxes	Income statement	EBIT
68	ebitda	Earnings before interest	Income statement	EBITDA
69	esub	Equity in earnings - unconsolidated subsidiaries	Income statement	Interest and other
70	esubc	Equity in net loss - earnings	Cash flow statement	Operating cash flow
71	exre	Exchange rate effect	Cash flow statement	Total cash flow
72	fatb	Property, plant, and equipment - buildings at cost	Balance sheet	Supplemental
73	fatc	Property, plant, and equipment - construction in progress at cost	Balance sheet	Supplemental
74	fate	Property, plant, and equipment - machinery and equipment at cost	Balance sheet	Supplemental
75	fatl	Property, plant, and equipment - leases at cost	Balance sheet	Supplemental
76	fatn	Property, plant, and equipment - natural resources at cost	Balance sheet	Supplemental
77	fato	Property, plant, and equipment - other at cost	Balance sheet	Supplemental
78	fatp	Property, plant, and equipment - land and improvements at cost	Balance sheet	Supplemental
79	fiao	Financing activities - other	Cash flow statement	Financing cash flow
80	fincf	Financing activities - net cash flow	Cash flow statement	Financing cash flow
81	fopo	Funds from operations - other	Cash flow statement	Operating cash flow
82	gp	Gross profit	Income statement	Operating expenses
83	ib	Income before extraordinary items	Income statement	Net income
84	ibadj	Income before extraordinary items - adjusted for common stock equivalents	Income statement	Net income
85	ibc	Income before extraordinary items (cash flow)	Cash flow statement	Operating cash flow
86	ibcom	Income before extraordinary items - available for common	Income statement	Net income
87	icapt	Invested capital - total	Balance sheet	Supplemental
88	idit	Interest and related income - total	Income statement	Interest and other

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
89	intan	Intangible assets - total	Balance sheet	Fixed assets
90	intc	Interest capitalized	Income statement	Interest and other
91	intpn	Interest paid - net	Cash flow statement	Operating cash flow
92	invch	Inventory - decrease (increase)	Cash flow statement	Operating cash flow
93	invfg	Inventories - finished goods	Balance sheet	Current assets
94	invo	Inventories - other	Balance sheet	Current assets
95	invrm	Inventories - raw materials	Balance sheet	Current assets
96	invt	Inventories - total	Balance sheet	Current assets
97	invwip	Inventories - work in process	Balance sheet	Current assets
98	itcb	Investment tax credit (balance sheet)	Balance sheet	Liabilities
99	itci	Investment tax credit (income account)	Income statement	Taxes
100	ivaco	Investing activities - other	Cash flow statement	Investing cash flow
101	ivaeq	Investment and advances - equity	Balance sheet	Fixed assets
102	ivao	Investment and advances - other	Balance sheet	Fixed assets
103	ivch	Increase in investments	Cash flow statement	Investing cash flow
104	ivncf	Investing activities - net cash flow	Cash flow statement	Investing cash flow
105	ivst	Short-term investments - total	Balance sheet	Current assets
106	ivstch	Short-term investments - change	Cash flow statement	Investing cash flow
107	lco	Current liabilities - other - total	Balance sheet	Liabilities
108	lcox	Current liabilities - other - sundry	Balance sheet	Liabilities
109	lct	Current liabilities - total	Balance sheet	Liabilities
110	lifr	Lifo reserve	Balance sheet	Supplemental

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
111	lo	Liabilities - other - total	Balance sheet	Liabilities
112	lse	Liabilities and stockholders equity - total	Balance sheet	Total liabilities and equity
113	lt	Liabilities - total	Balance sheet	Liabilities
114	mib	Noncontrolling interest (balance sheet)	Balance sheet	Liabilities
115	mii	Noncontrolling interest (income account)	Income statement	Interest and other
116	mrc1	Rental commitments - minimum - 1st year	Balance sheet	Supplemental
117	mrc2	Rental commitments - minimum - 2nd year	Balance sheet	Supplemental
118	mrc3	Rental commitments - minimum - 3rd year	Balance sheet	Supplemental
119	mrc4	Rental commitments - minimum - 4th year	Balance sheet	Supplemental
120	mrc5	Rental commitments - minimum - 5th year	Balance sheet	Supplemental
121	mrct	Rental commitments - minimum - 5 year total	Balance sheet	Supplemental
122	msa	Marketable securities adjustment	Balance sheet	Supplemental
123	ni	Net income (loss)	Income statement	Net income
124	niadj	Net income adjusted for common/ordinary stock (capital) equivalents	Income statement	Net income
125	nopi	Nonoperating income (expense)	Income statement	Interest and other
126	nopio	Nonoperating income (expense) - other	Income statement	Interest and other
127	np	Notes payable - short-term borrowings	Balance sheet	Liabilities
128	oancf	Operating activities - net cash flow	Cash flow statement	Operating cash flow
129	oiadp	Operating income after depreciation	Income statement	EBIT
130	oibdp	Operating income before depreciation	Income statement	EBITDA
131	pi	Pretax income	Income statement	EBT
132	ppegt	Property, plant and equipment - total (gross)	Balance sheet	Fixed assets

