APPENDIX

Appendix A. Additional Details on Supervised Learning Techniques and Forecast Formation

A.1 Supervised Learning Techniques

This section provides a more detailed description of the supervised learning techniques we explore.

Penalized linear estimators. The first set of estimators we explore falls within the class of penalized linear estimators. Given we are interested in capturing non-linearities, we include all two-way interactions (including squares) between variables in X_{it} as features when using penalized linear estimators. This results in approximately 8,000 features in each year. Penalized linear estimators consists in finding the linear combination of these features that minimizes MSE while minimizes over-fitting out of sample.

The first three commonly-used penalized linear estimators we explore are Lasso, Ridge, and Elastic Net, which are defined as follows for a vector of explanatory variables, X_{it}:

$$\mathcal{L}(\beta, \alpha_1, \alpha_2) \equiv \sum_i \left[\left(EPS_{it+h} - X_{it}'\beta \right)^2 \right] + \alpha_1 ||\beta||_1 + \alpha_2 ||\beta||_2,$$

$$\hat{\beta}^{\text{Lasso}} \equiv \arg\min_{\beta} \mathcal{L}(\beta, \alpha_1, 0), \quad \hat{\beta}^{\text{Ridge}}(\alpha_2) \equiv \arg\min_{\beta} \mathcal{L}(\beta, 0, \alpha_2),$$

$$\hat{\beta}^{\text{Elastic net}} \equiv \arg\min_{\beta} \mathcal{L}(\beta, \alpha_1, \alpha_2).$$

As evident from these definitions, the only difference between these estimators and OLS is the introduction of penalty terms on the norm of the coefficients, which are designed to improve stability. In order to choose the "hyper-parameters" α_1 and α_2, we use cross-validation on the training set, detailed in Appendix A.2. Intuitively, cross-validation consists of breaking up the training sample into smaller datasets, fitting models on these smaller datasets, and examining which values of the hyperparameters generate the best performance on the other parts of the training set. Importantly, cross-validation is done entirely using the training set to avoid introducing any look-ahead bias.

Which estimator will perform best depends on whether the true nature of the data is sparse or not. Under (approximate) sparsity conditions, Lasso can be shown to approximate an unknown function with asymptotically vanishing error (Belloni et al. 2011). This is because the L^1 norm will mechanically force several coefficients to zero, thereby revealing the true sparsity of the data. In contrast, Ridge tends to perform better in settings where there is minimal sparsity (Abadie and Kasy 2020). Recent evidence applying these techniques to estimate stochastic discount factors finds less sparse methods (e.g. Ridge) tend to perform best (in terms of minimizing pricing errors), suggesting SDFs are likely high-dimensional (Gu et al. 2018; Kozak et al. 2020; Bryzgalova et al. 2020).

We also explore two variations on Lasso: Post-Lasso and Iterative-Lasso. $\hat{\beta}^{\text{Post-Lasso}}$ consists of running an OLS regression using only variables with non-zero coefficients in the Lasso estimation. Post-Lasso has

been shown to improve performance, given that Lasso suffers from substantial regularization bias (Chernozhukov et al. 2016). Iterative-Lasso is calculated by solving the same optimization problem as Lasso, but an alternative iterative method (described in Appendix A.2) is used to estimate the penalty parameter instead of cross-validation. This iterative method has been shown to generate near-oracle rates of convergence (Belloni et al. 2011). Additionally, using this iterative method for the penalty choice has the added benefit of substantially reducing the likelihood of any over-fitting because the penalty parameter is chosen without regard to model performance on the training set (which is not the case for cross-validation).

Tree-based methods. We also consider two tree-based methods: Random Forests (RF) and Gradient-Boosted Trees (GBT). The building block of tree-based estimators are regression trees, which are nonparametric (unlike penalized linear estimators) regression estimators designed to capture arbitrary non-linearities among the variables in X_{it}.

We first describe regression trees, which are "grown" in sequential steps to approximate a function. The tree begins with an initial node containing all observations. Next, this initial node is split into two nodes: observations with $x_{it} < c$ and $x_{it} \geq c$. To make this split, the econometrician chooses the variable $x_{it} \in X_{it}$ and c to minimize MSE. This process of splitting based on a chosen covariate and value continues using the two new subsamples until a terminal criterion is satisfied (e.g. upper bound on the number of observations in each terminal node or the number of splits). The final regression values are then the averages of the outcome variable across all of the observations remaining in each of the terminal nodes.

The process of growing a regression tree immediately illustrates the potential problem with them: they are likely to overfit (i.e. they have high prediction variance), especially if they get extremely large. Without restrictions on the size of the tree, perfect fit in-sample fit could be achieved by having one observation in each terminal node, but this will perform terribly out-of-sample. To address this tendency to over-fit, many "ensemble" methods have been developed, which combine several decision trees with a form of regularization to make more accurate out-of-sample predictions. The two tree-based methods we consider, RF and GBT, are ensemble methods. The core idea behind RF and GBT is to grow a large number of uncorrelated trees and then average their predictions.

RF is constructed based on the intuition of bootstrapping. On each bootstrapped sample, a regression tree with a stopping criterion on the number of splits L with one adjustment - only a random subset of predictor variables are considered at each split.[24] These two steps are then repeated B times, generating B regression trees. Final predictions from the Random Forest are calculated by averaging predictions across the B regression trees. Averaging across many trees, which have different structures due to the randomness in the subset of predictor variables chosen, is the regularization in this method that limits over-fitting and reduces prediction variance. Similar to the penalized linear estimators, the two hyperparameters, $\{B, L\}$ can be chosen using cross-validation on the training set (see Appendix A.2 for details).

GBT starts by fitting a shallow tree of depth $d \in \{1, 2, 3\}$, and calculating the residuals from this regression tree. Then, a second shallow tree of depth d is fit on the residuals calculated from the first tree.[25] This

[24] If all variables are considered at each split, the procedure of forming many trees across bootstrapped samples is called bagging (i.e. bootstrap aggregation).

[25] Thinking of this procedure as operating on residuals from the trees conveys most of the intuition for why boosting works, but is a technically incorrect description. Gradient-Boosted Trees are a particular form

shallow tree is likely has terrible in-sample fit. To improve its fit, a second shallow tree of depth d is fit on the residuals calculated from the first tree. Predicted values are then formed by adding the predicted values from the two trees, shrinking the predicted values from the latter tree by a factor $\lambda \in (0,1)$ (regularization). This procedure is repeated B times, after which the predicted value will be a combination of the predicted value from the first tree and the predicted values of the $B-1$ trees scaled by λ. The sequential growing of trees on (pseudo-)residuals from the previous trees makes the trees less correlated, which is why averaging over trees limits over-fitting. This method has three hyperparameters, $\{B, d, \lambda\}$, which can be chosen using cross-validation on the training set (see Appendix A.2 for details).

A.2 Formation of Forecasts

This appendix describes the formation of our machine and machine + analyst forecasts, including details on the implementation of our machine learning estimators. For expositional simplicity, we present the procedure as pseudo-code. Additional details on implementation and cross-validation procedures are described at the end of this section.

Pseudo-code for penalized linear methods. To generate our machine using penalized linear methods at time t of EPS_{it+h}, the pseudo-code is as follows. For simplicity, denote X_{it}^{m} as the set of variables used in the machine forecast and X_{it}^{m+a} as the set of variables used in the machine + analyst forecast. As described in Section 1, these variables are defined as follows:

$$X_{it}^{m} = \{X_{it}^{Compustat}, X_{it-1}^{Compustat}, X_{it-2}^{Compustat}, X_{it}^{CRSP}, EPS_{it}\},$$
$$X_{it}^{m+a} = \{X_{it}^{Compustat}, X_{it-1}^{Compustat}, X_{it-2}^{Compustat}, X_{it}^{CRSP}, EPS_{it}, F_t EPS_{it+h}\},$$

where $x_{it}^{Compustat}$ and x_{it}^{CRSP} are defined in Table A1. We describe our procedure below for our machine forecasts, but an analogous procedure is used to form machine + analyst forecasts, replacing x_{it}^{m} with X_{it}^{m+a}.

1. Start with the above dataset that contains X_{it}^{m} and EPS_{it+h}, and $F_t EPS_{it+h}$ for each firm-year
2. Scale all elements of X_{it} by total assets at fiscal year end t, except for total assets, EPS_t, SIC codes, $PRCC\_F$, and X_{it}^{CRSP} (and $F_t EPS_{it+h}$, if using X_{it}^{m+a})
3. Replacing all missing values in $X_{it}^{Compustat}$ with 0[26]
4. Replacing all missing values in $X_{it-1}^{Compustat}$ with matching variable values form $X_{it}^{Compustat}$
5. Replacing all missing values in $X_{it-2}^{Compustat}$ with matching variable values form $X_{it-1}^{Compustat}$
6. Create all possible two way interactions between elements of X_{it}^{m}, including squares
7. Initialize $s = 1995$
   (a) Create a **training** dataset of observations indexed by i, s in the following set: $\{(i,t) : t \in \{s-5, ..., s-1\}\}$

of boosting, where trees are successively fit on pseudo-residuals instead of residuals. Pseudo-residuals are defined as the gradient of the objective function, evaluated at each data point.

[26] Filling missing values based on means or medians (as in Gu et al. 2018) does not affect our results.

(b) Create a **test** dataset of obserations indexed by i, t in the following set: $\{(i,t) : t = s\}$

(c) Trim all independent variables in the **training** dataset based on 5 times the interquartile range

(d) Trim all independent variables in the **test** dataset based on 5 times the interquartile range, with the interquartile range *calculated from the training set*

(e) Standardize all independent variables in the **training** set to have zero mean and unit variance

(f) Standardize all independent variables in the **test** based on means and variances *calculated from the training set*

(g) Fit a machine learning estimator that is one of the following on the training set, using cross-validation described at the end of this section:

- Lasso
- Ridge
- Elastic Net

(h) Generate forecasts on the test set. Calculating the MSE of these forecasts yields the MSEs for our three forecasts for year s.

(i) Stop if $s = 2020$, otherwise set $s = s + 1$ and continue back to (a)

Pseudo-code for tree-based methods. To generate our machine and machine + analyst forecasts using tree-based methods at time t of EPS_{it+h}, the pseudo-code is as follows.

1. Start with the above dataset that contains X_{it}^{m} and EPS_{it+h}, and F_tEPS_{it+h} for each firm-year
2. Scale all elements of $xXit$ by total assets at fiscal year end t, except for total assets, EPS_t, SIC codes, $PRCC\_F$, and X_{it}^{CRSP} (and F_tEPS_{it+h}, if using X_{it}^{m+a})
3. Replacing all missing values in $X_{it}^{Compustat}$ with 0
4. Replacing all missing values in $X_{it-1}^{Compustat}$ with matching variable values form $X_{it}^{Compustat}$
5. Replacing all missing values in $X_{it-2}^{Compustat}$ with matching variable values form $X_{it-1}^{Compustat}$
6. Initialize $s = 1995$

   (a) Create a **training** dataset of observations indexed by i, s in the following set: $\{(i,t) : t \in \{s-5, ..., s-1\}\}$

   (b) Create a **test** dataset of obserations indexed by i, t in the following set: $\{(i,t) : t = s\}$

   (c) Trim all independent variables in the **training** dataset based on 5 times the interquartile range

   (d) Trim all independent variables in the **test** dataset based on 5 times the interquartile range, with the interquartile range *calculated from the training set*

   (e) Standardize all independent variables in the **training** set to have zero mean and unit variance

   (f) Standardize all independent variables in the **test** based on means and variances *calculated from the training set*

(g) Fit a machine learning estimator that is one of the following on the training set, using cross-validation described at the end of this section:

- Random Forest
- Gradient-Boosted Trees

(h) Generate forecasts on the test set. Calculating the MSE of these forecasts yields the MSEs for our three forecasts for year s.

(i) Stop if $s = 2020$, otherwise set $s = s + 1$ and continue back to (a)

Cross-validation and implementation details by estimator. We use the following cross-validation and implementation procedures for each machine learning algorithm on our training sets for each model. All procedures are implemented using the `sklearn` package in `Python 3.6`. We use default inputs to all `sklearn` functions mentioned below, unless otherwise specified.

- Lasso: We use 5-fold cross-validation on the training set, implemented using the `LassoCV` function in `sklearn`. We search over a grid of `alphas` $\in -[10^{-6}, 10^{6}]$.
- Post-Lasso: Same cross-validation technique as for Lasso. The only difference from Lasso is that an OLS regression is run using the set of variables that were not penalized to zero to generate the final model; hyperparameters are the same.
- Iterative Lasso: To choose the $\mathcal{L}^1$ penalty parameter, we use the iterative procedure described in the Appendix of Belloni et al. (2011). Specifically, we use their algorithm that generates an "X-dependent" penalty level, using 1000 simulations and the parameters as follows: $c = 1.1$, $\gamma = 0.5$, $\phi = 0.1$, $\nu = 10^{-8}$, $K = 20$. As Belloni et al. (2011) show, Iterative Lasso obtains a near-oracle rate of convergence that suggests it might perform better than standard Lasso implemented with cross-validation.
- Ridge: We use generalized cross-validation to choose the $\mathcal{L}^2$ penalty parameter, implemented using the `RidgeCV` function in `sklearn`. Generalized cross-validation is an form of "leave one out cross-validation", which is generally infeasible. However, in the special case of Ridge, this penalty parameter can be estimated directly in the data (see Section 4.1 van Wieringen 2020).
- Elastic Net: We use 5-fold cross-validation on the training set, implemented using the `ElasticNetCV` function in `sklearn`. We search over a grid of the parameter `l1_ratio` $\in [0.1, 0.99]$, which corresponds to the ratio of the $\mathcal{L}^1$ to $\mathcal{L}^2$ penalty parameters.
- Random Forest: We use 5-fold cross-validation on the training set, implemented using the `GridSearchCV` function for `RandomForestRegressor` in `sklearn`. We set `n_estimators` to 1000, corresponding to the number of decision trees in the ensemble, and search over the following grid for each parameter: `max_depth` $\in [4, 8]$, `max_features` $\in [0.3, 1]$, `min_samples_leaf` $\in [1, 5]$, and `min_samples_splits` $\in [2, 10]$. We use bootstrap samples for each decision tree. These parameter choices are similar to Gu et al. (2018) and Hansen and Thimsen (2020).
- Gradient-Boosted Trees: We use 5-fold cross-validation on the training set, implemented using the `GridSearchCV` function for `GradientBoostingRegressor` in `sklearn`. We search over the following grid for each parameter: `n_estimators` $\in [500, 10000]$, `max_depth` $\in [1, 3]$, and `learning_rate` $\in [0.001, 0.1]$. These parameter choices are similar to Gu et al. (2018).

Appendix B. Details on Conditional Estimation

This section discusses the estimation and identification of conditional versions of our decomposition in Section 3.3 that is used to produce the results in Sections 3.5 and 5.2. This approach is designed to impose minimal assumptions beyond those required for identification and estimation in Section 3.3.

Identification of conditional parameters. Let $\mathcal{F}$ be a conditioning event, such $\{\text{volatility}_{it} = 15\%\}$ or $\{\text{number of analysts}_{it} = 3\}$. Our objective is to generate conditional estimates of our three parameters, defined as follows: $\Theta(\mathcal{F}) \equiv E\left(Z_{it}^2|\mathcal{F}\right)$, $\Delta(\mathcal{F}) \equiv E\left[\left(x_{it}^h - g_h(X_{it})\right)^2|\mathcal{F}\right]$, and $\Sigma_\eta(\mathcal{F}) \equiv E\left(\eta_{it}^2|\mathcal{F}\right)$. The following proposition shows our conditional moments are identified using the moments in the following proposition, under a set of identifying assumptions given in Assumption B1.

Assumption B1. *For a conditioning event , $\mathcal{F}$, the following are true:*

$$cov(Z_{it}, \eta_{it}|\mathcal{F}) = 0, \tag{13}$$
$$cov(\varepsilon_{it+1}, \eta_{it}|\mathcal{F}) = 0, \tag{14}$$
$$E(\eta_{it}|\mathcal{F}) = 0, \tag{15}$$
$$cov(Z_{it}, \varepsilon_{it+1}|\mathcal{F}) = 0. \tag{16}$$

Proposition B1. *Denote F_t^* and EPS_{it+1}^* as in Proposition 2. If Assumption B1 holds, then the bias, information advantage and noise conditional on an event $\mathcal{F}$ are given by the following three equations:*

$$\Delta(\mathcal{F}) = E\left[\left(E(F_t EPS_{it+1}|X_{it}) - E(EPS_{it+1}|X_{it})\right)^2|\mathcal{F}\right],$$
$$\Theta(\mathcal{F}) = cov(F_t^* EPS_{it+1}, EPS_{it+1}^*|\mathcal{F}) + E\left(F_t^* EPS_{it+1}|\mathcal{F}\right)^2,$$
$$\Sigma_\eta(\mathcal{F}) = var(F_t^* EPS_{it+1}|\mathcal{F}) - cov(F_t^* EPS_{it+1}, EPS_{it+1}^*|\mathcal{F}).$$

First, we describe the content of Assumption B1. Equations (13)-(15) can be justified by the definition of an expectation noise term, which should be mean zero, orthogonal to private information, and orthogonal to innovations regardless of whether we condition on time or firm characteristics (as we will do below). Equation (16) is the core identifying assumption for our methodology. Intuitively, it requires that the analyst be unable to forecast the innovation term conditional on the event $\mathcal{F}$ (which the analyst can't do unconditionally by the normalization in Section 3.3). Note that Equation (16) is implied by our assumptions in Section 3.3 if the conditioning event involves random variables that are subsets of X_{it} by the law of iterated expectations.[27] For this reason, we believe it is a reasonable assumption, given our conditioning variables are reasonably well-spanned by X_{it}.

The results in Proposition B1 are a natural extension of those in Proposition 2, with two differences. First, an additional term shows up in the equation for $\Theta(\mathcal{F})$: $E\left(F_t^* EPS_{it+1}|\mathcal{F}\right)^2$. This term reflects the fact that the analyst's private signal, Z_{it}, may not be mean zero after conditioning on $\mathcal{F}$, despite being mean zero unconditionally (which is just a normalization). Secondly, all moments are now conditional

[27] Take a set of variables $Y_{it} \subset X_{it}$. Then $E(Z_{it}\varepsilon_{it+1}|Y_{it}) = E(Z_{it}E(\varepsilon_{it+1}|Z_{it}, Y_{it})) = 0$, and $E(Z_{it}|Y_{it})E(\varepsilon_{it+1}|Y_{it}) = 0$.

rather than unconditional. Thus, we need to estimate five conditional expectations to identify our three conditional parameters: $E(EPS_{it+1}|\mathcal{F})$, $E(F_t^* EPS_{it+1}|\mathcal{F})$, $E(EPS_{it+1}F_t^* EPS_{it+1}|\mathcal{F})$, $E(F_t^* EPS_{it+1}^2|\mathcal{F})$, and $E\left[\left(E(F_t EPS_{it+1}|X_{it}) - E(EPS_{it+1}|X_{it})|\right)^2 \mathcal{F}\right]$. We next describe how we do this.