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
133	ppent	Property, plant and equipment - total (net)	Balance sheet	Fixed assets
134	ppeveb	Property, plant, and equipment - ending balance (schedule v)	Balance sheet	Supplemental
135	prstkc	Purchase of common and preferred stock	Cash flow statement	Financing cash flow
136	pstk	Preferred/preference stock (capital) - total	Balance sheet	Equity
137	pstkc	Preferred stock - convertible	Balance sheet	Equity
138	pstkl	Preferred stock - liquidating value	Balance sheet	Supplemental
139	pstkn	Preferred/preference stock - nonredeemable	Balance sheet	Equity
140	pstkr	Preferred/preference stock - redeemable	Balance sheet	Equity
141	pstkrv	Preferred stock - redemption value	Balance sheet	Supplemental
142	re	Retained earnings	Balance sheet	Equity
143	rea	Retained earnings - restatement	Balance sheet	Supplemental
144	reajo	Retained earnings - other adjustments	Balance sheet	Supplemental
145	recch	Accounts receivable - decrease (increase)	Cash flow statement	Operating cash flow
146	recco	Receivables - current - other	Balance sheet	Current assets
147	recd	Receivables - estimated doubtful	Balance sheet	Current assets
148	rect	Receivables - tota	Balance sheet	Current assets
149	recta	Retained earnings - cumulative translation adjustment	Balance sheet	Supplemental
150	rectr	Receivables - trade	Balance sheet	Current assets
151	reuna	Retained earnings - unadjusted	Balance sheet	Equity
152	revt	Revenue - total	Income statement	Sales
153	sale	Sales/turnover (net)	Income statement	Sales
154	seq	Stockholders equity - parent	Balance sheet	Equity

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
155	siv	Sale of investments	Cash flow statement	Investing cash flow
156	spi	Special items	Income statement	Interest and other
157	sppe	Sale of property	Cash flow statement	Operating cash flow
158	sppiv	Sale of property, plant and equipment and investments - gain (loss)	Cash flow statement	Operating cash flow
159	sstk	Sale of common and preferred stock	Cash flow statement	Financing cash flow
160	tlcf	Tax loss carry forward	Balance sheet	Supplemental
161	tstk	Treasury stock - total (all capital)	Balance sheet	Equity
162	tstkc	Treasury stock - common	Balance sheet	Equity
163	tstkp	Treasury stock - preferrred	Balance sheet	Equity
164	txach	Income taxes - accrued - increase/(decrease)	Cash flow statement	Operating cash flow
165	txc	Income taxes - current	Income statement	Taxes
166	txdb	Deferred taxes (balance sheet)	Balance sheet	Liabilities
167	txdc	Deferred taxes (cash flow)	Income statement	Operating cash flow
168	txdfed	Deferred taxes-federal	Income statement	Taxes
169	txdfo	Deferred taxes-foreign	Income statement	Taxes
170	txdi	Income taxes - deferred	Income statement	Taxes
171	txditc	Deferred taxes and investment tax credit	Balance sheet	Liabilities
172	txds	Deferred taxes-state	Income statement	Taxes
173	txfed	Income taxes - federal	Income statement	Taxes
174	txfo	Income taxes - foreign	Income statement	Taxes
175	txo	Income taxes - other	Income statement	Taxes
176	txp	Income taxes payable	Balance sheet	Liabilities

Continued on next page

Table B.2: Predictor variables for the machine learning models

	Variable	Compustat description	Financial statement	Component
177	txpd	Income taxes paid	Cash flow statement	Operating cash flow
178	txr	Income tax refund	Balance sheet	Current assets
179	txs	Income taxes - state	Income statement	Taxes
180	txt	Income taxes - total	Income statement	Taxes
181	txw	Excise taxes	Income statement	Taxes
182	wcap	Working capital (balance sheet)	Balance sheet	Supplemental
183	xacc	Accrued expenses	Balance sheet	Liabilities
184	xi	Extraordinary items	Income statement	Interest and other
185	xido	Extraordinary items and discontinued operations	Income statement	Interest and other
186	xidoc	Extraordinary items and discontinued operations (cash flow)	Cash flow statement	Operating cash flow
187	xint	Interest and related expense - total	Income statement	Interest and other
188	xopr	Operating expenses - total	Income statement	Operating expenses
189	xpp	Prepaid expenses	Balance sheet	Current assets
190	xpr	Pension and retirement expense	Income statement	Operating expenses
191	xrent	Rental expense	Income statement	Operating expenses
192	xsga	Selling, general and administrative expense	Income statement	Operating expenses

This table reports the input variables used in our machine learning models. We also report the Compustat description, the financial statement group as well as the financial statement component group we assign to the respective variable. EBITDA denotes earnings before interest, taxes, depreciation and amortization. EBIT denotes earnings before interest and taxes. EBT denotes earnings before taxes. We scale all variables by common shares outstanding.