Estimation of conditional parameters. We use the following procedure to estimate conditional expectations. First, for each firm-year observation, we use the following conditioning variables, $\mathcal{F}$:

- Section 3.5 – evenly-spaced bins of the inverse of the number of distinct analysts that issue forecasts used in forming our consensus forecast;
- Section 5.2 – evenly-spaced bins of equity volatility calculated using monthly data over the five-years prior to the most recent fiscal year-end, taking into account delisting returns.

We choose to use bins so that our estimation strategy perfectly approximates the true conditional expectation function in population.

Next, for each conditional expectation, $E(Y_{it}|\mathcal{F})$, we regress Y_{it} onto the indicators for the bins described above. Finally, we use our estimation to impute estimates of Θ, Δ and Σ_η at the firm-year level using Proposition B1.

Appendix C. Additional Tables

Table A1. Variables Collected from Compustat and CRSP

This table lists the set of variables that are collected from COMPUSTAT and CRSP to form our statistical forecast. As described in Section 1.3, we use up to two lags of each of these variables. In our penalized linear estimators, we use all two-way interactions. See Appendix A for a detailed discussion of how we use these variables.

Variable	Included in	Required non-missing?
Panel A: Collected from Compustat		
Total assets	$X_{it}^{Compustat}$	✓
Total liabilities	$X_{it}^{Compustat}$	✓
Revenue	$X_{it}^{Compustat}$	✓
SG&A expense	$X_{it}^{Compustat}$	
R&D expense	$X_{it}^{Compustat}$	
Cost of goods sold	$X_{it}^{Compustat}$	✓
Current assets	$X_{it}^{Compustat}$	
Current liabilities	$X_{it}^{Compustat}$	
Cash	$X_{it}^{Compustat}$	
Cash and short-term investments	$X_{it}^{Compustat}$	
Income tax expense	$X_{it}^{Compustat}$	
Total long-term debt	$X_{it}^{Compustat}$	
Total long-term debt due within one-year	$X_{it}^{Compustat}$	
Debt in current liabilities	$X_{it}^{Compustat}$	
Depreciation expense	$X_{it}^{Compustat}$	
EBIT	$X_{it}^{Compustat}$	
EBITDA	$X_{it}^{Compustat}$	✓
Interest expense	$X_{it}^{Compustat}$	
Interest paid	$X_{it}^{Compustat}$	
Capital expenditures	$X_{it}^{Compustat}$	
Income tax payable	$X_{it}^{Compustat}$	
Income tax expense	$X_{it}^{Compustat}$	
Total income tax	$X_{it}^{Compustat}$	
Net income	$X_{it}^{Compustat}$	
Common dividends	$X_{it}^{Compustat}$	
Purchase of common and preferred stock	$X_{it}^{Compustat}$	
Sale of common and preferred stock	$X_{it}^{Compustat}$	
Subordinated debt	$X_{it}^{Compustat}$	
Gross profit	$X_{it}^{Compustat}$	✓
Operating cash flow	$X_{it}^{Compustat}$	✓
Common shares outstanding	$X_{it}^{Compustat}$	
Stock price at fiscal year end	$X_{it}^{Compustat}$	✓

Table A1. Variables Collected from Compustat and CRSP (continued)

Extraordinary items	$X_{it}^{Compustat}$	
Special items	$X_{it}^{Compustat}$	
Acquisitions	$X_{it}^{Compustat}$	
Capitalized leases (due within two-years)	$X_{it}^{Compustat}$	
Capitalized leases (due within three-years)	$X_{it}^{Compustat}$	
Capitalized leases (due within four years)	$X_{it}^{Compustat}$	
Capitalized leases (due within five years)	$X_{it}^{Compustat}$	
Common ESOP obligation	$X_{it}^{Compustat}$	
Goodwill	$X_{it}^{Compustat}$	
Interest and related income (total)	$X_{it}^{Compustat}$	
Total intangible assets	$X_{it}^{Compustat}$	
Marketable securities adjustment	$X_{it}^{Compustat}$	
Net PPE	$X_{it}^{Compustat}$	✓
Nonoperating income	$X_{it}^{Compustat}$	
Tax loss carryforward	$X_{it}^{Compustat}$	
Pension and retirement expense	$X_{it}^{Compustat}$	
Preferred stock value	$X_{it}^{Compustat}$	

Panel B: Collected from CRSP		
SIC 2-digit industry code dummies	X_{it}^{CRSP}	✓
Return over prior month to fiscal year end t	X_{it}^{CRSP}	✓
Return over year prior to fiscal year end t, excluding last month	X_{it}^{CRSP}	✓
Market capitalization at the end of year t	X_{it}^{CRSP}	✓

Table A2. Table of Machine MSEs by Year (One-Year Horizon Forecasts)

This table contains the mean squared error of various machine forecasts. All mean squared errors are normalized by the average squared EPS across all firm-years within each year. The final sample used in this figure contains 47,542 firm-year observations. See Table 1 for additional notes and information.

Year	Analyst	Random Walk	OLS	Lasso	Post-Lasso	Iterative Lasso	Ridge	Elastic Net	Random Forest	Gradient-Boosted Trees
1995	13.66%	17.55%	15.52%	16.05%	16.25%	16.56%	17.47%	16.07%	15.28%	15.27%
1996	17.86%	19.22%	17.81%	15.44%	15.63%	16.16%	15.98%	15.44%	15.24%	15.41%
1997	12.04%	15.68%	17.76%	13.76%	13.98%	14.89%	14.73%	13.76%	13.38%	13.88%
1998	19.92%	22.2%	25.03%	19.77%	20.01%	20.79%	20.8%	19.71%	20.0%	20.21%
1999	16.56%	23.44%	27.87%	19.99%	20.35%	20.36%	21.64%	20.1%	20.29%	20.91%
2000	15.33%	28.23%	40.46%	24.2%	24.05%	25.3%	27.29%	24.38%	24.06%	24.0%
2001	25.15%	34.01%	53.7%	29.26%	30.4%	29.86%	33.04%	29.28%	29.67%	31.07%
2002	13.1%	35.75%	46.38%	25.17%	25.88%	27.52%	26.16%	25.17%	25.12%	25.35%
2003	11.8%	23.39%	23.58%	17.82%	20.15%	17.15%	18.87%	17.82%	16.96%	17.98%
2004	9.26%	22.54%	20.16%	17.51%	18.64%	18.56%	19.31%	17.51%	18.21%	18.16%
2005	8.77%	17.61%	16.18%	14.51%	15.09%	15.19%	16.69%	14.51%	14.93%	15.52%
2006	8.33%	14.76%	15.48%	13.05%	13.59%	13.47%	14.83%	13.05%	13.12%	13.29%
2007	7.74%	15.14%	15.78%	11.74%	11.98%	12.04%	12.64%	11.74%	11.47%	12.05%
2008	12.19%	21.57%	20.38%	16.0%	15.92%	16.78%	16.4%	15.87%	15.42%	15.31%
2009	17.55%	53.56%	33.82%	26.78%	27.94%	27.59%	28.25%	26.76%	25.21%	26.16%
2010	8.19%	26.12%	24.39%	16.55%	17.26%	17.35%	18.03%	16.46%	16.12%	16.7%
2011	5.53%	12.83%	11.85%	9.75%	10.14%	10.35%	10.96%	9.88%	10.13%	9.85%
2012	6.51%	11.36%	11.47%	9.29%	9.81%	9.75%	9.99%	9.3%	9.01%	9.13%
2013	6.08%	10.1%	10.58%	9.43%	10.13%	9.56%	10.65%	9.48%	8.91%	9.34%
2014	6.34%	11.16%	12.16%	9.37%	10.02%	9.57%	10.58%	9.36%	9.28%	9.49%
2015	7.94%	18.75%	24.14%	17.11%	17.35%	17.84%	18.28%	17.19%	17.15%	17.38%
2016	5.65%	12.03%	18.86%	10.24%	10.34%	10.85%	11.15%	10.24%	10.12%	9.82%
2017	5.14%	10.69%	13.75%	8.39%	8.67%	8.9%	9.55%	8.39%	8.59%	8.72%
2018	5.57%	13.4%	14.71%	10.61%	11.02%	11.12%	11.61%	10.62%	11.21%	11.25%
2019	4.38%	10.15%	13.06%	8.34%	8.7%	8.65%	9.33%	8.31%	8.55%	8.93%
2020	4.32%	5.38%	5.59%	4.88%	5.6%	4.53%	5.22%	4.91%	4.13%	4.34%

Table A3. Table of Machine + Analyst MSEs by Year (One-Year Horizon Forecasts)

This table contains the mean squared error of various machine + analyst forecasts. All mean squared errors are normalized by the average squared EPS across all firm-years within each year. The final sample used in this figure contains 47,542 firm-year observations. See Table 1 for additional notes and information.

Year	Analyst	Lasso	Post-Lasso	Iterative Lasso	Ridge	Elastic Net	Random Forest	Gradient-Boosted Trees
1995	13.66%	11.79%	11.98%	12.12%	12.82%	11.81%	11.62%	11.63%
1996	17.86%	13.23%	13.25%	13.81%	13.41%	13.23%	13.23%	13.18%
1997	12.04%	10.0%	10.34%	10.48%	11.08%	10.02%	10.1%	10.34%
1998	19.92%	14.9%	15.18%	15.49%	16.08%	14.89%	15.11%	15.12%
1999	16.56%	13.7%	14.23%	13.75%	15.42%	13.76%	14.53%	14.37%
2000	15.33%	13.41%	14.0%	13.86%	17.5%	13.73%	15.64%	15.48%
2001	25.15%	18.62%	19.22%	19.68%	22.84%	18.63%	18.21%	18.4%
2002	13.1%	11.46%	12.08%	11.61%	13.9%	11.46%	11.95%	11.47%
2003	11.8%	11.4%	12.28%	11.25%	12.57%	11.4%	11.09%	11.06%
2004	9.26%	10.47%	11.13%	10.77%	11.91%	10.51%	11.53%	11.54%
2005	8.77%	9.08%	9.25%	9.35%	10.68%	9.08%	10.06%	10.61%
2006	8.33%	8.36%	8.68%	8.57%	9.74%	8.36%	9.28%	9.42%
2007	7.74%	7.0%	7.03%	7.12%	7.8%	7.0%	7.02%	7.09%
2008	12.19%	10.88%	10.87%	11.0%	11.7%	10.86%	11.18%	11.01%
2009	17.55%	13.81%	14.07%	13.71%	15.51%	13.81%	13.7%	13.25%
2010	8.19%	8.94%	9.06%	9.14%	10.28%	8.96%	9.59%	9.31%
2011	5.53%	5.33%	5.56%	5.56%	6.27%	5.35%	5.28%	5.44%
2012	6.51%	5.9%	6.1%	5.87%	6.51%	5.9%	5.83%	5.85%
2013	6.08%	5.83%	6.06%	5.84%	6.96%	5.86%	5.87%	5.89%
2014	6.34%	5.85%	5.89%	6.0%	6.84%	5.85%	5.93%	5.94%
2015	7.94%	7.49%	7.55%	7.69%	9.89%	7.54%	8.47%	7.97%
2016	5.65%	5.47%	5.75%	5.48%	7.0%	5.47%	5.76%	5.62%
2017	5.14%	4.92%	4.96%	5.09%	5.6%	4.93%	5.05%	5.25%
2018	5.57%	5.56%	5.74%	5.72%	6.65%	5.56%	6.35%	6.48%
2019	4.38%	3.92%	4.24%	3.95%	5.03%	3.92%	4.09%	4.1%
2020	4.32%	3.96%	4.15%	3.8%	4.82%	3.96%	3.73%	3.45%

Table A4. Table of Machine MSEs by Year (Two-Year Horizon Forecasts)

This table contains the mean squared error of various machine forecasts. All mean squared errors are normalized by the average squared EPS across all firm-years within each year. The final sample used in this figure contains 39,973 firm-year observations. See Table 1 for additional notes and information.

Year	Analyst	Random Walk	OLS	Lasso	Post-Lasso	Iterative Lasso	Ridge	Elastic Net	Random Forest	Gradient-Boosted Trees
1996	45.81%	31.02%	26.98%	27.41%	29.55%	28.08%	27.98%	27.34%	25.05%	25.6%
1997	38.86%	27.42%	24.86%	23.98%	25.78%	24.58%	24.2%	23.88%	22.12%	23.24%
1998	53.08%	36.57%	36.73%	32.16%	32.87%	34.17%	32.17%	32.15%	30.36%	31.82%
1999	68.7%	40.1%	39.49%	34.08%	34.8%	36.47%	34.09%	34.09%	32.36%	33.87%
2000	49.77%	46.1%	43.19%	38.46%	39.63%	39.45%	39.54%	38.55%	36.64%	37.86%
2001	87.43%	54.66%	66.21%	47.11%	48.87%	48.17%	49.7%	47.45%	46.09%	47.81%
2002	93.91%	64.83%	54.58%	44.05%	43.87%	47.6%	44.06%	44.05%	39.47%	42.2%
2003	46.95%	55.21%	45.22%	32.39%	36.6%	34.24%	33.06%	32.27%	31.4%	32.12%
2004	21.97%	35.56%	30.28%	29.29%	32.64%	29.19%	29.72%	29.26%	27.1%	27.61%
2005	23.03%	34.12%	25.74%	27.58%	29.16%	29.7%	28.54%	27.68%	28.0%	28.11%
2006	21.12%	27.62%	24.15%	23.98%	25.45%	25.8%	25.5%	24.24%	23.97%	24.01%
2007	22.02%	24.87%	23.95%	21.27%	22.73%	21.83%	22.17%	21.27%	20.46%	21.76%
2008	34.42%	36.87%	35.06%	30.47%	30.98%	31.4%	30.99%	30.38%	29.02%	29.78%
2009	79.45%	62.96%	53.81%	53.06%	52.61%	54.57%	50.41%	52.95%	48.72%	52.32%
2010	27.84%	48.41%	33.83%	27.4%	29.24%	25.96%	27.6%	26.98%	24.03%	29.64%
2011	14.44%	35.1%	25.54%	23.99%	26.91%	25.5%	25.23%	24.01%	23.3%	23.27%
2012	18.4%	22.07%	18.71%	17.11%	18.31%	18.63%	17.97%	17.22%	17.02%	17.27%
2013	18.38%	18.81%	16.83%	16.46%	17.54%	17.51%	17.55%	16.46%	15.66%	16.13%
2014	13.73%	15.12%	14.72%	14.37%	15.78%	15.12%	15.47%	14.37%	13.5%	14.31%
2015	30.9%	27.15%	30.82%	26.31%	27.04%	26.47%	26.55%	26.21%	23.36%	25.38%
2016	22.79%	31.01%	31.81%	26.24%	26.44%	27.89%	26.32%	26.14%	22.7%	22.68%
2017	14.67%	19.29%	22.63%	16.46%	17.94%	16.9%	16.89%	16.41%	15.14%	15.79%
2018	12.22%	23.58%	19.46%	17.81%	19.64%	18.88%	18.77%	17.92%	17.49%	18.07%
2019	14.63%	19.56%	16.71%	14.09%	15.5%	14.96%	14.44%	14.11%	13.56%	13.78%
2020	13.43%	12.72%	12.56%	11.84%	12.75%	11.4%	11.63%	11.83%	11.13%	11.33%

Table A5. Table of Machine + Analyst MSEs by Year (Two-Year Horizon Forecasts)

This table contains the mean squared error of various machine + analyst forecasts. All mean squared errors are normalized by the average squared EPS across all firm-years within each year. The final sample used in this figure contains 39,973 firm-year observations. See Table 1 for additional notes and information.

Year	Analyst	Lasso	Post-Lasso	Iterative Lasso	Ridge	Elastic Net	Random Forest	Gradient-Boosted Trees
1996	45.81%	26.37%	27.42%	27.08%	26.5%	26.33%	24.13%	24.81%
1997	38.86%	22.47%	23.82%	23.3%	22.77%	22.47%	21.18%	21.99%
1998	53.08%	30.36%	31.95%	32.38%	30.62%	30.36%	28.87%	30.29%
1999	68.7%	32.67%	33.53%	35.15%	32.57%	32.57%	31.43%	31.76%
2000	49.77%	35.46%	36.55%	36.69%	36.74%	35.47%	34.44%	34.53%
2001	87.43%	44.61%	44.75%	48.87%	47.64%	44.61%	43.62%	43.88%
2002	93.91%	39.11%	39.94%	46.0%	41.44%	39.11%	36.31%	38.33%
2003	46.95%	30.47%	34.67%	30.7%	30.35%	30.47%	29.38%	29.89%
2004	21.97%	26.74%	30.46%	26.58%	26.87%	26.73%	24.95%	25.47%
2005	23.03%	25.28%	27.09%	26.69%	25.88%	25.37%	26.07%	25.63%
2006	21.12%	22.42%	24.69%	22.94%	23.23%	22.46%	22.85%	23.16%
2007	22.02%	18.94%	20.07%	19.23%	19.6%	18.9%	18.73%	19.78%
2008	34.42%	27.68%	27.86%	28.04%	27.95%	27.56%	26.89%	27.15%
2009	79.45%	53.44%	52.99%	55.08%	49.84%	52.03%	49.25%	51.43%
2010	27.84%	25.93%	30.22%	22.42%	25.89%	26.12%	21.72%	27.38%
2011	14.44%	18.32%	21.13%	19.05%	19.81%	18.22%	18.86%	18.25%
2012	18.4%	14.87%	16.06%	15.49%	15.14%	14.78%	14.21%	15.0%
2013	18.38%	14.0%	14.99%	14.25%	14.73%	14.08%	13.8%	14.32%
2014	13.73%	11.97%	13.32%	12.5%	12.9%	12.04%	11.74%	12.13%
2015	30.9%	23.73%	23.66%	23.95%	23.79%	23.48%	22.11%	23.03%
2016	22.79%	17.78%	18.37%	19.02%	19.42%	18.11%	18.28%	19.18%
2017	14.67%	13.57%	15.29%	13.44%	14.5%	13.57%	13.15%	13.74%
2018	12.22%	14.3%	15.35%	14.74%	15.33%	14.29%	14.7%	14.68%
2019	14.63%	11.83%	12.33%	12.23%	12.09%	11.73%	11.63%	11.71%
2020	13.43%	10.48%	11.22%	10.14%	10.11%	10.47%	9.68%	10.35%

Table A6. Table of Machine MSEs by Year (Three-Year Horizon Forecasts)

This table contains the mean squared error of various machine forecasts. All mean squared errors are normalized by the average squared EPS across all firm-years within each year. The final sample used in this figure contains 10,831 firm-year observations. See Table 1 for additional notes and information.