Appendix C Implied Cost of Capital Models

Table C.1: Implied cost of capital models

Name	Model/Description
GLS	$P_t = B_t + \Sigma_{\tau=1}^{11} \frac{\mathbb{E}_t[(ROE_{t+\tau} - ICC_{GLS}) \cdot B_{t+\tau-1}]}{(1+ICC_{GLS})^{\tau}} + \frac{\mathbb{E}_t[ROE_{t+12} - ICC_{GLS}) \cdot B_{t+11}]}{(1+ICC_{GLS})^{11} \cdot ICC_{GLS}}$ This model is given by Gebhardt et al. (2001). P_t denotes the stock price as of the estimation date in t, ICC_{GLS} denotes the implied cost of capital (ICC), B_t denotes the book value of equity per share in t and ROE_t is the return on equity in t. B_t is calculated using the clean surplus relation following Hou et al. (2012). ROE_t is calculated by dividing earnings per share (forecasts) E_t by B_{t-1}. For $ROE_{t+\tau}$ up to $\tau = 3$ we use the respective models earnings per share forecast. Afterwards, we assume $ROE_{t+\tau}$ to revert to the historical industry median by $\tau = 11$ (e.g., Hou et al., 2012). The industry median of ROE is derived using 10 years of data while excluding loss firms (e.g., Gebhardt et al., 2001). We expect ROE to be constant after $\tau = 11$.
CT	$P_t = B_t + \sum_{\tau=1}^{5} \frac{\mathbb{E}_t[(ROE_{t+\tau} - ICC_{CT}) \cdot B_{t+\tau-1}]}{(1+ICC_{CT})^{\tau}} + \frac{\mathbb{E}_t[(ROE_{t+5} - ICC_{CT}) \cdot B_{t+4}] \cdot (1+g)}{(1+ICC_{CT})^{5} \cdot (ICC_{CT} - g)}$ This model is given by Claus and Thomas (2001). P_t denotes the stock price as of the estimation date in t, ICC_{CT} denotes the implied cost of capital (ICC), B_t denotes the book value of equity per share in t, ROE_t is the return on equity in t and g is the perpetuity growth rate. B_t is calculated using the clean surplus relation following Hou et al. (2012). ROE_t is calculated by dividing earnings per share (forecasts) E_t by B_{t-1}. g is calculated as the current risk-free rate minus 3% (e.g., Hou et al., 2012).
OJ	$P_t = \frac{\mathbb{E}_t[E_{t+1}] \cdot (g_{st} - (\gamma - 1))}{(R-A)^{-} A^2}$, with $A = 0.5((\gamma - 1)\frac{\mathbb{E}_t[E_{t+1}] \cdot payout}{P_t})$, $g_{st} = 0.5(\frac{\mathbb{E}_t[E_{t+3}] - \mathbb{E}_t[E_{t+2}]}{\mathbb{E}_t[E_{t+2}]} - \frac{\mathbb{E}_t[E_{t+5}] - \mathbb{E}_t[E_{t+4}]}{\mathbb{E}_t[E_{t+4}]})$ This model is given by Ohlson and Juettner-Nauroth (2005). P_t denotes the stock price as of the estimation date in t, ICC_{OJ} denotes the implied cost of capital (ICC), $E_{t+\tau}$ is the earnings forecast for $t + \tau$, g_{st} is the short-term growth rate, γ is the perpetual growth rate and *payout* is the current payout ratio. g_{st} is calculated as the mean of forecasted earnings growth in $\tau = 3$ and $\tau = 5$ (e.g., Hou et al., 2012). γ is the current risk-free rate minus 3% (e.g., Hou et al., 2012). *payout* is calculated as dividends divided by earnings for profit firms and as dividends divided by $0.06 \cdot$ total assets for loss firms (e.g., Hou et al. (2012)).
MPEG	$P_t = \frac{\mathbb{E}_t[E_{t+2}] + (ICC_{MPEG} \cdot payout - 1) \cdot \mathbb{E}_t[E_{t+1}]}{ICC_{MPEG}^2}$ This model is given by Easton (2004). P_t denotes the stock price as of the estimation date in t, ICC_{MPEG} denotes the implied cost of capital (ICC) and $E_{t+\tau}$ denotes the earnings per share forecast for $t + \tau$. *payout* is derived as dividends divided by earnings for profit firms and as dividends divided by $0.06 \cdot$ total assets for loss firms (e.g., Hou et al., 2012).
GG	$P_t = \frac{\mathbb{E}_t[E_{t+1}]}{ICC_{GG}}$ This model is given by Gordon and Gordon (1997). P_t denotes the stock price as of the estimation date in t, ICC_{GG} denotes the implied cost of capital (ICC) and E_{t+1} denotes the earnings per share forecast for $t + 1$.

This table gives implied cost of capital (ICC) models that we base our composite ICC on. For simplicity, we drop the firm index i. The composite ICC that we use is derived as the average of these ICC.