Year	Analyst	Random Walk	OLS	Lasso	Post-Lasso	Iterative Lasso	Ridge	Elastic Net	Random Forest	Gradient-Boosted Trees
1997	44.78%	22.78%	17.88%	17.73%	19.25%	22.47%	18.49%	17.64%	15.53%	16.97%
1998	38.05%	20.64%	21.66%	19.65%	20.63%	21.12%	20.76%	19.71%	18.23%	19.11%
1999	89.82%	32.52%	37.87%	38.29%	41.37%	43.33%	38.97%	39.25%	37.29%	38.57%
2000	104.99%	53.77%	47.18%	47.82%	46.84%	50.52%	44.49%	45.64%	43.32%	44.68%
2001	164.49%	65.31%	84.9%	59.67%	63.17%	61.63%	63.16%	60.54%	57.43%	58.43%
2002	123.91%	52.82%	71.62%	47.78%	50.5%	50.16%	52.88%	46.74%	48.38%	50.57%
2003	148.13%	90.46%	73.9%	58.98%	63.2%	63.84%	56.6%	59.57%	55.11%	53.29%
2004	65.77%	50.71%	46.97%	39.41%	48.09%	42.5%	42.26%	39.48%	37.89%	42.0%
2005	27.96%	45.29%	32.54%	36.53%	36.37%	42.75%	40.7%	36.99%	34.1%	36.86%
2006	27.75%	40.69%	25.26%	29.45%	35.19%	37.17%	29.34%	29.66%	27.44%	27.19%
2007	26.61%	29.14%	24.31%	23.48%	25.97%	26.54%	25.39%	23.77%	22.0%	22.52%
2008	39.0%	34.08%	34.72%	29.96%	33.7%	30.82%	30.31%	29.56%	27.62%	28.73%
2009	94.1%	68.98%	68.78%	62.44%	65.36%	60.98%	56.42%	60.64%	54.14%	57.16%
2010	58.91%	41.99%	34.26%	29.21%	32.97%	29.82%	28.5%	28.71%	24.59%	28.07%
2011	26.19%	40.15%	25.59%	27.2%	31.54%	24.46%	27.08%	27.19%	22.49%	24.46%
2012	25.36%	35.58%	20.75%	22.13%	23.15%	24.61%	23.24%	22.2%	21.33%	22.86%
2013	33.55%	27.69%	22.42%	21.47%	22.02%	23.8%	22.19%	21.48%	20.42%	21.15%
2014	27.83%	21.54%	18.78%	20.48%	21.77%	21.53%	21.04%	20.51%	18.01%	19.27%
2015	36.54%	23.87%	23.27%	21.68%	23.16%	22.23%	21.72%	21.64%	18.8%	20.83%
2016	56.57%	37.03%	36.99%	34.62%	35.2%	35.04%	33.18%	34.3%	27.37%	28.1%
2017	32.41%	30.25%	28.49%	22.14%	22.2%	25.09%	22.03%	22.48%	19.32%	20.53%
2018	21.4%	27.91%	26.71%	23.51%	26.45%	22.43%	24.37%	23.74%	19.76%	22.23%
2019	17.48%	25.17%	17.73%	17.32%	19.14%	18.77%	17.58%	17.22%	15.54%	15.65%
2020	15.5%	18.06%	14.03%	13.88%	15.44%	14.88%	13.66%	13.75%	12.69%	12.95%

Table A7. Table of Machine + Analyst MSEs by Year (Three-Year Horizon Forecasts)

This table contains the mean squared error of various machine + analyst forecasts. All mean squared errors are normalized by the average squared EPS across all firm-years within each year. The final sample used in this figure contains 10,831 firm-year observations. See Table 1 for additional notes and information.

Year	Analyst	Lasso	Post-Lasso	Iterative Lasso	Ridge	Elastic Net	Random Forest	Gradient-Boosted Trees
1997	44.78%	16.2%	17.46%	20.29%	16.55%	16.2%	14.97%	15.62%
1998	38.05%	18.35%	19.16%	20.39%	19.41%	18.44%	16.72%	16.98%
1999	89.82%	37.47%	38.36%	42.52%	36.31%	37.48%	36.72%	39.69%
2000	104.99%	48.04%	47.97%	51.54%	44.64%	45.63%	41.05%	40.86%
2001	164.49%	64.68%	71.06%	63.98%	67.04%	65.4%	60.43%	66.09%
2002	123.91%	47.58%	50.58%	50.06%	52.73%	47.05%	48.81%	49.71%
2003	148.13%	58.6%	61.46%	63.69%	55.73%	59.53%	55.86%	50.76%
2004	65.77%	38.48%	48.77%	42.01%	40.27%	38.48%	36.28%	41.09%
2005	27.96%	36.6%	37.11%	41.97%	40.66%	36.98%	33.29%	36.62%
2006	27.75%	28.79%	35.61%	34.62%	29.89%	28.73%	27.18%	25.89%
2007	26.61%	21.82%	24.18%	24.5%	23.47%	21.99%	20.98%	21.73%
2008	39.0%	28.73%	32.3%	29.6%	28.74%	28.45%	26.92%	27.45%
2009	94.1%	62.24%	63.09%	61.27%	57.15%	60.4%	55.07%	58.21%
2010	58.91%	27.83%	31.16%	28.3%	27.27%	27.39%	24.03%	27.82%
2011	26.19%	26.37%	31.39%	22.75%	25.94%	26.18%	21.78%	25.48%
2012	25.36%	20.16%	22.58%	21.95%	20.98%	20.35%	19.68%	21.27%
2013	33.55%	20.03%	21.15%	21.54%	20.46%	20.03%	18.97%	19.8%
2014	27.83%	18.64%	21.33%	18.84%	18.8%	18.49%	16.6%	18.46%
2015	36.54%	21.31%	21.61%	21.82%	20.99%	21.3%	18.6%	19.45%
2016	56.57%	31.18%	31.34%	33.0%	31.36%	31.75%	26.22%	26.56%
2017	32.41%	19.78%	22.32%	22.02%	20.0%	19.78%	18.48%	20.88%
2018	21.4%	21.76%	24.78%	21.11%	22.95%	22.18%	19.33%	22.08%
2019	17.48%	15.77%	17.31%	16.95%	16.07%	15.77%	14.47%	14.61%
2020	15.5%	13.43%	15.19%	13.95%	12.92%	13.18%	12.33%	11.57%

Table A8. Validity of AR(1) Assumption

This table provides evidence consistent with the fact that an AR(1) model is a reasonable approximation to the data generating process used in the existing models presented in Section 4. The table below presents least-squares regressions on the sub-sample of firm-year observations for which we have analyst forecast at all three horizons.

	Dependent Variable			
	$F_t^m EPS_{it+2}$	$F_t^m EPS_{it+3}$	$F_t^m EPS_{it+3}$	$F_t^m EPS_{it+3}$
$F_t^m EPS_{it+1}$	0.928^{***}	0.848^{***}		-0.260^{***}
	(0.002)	(0.004)		(0.012)
$F_t^m EPS_{it+2}$			0.932^{***}	1.194^{***}
			(0.003)	(0.013)
Constant	0.006^{***}	0.013^{***}	0.007^{***}	0.006^{***}
	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Observations	10,831	10,831	10,831	10,831
R^2	0.934	0.783	0.873	0.878
Adjusted R^2	0.934	0.783	0.873	0.878

Appendix D. Model Extension: Bias on Private Information

In this appendix, we extend our baseline model presented in Section 3 to allow for bias in processing soft information by the analyst and estimate this bias. The layout of this section parallels that of Section 3 for ease of comparison. In sum, we estimate negligible bias on soft information, which is why we leave it out of our analysis in the main text.

D.1 Setup

As in Section 3.1, we decompose EPS_{it+h} as:

$$EPS_{it+h} = x^h_{it} + z^h_{it} + \varepsilon_{it+h}.$$

Unlike in our baseline model, we model the forecasts of the individual analysts forecasting EPS_{it+h} directly. As will become clear, variation in the number of analysts forecasting EPS for firm i in year $t+h$, denoted N^h_{it}, is crucial for our identification strategy. All analysts observe both X_{it} and Z_{it} as in our baseline model[28], but allow analysts to be biased with respect to public *and* private information. Specifically, we assume the forecast of an analyst $j \in \{1, ..., N^h_{it}\}$ is generated as follows:

$$F^j_t EPS_{it+h} = g^j_h(X_{it}) + \alpha^j_h z^h_{it} + \eta^{hj}_{it}. \tag{17}$$

In eq. (17), there are three possible deviations from full-information rationality: $g^j_h(X_{it}) \neq x^h_{it}$, $\alpha^j_h \neq 1$, and $\eta^j_{it} \neq 0$. Our estimation results in Section 3.4 suggest the first and third deviations are sizeable (on average), but smaller than the size of the analyst information advantage. In this section, we attempt to quantify the size of the second possible deviation from full-information rationality: $\alpha^j_h \neq 1$.

Aggregating individual analyst forecasts into the consensus forecast (which we still refer to as "the analyst forecast"), we obtain a version of eq. (5) that allows for bias in processing soft information:

$$F_t EPS_{it+h} = g_h(X_{it}) + \alpha_h z^h_{it} + \underbrace{\frac{1}{N^h_{it}} \sum_{j=1}^{N^h_{it}} \eta^{hj}_{it}}_{\equiv \eta^h_{it}}, \tag{18}$$

where $g_h(X_{it}) \equiv \frac{1}{N^h_{it}} \sum_{j=1}^{N^h_{it}} g^j_h(X_{it})$ is the average subjective weight on public information and $\alpha_h = \frac{1}{N^h_{it}} \sum_{j=1}^{N^h_{it}} \alpha^j_h$ is the average subjective weight on soft information. We define the analyst information advantage, bias, and noise as in Definition 1. The expectation noise in the consensus forecast, denoted Σ, is now related to the expectation noise in the individual analyst forecasts as follows: $\Sigma_h = E(1/N^h_{it})\Sigma^j_h$.

[28] In principle analysts could observe different pieces of soft information, Z^j_{it}. This would not affect our analysis, as $var(z^h_{it})$ would have the interpretation as the informational advantage of the average signal.

Using eq. (18), the decomposition of the difference in mean squared error between the machine and the analyst given in eq. (6) becomes:

$$MSE^a_{t+h} - MSE^m_{t+h} = \Delta_h + \Sigma_h - \Theta_h + \underbrace{(1-\alpha_h)^2\Theta_h}_{\text{effect of bias on private information}} . \tag{19}$$

The final term in this decomposition illustrates bias in processing soft information improves the forecasting performance of the machine relative to the analyst. If this bias is sufficiently large (i.e. $(1-\alpha)^2 > 1$), then the benefits for forecasting performance of the analyst's extra information is swamped by the bias in her processing of it.

D.2 Effect of Bias on Z_{it} on Parameter Estimates

Recall the following definitions of F^*_t and EPS^*_{it+h} from Proposition 2:

$$F^*_t EPS_{it+h} \equiv F_t EPS_{it+h} - E\left(F_t EPS_{it+h}|X_{it}\right) = z^h_{it} + \eta^h_{it}$$
$$EPS^*_{it+h} \equiv EPS_{it+h} - E\left(EPS_{it+h}|X_{it}\right) = z^h_{it} + \varepsilon_{t+h}$$

The following proposition implicitly characterizes the four parameters using the three moments proposed in Proposition 2, which simplify to the explicit solutions for Θ_h, Δ_h and Σ_h when $\alpha_h = 1$:

Proposition D2. *The analyst bias, information advantage, and noise satisfy the following moment conditions:*

$$E\left[\left(E(F_t EPS_{it+h}|X_{it}) - E(EPS_{it+h}|X_{it})\right)^2\right] = \Delta_h$$
$$cov(F^*_t EPS_{it+h}, EPS^*_{it+h}) = \alpha_h\Theta_h$$
$$var(F^*_t EPS_{it+h}) = \alpha^2_h\Theta_h + \Sigma_h$$

From Proposition D2, its clear that if we want to identify α_h in addition to our other three parameters, we need to impose additional identifying restrictions because we have four parameters with only three moments. In the main text of this paper we impose the identifying assumption that $\alpha_h = 1$, but Proposition D2 shows how the failure of this assumption would affect our parameter estimates.

The first equation of Proposition D2 shows Δ_h is unaffected, which is intuitive because it depends only on x_{it} (which is orthogonal to Z_{it}). The second equation shows that the effect on Θ_h depends on whether the analyst underreacts ($\alpha_h < 1$) or overreacts ($\alpha_h > 1$) to private information. If the analyst underreacts, the estimated information advantage will be understated, as the analyst's information is essentially "hidden". If the analyst overreacts, the estimated information advantage will be overstated, as the analyst induces excessive correlation between $F^*_t EPS_{it+h}$ and EPS^*_{it+h} that is costly in terms of MSE (eq. (19)). Finally, the effect on Σ_h also depends on the sign of the bias on Z_{it}. Since expectation noise is essentially the residual variance in $F^*_t EPS_{it+h}$ after accounting for Θ_h, overreaction (underreaction) upward (downward) biases estimates of it.

D.3 Moment Selection and Identification

Proposition D2 shows Δ_h is unaffected by relaxing the identifying assumption that $\alpha_h = 1$, so we can estimate it exactly as described in Section 3.3. The essence of our identification strategy for the remaining three parameters is to generate an estimate of Σ_h using a moment that is linearly independent from those provided in Proposition D2, and then use the latter moments to estimate Θ_h and α_h. The following proposition expresses these three parameters in terms of the three moments we use:

Proposition D3. *Assume N_{it}^h is independent of z_{it}^h and $E(N_{it}^h)$ is finite. Then the noise, the bias on Z_{it}, and the information advantage are given (recursively) by the following three equations:*

$$\Sigma_h = -cov\left(\left(F_t^* EPS_{it+h}\right)^2, N_{it}^h\right) * C$$
$$\alpha_h = \frac{var\left(F_t^* EPS_{it+h}\right) - \Sigma_h}{cov\left(F_t^* EPS_{it+h}, EPS_{it+h}^*\right)}$$
$$\Theta_h = \frac{cov\left(F_t^* EPS_{it+h}, EPS_{it+h}^*\right)}{\alpha_h},$$

where $C > 0$ is a constant equal to $\frac{E\left(\frac{1}{N_{it}^h}\right)}{\left[E\left(N_{it}^h\right)E\left(\frac{1}{N_{it}^h}\right)-1\right]}$. Additionally, the bias on public information is given by

$$\Delta_h = E\left[\left(E(F_t EPS_{it+h}|X_{it}) - E(EPS_{it+h}|X_{it})\right)^2\right].$$

The final two equations of Proposition D3 follow directly from Proposition D2. The new aspect of this proposition is the identification of Σ_h using the covariance between the square of $F_t^* EPS_{it+h}$ and the number of analysts, N_{it}^h (modulo the constant C). The intuition for this covariance identifying expectation noise is that as the number of *individual* analysts gets large, the sum of their individual noise terms will tend to zero because they are mutually uncorrelated.[29] This subsequently lowers the noise in the consensus forecast, which depends on the average noise across all analysts, reducing $(F_t^* EPS_{it+h})^2$.

D.4 Estimation

Our estimation of the four quantities begins by performing the first three steps outlined in Section 3.4. We then replace the four step with the procedure described in Proposition D3. The results are presented in Table A9.

[29] This is indeed true in the data. If we estimate a conditional version of our model across subsamples of firm-years with a different number of analysts, we indeed find Σ_h is decreasing in N_{it}^h.

Table A9. Estimated Model Parameters: Extended Model

This table presents our results from estimating the parameters in our extended model that allows for bias in processing soft information, using the estimation procedure described above. Θ_h represents the analyst information advantage; Δ_h represents the analyst bias; Σ_h is the analyst noise. Each row corresponds to the use of a different ML estimator to form our statistical forecasts. Panel A focuses on one-year ahead forecasts, while Panels B and C produce the same analysis for two-year and three-year ahead forecasts, respectively. Parameter estimates are normalized by the average squared earnings-per-share divided by price, calculated across the entire sample. Standard errors are calculated using a clustered bootstrap at the firm level with 1,000 iterations. For each estimation, the firm-year level variables in Proposition 2 are trimmed at 5 times the interquartile range, after they are scaled by price. The final samples in Panels A, B, and C contain and 41,208, 35,458, and 11,533 firm-year observations, respectively.

Panel A: One-Year Horizon Forecasts

Estimator	Θ	s.e.	Δ	s.e.	α	s.e.	$(1-\alpha)^2\Theta$	s.e.	Σ	s.e.
Lasso	5.13%	(0.22%)	2.12%	(0.02%)	1.09	(0.02)	0.04%	(0.02%)	0.97%	(0.13%)
Post-Lasso	5.63%	(0.24%)	3.31%	(0.03%)	1.11	(0.02)	0.07%	(0.03%)	1.21%	(0.15%)
Iterative Lasso	5.9%	(0.25%)	1.56%	(0.01%)	1.06	(0.02)	0.02%	(0.04%)	1.13%	(0.14%)
Ridge	6.33%	(0.28%)	3.27%	(0.03%)	1.11	(0.03)	0.07%	(0.03%)	1.41%	(0.18%)
Elastic Net	5.21%	(0.21%)	2.1%	(0.02%)	1.08	(0.02)	0.03%	(0.02%)	0.97%	(0.14%)
Random Forest	3.75%	(0.15%)	1.99%	(0.02%)	1.15	(0.03)	0.08%	(0.01%)	0.53%	(0.11%)
Gradient-Boosted Trees	4.15%	(0.2%)	2.36%	(0.02%)	1.14	(0.03)	0.08%	(0.01%)	0.79%	(0.13%)
Mean	5.16%		2.39%		1.10		0.06%		1.0%	

Panel B: Two-Year Horizon Forecasts

Estimator	Θ	s.e.	Δ	s.e.	α	s.e.	$(1-\alpha)^2\Theta$	s.e.	Σ	s.e.
Lasso	2.81%	(0.25%)	10.62%	(0.08%)	1.64	(0.05)	1.14%	(0.08%)	1.92%	(0.17%)
Post-Lasso	2.89%	(0.27%)	13.53%	(0.13%)	1.74	(0.06)	1.6%	(0.12%)	2.26%	(0.22%)
Iterative Lasso	3.77%	(0.31%)	8.73%	(0.07%)	1.56	(0.05)	1.17%	(0.09%)	2.44%	(0.24%)
Ridge	3.27%	(0.28%)	12.3%	(0.11%)	1.69	(0.05)	1.54%	(0.12%)	2.52%	(0.23%)
Elastic Net	2.85%	(0.27%)	10.65%	(0.08%)	1.62	(0.06)	1.1%	(0.09%)	1.94%	(0.18%)
Random Forest	1.56%	(0.15%)	11.66%	(0.08%)	1.99	(0.08)	1.54%	(0.14%)	1.32%	(0.18%)
Gradient-Boosted Trees	2.05%	(0.18%)	12.0%	(0.09%)	1.78	(0.06)	1.24%	(0.11%)	1.4%	(0.16%)
Mean	2.74%		11.36%		1.72		1.33%		1.97%	

Panel C: Three-Year Horizon Forecasts

Estimator	Θ	s.e.	Δ	s.e.	α	s.e.	$(1-\alpha)^2\Theta$	s.e.	Σ	s.e.
Lasso	2.37%	(0.3%)	11.46%	(0.15%)	2.33	(0.17)	4.17%	(0.61%)	1.1%	(0.62%)
Post-Lasso	2.85%	(0.41%)	16.97%	(0.22%)	2.22	(0.15)	4.23%	(0.61%)	1.86%	(0.71%)
Iterative Lasso	4.31%	(0.44%)	8.4%	(0.1%)	1.90	(0.14)	3.46%	(0.73%)	1.48%	(0.84%)
Ridge	2.21%	(0.35%)	13.52%	(0.15%)	2.49	(0.19)	4.92%	(0.65%)	2.0%	(0.67%)
Elastic Net	2.53%	(0.31%)	11.43%	(0.15%)	2.25	(0.16)	3.94%	(0.58%)	0.99%	(0.63%)
Random Forest	1.66%	(0.21%)	11.54%	(0.15%)	2.49	(0.17)	3.71%	(0.47%)	0.9%	(0.54%)
Gradient-Boosted Trees	1.44%	(0.25%)	14.82%	(0.23%)	2.78	(0.2)	4.56%	(0.48%)	1.26%	(0.48%)
Mean	2.48%		12.59%		2.35		4.14%		1.37%	

Panel A of Table A9 shows that we find limited evidence of bias in processing soft information at the one-year horizon. This means that bias in processing soft information constitutes a very small profit loss in forecasting relative to the gain of exploiting the analyst's information advantage. We find larger bias in processing soft information at the two-year and three-year horizons. Our findings that $\alpha_h \leq 1$ in Panel A and $\alpha_h > 1$ in Panels B and C are consistent with the literature on expectations formation, which generally finds short-run underreaction (e.g. Bouchaud et al. 2019) and long-run overreaction (e.g. Giglio and Kelly 2018; Bordalo, Gennaioli, Ma, and Shleifer 2018). Moreover, the slope of the upward term structure in $(1-\alpha_h)^2\Theta_h$ suggests the mechanism that generates an upward slope in Δ_h could also be at play on private information. As expected from Proposition D2, we find a slightly lower estimated information advantage, given some of the information is lost due to bias in processing it. The estimated bias on public information is unchanged by definition.

Finally, Panels B and C show that we find less noise, but the term structure still appears somewhat upward sloping. Although this is somewhat at odds with the result in Table 2, we note that the source of variation to identify the noise is very different here – its coming essentially from a cross-sectional regression of $F_t^* EPS_{it+h}$ onto N_{it}^h. We are less comfortable using this moment to identify Σ_h, given there is plenty of unmodeled firm heterogeneity that would pollute this moment and break the assumption that $N_{it}^h \perp Z_{it}$, which is required for Proposition D2 to hold. Thus, we conclude that the upward sloping term structure of noise we document in the main text is likely still upward sloping once bias on private information is allowed, but the exact level of this slope is unclear.

Appendix E. Afrouzi et al. (2021) Model

In this section, we apply the model proposed by Afrouzi et al. (2021) to our setting. We first briefly describe it, referring the reader to Afrouzi et al. (2021) Section 5 for additional details. We then show it qualitatively delivers an upward sloping term structure of bias and noise, but fails quantitatively in our setting for an intuitive reason.

E.1 Model Description

Consider the same setup as in Section 4.1, including Assumption 1. In addition, assume further that $u_{it} \sim \mathcal{N}\left(0, \sigma_u^2\right)$. At time t, the analyst needs to form forecasts of EPS_{it+h} for $h \geq 1$. The analyst costlessly observes x_{it-1}, z_{it}^h, and ρ. Importantly, the analyst does not know μ, but has the prior $\mu | x_{it-1} \sim \mathcal{N}(x_{it-1}, \underline{\tau})$. The analyst also knows the DGP for EPS_{it}. Because x_{it}^h and z_{it}^h are orthogonal, we assume the analyst forms forecasts of EPS_{it+h} by first forming optimal forecasts of x_{it+h-1} and z_{it}^h individually, then adding them together. Denote her forecasts by F_t. Since $F_t z_{it}^h = z_{it}^h$ by assumption, we focus on characterizing $F_t x_{it+h}$.

To form $F_t x_{it+h}$, the analyst has access to two sources of information. First, the agent costlessly observes x_{it-1}, which is "on top of his mind". Secondly, the agent can retrieve additional information from past data to learn about μ, but doing so is costly. The chooses the level of information acquisition in order to minimize the expected squared error of his forecasts, subject to the cost of information retrieval. Formally, the analyst solves the following problem:

$$\min_{S_{it}} E\left[\min_{F_t x_{it+h}} E\left[\left(F_t x_{it+h} - x_{it+h}\right)^2 | S_{it}\right] + C(S_{it})\right], \tag{20}$$
$$s.t. \quad \{x_{it-1}\} \in S_{it} \subset \mathcal{S}_{it}, \quad \mathcal{S}_{it} = \{s : s \perp \mu | x_{it-1}, x_{it-2}, ...\}.$$

The solution to the inner maximization problem problem is $F_t x_{it+h} = E(x_{it+h} | S_{it})$. As shown in Afrouzi et al. (2021), the assumption that u_{it} is normally distributed implies the outer maximization problem can be reduced to choosing a belief prediction about μ, denoted τ. Denote the solution to this problem as $\tau^*(h)$.

To further characterize the solution to this problem, Afrouzi et al. (2021) assume the cost of information retrieval takes the following form:

$$C_t\left(S_t\right) \equiv \omega \frac{\exp\left(2\ln(2) \cdot \gamma \cdot \mathbb{I}\left(S_t, \mu \mid x_{it-1}\right)\right) - 1}{\gamma}.$$

In this expression, $\mathbb{I}(\cdot, \cdot)$ denotes Shannon mutual information, $\omega \geq 0$ governs the overall cost of retrieval, and $\gamma \geq 0$ measures convexity of the cost function in mutual information. Given this cost function, Afrouzi

et al. (2021) show the choice of belief precision, τ, that solves (20) is:

$$\tau^*(h) = \underline{\tau} \max\left\{1, \left(\frac{\left(1-\rho^{h+1}\right)^2}{\omega \underline{\tau}}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{21}$$

This choice of $\tau = \tau^*(h)$ implies the analyst's forecast is the following:

$$F_t x_{it+h} = (1-\rho^{h+1})\left(1 - \frac{\underline{\tau}}{\tau^*(h)}\right)\mu + \left(\rho^{h+1} + (1-\rho^{h+1})\frac{\underline{\tau}}{\tau^*(h)}\right)x_{it-1} + \eta_{it}^h,$$
$$\eta_{it}^h \sim \mathcal{N}\left(0, (1-\rho^{h+1})^2 \frac{1}{\tau^*(h)}\left(1 - \frac{\underline{\tau}}{\tau^*(h)}\right)\right).$$

Note that $F_t x_{it+h}$ is a random variable because of the expectation noise, η_{it}^h.

E.2 Term Structure of Bias and Noise

Now that we have characterized the analyst's forecasts, we can determine the term structure of bias and noise in this model.

Proposition E4. *In the Afrouzi et al. (2021) model, the term structure of bias and noise are given by:*

$$\Delta_h = \left[(1-\rho^h)\frac{\underline{\tau}}{\tau^*(h-1)}\right]^2 \frac{\sigma_u^2}{1-\rho^2}, \qquad \Sigma_h = \left(1-\rho^h\right)^2 \frac{1}{\tau^*(h-1)}\left(1 - \frac{\underline{\tau}}{\tau^*(h-1)}\right).$$

Thus, the Afrouzi et al. (2021) model produces an upward sloping term structure of bias and noise.

Proposition E4 shows noise in increasing with the horizon because of $\tau^*(h-1)$ increasing with the horizon (see eq. (21)). The intuition here is that at longer horizons, it is more useful to know the long-run mean, so the analyst engages in more information retrieval. This greater retrieval provides the analyst with more noisy information, creating large noise in her forecasts.

Additionally, the formula for Δ_h in Proposition E4 shows bias is increasing with the horizon, since $\tau^*(h-1)$ is increasing in h at a slower rate than $1-\rho^h$. This upward term structure of bias results from a combination of two effects. On the one hand, the agent engages in more retrieval, moving her forecast closer to the conditional expectation. On the other hand, the analyst still overweights x_{it-1} in her estimate of the long-run mean, and this overweighting is magnified at longer horizons because the analyst's forecast moves closer to her subjective expectation of the long-run mean. This second effect dominates, generating an upward sloping term structure of bias.

In sum, the Afrouzi et al. (2021) model generates an upward sloping term structure of bias and noise. We now discuss how this model cannot *quantitatively* fit the data.

E.3 Quantitative Model Fit

The Afrouzi et al. (2021) model has 5 parameters: ρ, σ_u, ω, γ, and $\underline{\tau}$. The key endogenous parameter is $\tau^*(h)$. Rewriting the equation for Δ_h in Proposition E4, we obtain an expression for $\tau^*(h)$ as a function of the bias at horizon h:

$$\frac{\tau^*(h-1)}{\underline{\tau}} = \frac{\left(1-\rho^h\right)\sigma_u}{\sqrt{\Delta_h\left(1-\rho^2\right)}}. \tag{22}$$

Equation (22) is useful because it express $\frac{\tau^*(h-1)}{\underline{\tau}}$, which is the crucial to this model, as a function of data moments we have already estimated. Using $\rho = 0.928$ from Table A8, $\{\Delta_1, \Delta_2, \Delta_3\} = \left\{4.3\times10^{-5}, 2.5\times10^{-4}, 5.8\times10^{-4}\right\}$ from Figure 1[30], and $\sigma_u = 0.01$ from Table A8[31], we obtain

$$\frac{\left(1-\rho^1\right)\sigma_u}{\sqrt{\Delta_1\left(1-\rho^2\right)}} \approx 0.27, \quad \frac{\left(1-\rho^2\right)\sigma_u}{\sqrt{\Delta_2\left(1-\rho^2\right)}} \approx 0.21, \quad \frac{\left(1-\rho^3\right)\sigma_u}{\sqrt{\Delta_3\left(1-\rho^2\right)}} \approx 0.20.$$

By eq. (22), we obtain

$$\frac{\tau^*(1-1)}{\underline{\tau}} \approx 0.27, \quad \frac{\tau^*(2-1)}{\underline{\tau}} \approx 0.21, \frac{\tau^*(3-1)}{\underline{\tau}} \approx 0.20.$$

However, this contradicts eq. (21), which shows $\forall h$, $\frac{\tau^*(h)}{\underline{\tau}} \geq 1$. Thus, these numerical results illustrate that the Afrouzi et al. (2021) model cannot match our data.

The intuition of the quantitative failure of the Afrouzi et al. (2021) model to match our data is the following. From Proposition E4, the bias at horizon h is increasing in σ_u: as $\sigma_u \to 0$, there will be no bias because x_{it} will be a deterministic process. In the data, we in fact find σ_u is quite low, given the R^2 in column (1) of Table A8 is around 93%. However, despite a low σ_u, we still find substantial forecasting bias – the bias is so high that this model would require the analyst to *forget* information to match this level of bias. The heart of this problem is that the model gives the analyst access to x_{it-1}, which is really close to x_{it+h} because σ_u is low. In other words, the fact that the analyst gets to see the machine forecast from the last period, $F^m_{t-1}EPS_{it} = x_{it-1}$, gives her too much knowledge to be as biased (and noisy) as we find in the data.

[30] These are the values Δ_h in Figure 1 before we use the normalization by average squared earnings-per-share over price for result presentation.

[31] This isn't shown exactly in Table A8, but it's simply the standard error of the regression residual in column (1).

Appendix F. Additional Derivations and Proofs

Proposition 1. Applying the conditional mean independence restrictions from Section 3.1 and the fact that Z_{it} and η_{it} have zero unconditional means we obtain:

$$\begin{aligned}
MSE^a_{t+h} &\equiv E\left[\left(EPS_{it+h} - F_t EPS_{it+h}\right)^2\right] \\
&= E\left[\left(x^h_{it} - g_h(X_{it} + z^h_{it} - z^h_{it} + \varepsilon_{it+h} - \eta_{it+h}\right)^2\right] \\
&= \Delta_h + var\left(\varepsilon_{it+h}\right) + \Sigma_h \\
MSE^m_{t+h} &\equiv E\left[\left(EPS_{it+h} - E\left(EPS_{it+h} \mid X_{it}\right)\right)^2\right) \\
&= E\left[\left(x^h_{it} - x^h_{it} + z^h_{it} + \varepsilon_{it+h}\right)^2\right] \\
&= \Theta_h + var\left(\varepsilon_{it+h}\right).
\end{aligned}$$

Subtracting these two gives the desired result. □

Proposition 2. The expression for Δ_h follows directly. The expressions for Θ_h and Σ_h follow from the fact that

$$var\left(F^*_t EPS_{it+h}\right) = \Theta_h + \Sigma_h.$$

□

Proposition 3. First note that combining Assumption 1 with the law of iterated expectations implies

$$\begin{aligned}
E\left(EPS_{it+h} \mid X_{it}\right) \equiv x^h_{it} &= \left(1 - \rho^{h-1}\right)\mu + \rho^{h-1} x_{it}, \\
&\equiv \left(1 - \rho^{h-1}\right)\mu + \rho^{h-1} E\left(EPS_{it+1} \mid X_{it}\right).
\end{aligned}$$

Therefore, at horizon h, the bias is

$$\begin{aligned}
\Delta_h &= E\left[\left(E(EPS_{t+h}|X_{it}) - E(F_t EPS_{t+h}|X_{it})\right)^2\right], \\
&= E\left[\left(\rho^{h-1} x_{it} - \rho^{h-1} E(F_t x_{it}|X_{it})\right)^2\right] = \rho^{2(h-1)} E\left[\left(x_{it} - E(F_t x_{it}|X_{it})\right)^2\right] = \rho^{2(h-1)}\Delta^1.
\end{aligned}$$

The noise is

$$\begin{aligned}
\Sigma_h &= var(\eta_{t,h}) = var\left(F_t EPS_{t+h} - E\left(F_t EPS_{t+h}|X_{it}, Z_{it}\right)\right), \\
&= var\left(F_t x_{t+h} - E(F_t x_{t+h}|X_{it}, Z_{it})\right) = \rho^{2(h-1)} var(\eta_{t,1}) = \rho^{2(h-1)}\Sigma^1_\eta.
\end{aligned}$$

The result follows because $\rho < 1$ by assumption. □

Proposition B1. By definition, $cov(F^*_t EPS_{it+1}, EPS^*_{it+1}|\mathcal{F}) = var(Z_{it}|\mathcal{F}) + cov(Z_{it}, \varepsilon_{it+1}|\mathcal{F}) + cov(Z_{it}, \eta_{it}|\mathcal{F}) + cov(\eta_{it}, \varepsilon_{it+1}|\mathcal{F}) = var(Z_{it}|\mathcal{F})$, where the final equality follows from Assumption B1. Thus,

$$\Theta(\mathcal{F}) = cov(F^*_t EPS_{it+1}, EPS^*_{it+1}|\mathcal{F}) + E(Z_{it}|\mathcal{F})^2 = cov(F^*_t EPS_{it+1}, EPS^*_{it+1}|\mathcal{F}) + E(F^*_t EPS_{it+1}|\mathcal{F})^2,$$

which delivers the first equation in Proposition B1. By definition, $var(F_t^* EPS_{it+1}|\mathcal{F}) = var(Z_{it}|\mathcal{F}) + \Sigma(\mathcal{F})$. Using the first sequence of equalities above,

$$\Sigma(\mathcal{F}) = var(F_t^* EPS_{it+1}|\mathcal{F}) - cov(F_t^* EPS_{it+1}, EPS_{it+1}^*|\mathcal{F}).$$

The expression for $\Delta(\mathcal{F})$ follows directly. □

Proposition D2. Given that α_h is a constant, the expressions follow directly as in the derivation of Proposition 2. □

Proposition D3. The expression for Δ_h follows directly, since Δ_h is unaffected by $\alpha_h \neq 1$. To obtain the remaining expression, we begin by noting two facts. First,

$$E\left(\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right) = E(N_{it}^h)E\left(\eta_{it}^{hj}\right) = 0.$$

The first equality in this expression follows from Wald's identity, which applies given the assumptions made in Proposition D3. Secondly,

$$var\left(\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right) = E(N_{it}^h)var(\eta_{it}^{hj}) + E\left(\eta_{it}^{hj}\right)^2 var(N_{it}^h),$$

by a similar application of the law of iterated expectations and Wald's identity. We will use these two facts throughout the derivations below.

Now, note that

$$var\left(F_t^* EPS_{it+h}\right) = \alpha_h^2 \Theta_h + var\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right) + \alpha_h cov\left(z_{it}^h, \frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right). \tag{23}$$

We now simplify each of the latter two elements of (23). First, using the facts above, we obtain:

$$\begin{aligned}
var\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right) &= E\left[\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right)^2\right] - E\left[\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right]^2 \\
&= E\left[E\left[\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj}\right)^2 \mid N_{it}^h = N\right]\right] - E\left[E\left[\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h} \eta_{it}^{hj} \mid N_{it}^h = N\right]\right] \\
&= E\left[E\left[\left(\frac{1}{N}\sum_{j=1}^{N} \eta_{it}^{hj}\right)^2\right] \mid N_{it}^h = N\right] - 0 \\
&= E\left[var\left(\frac{1}{N}\sum_{j=1}^{N} \eta_{it}^{hj}\right) \mid N_{it}^h = N\right] = E\left(\frac{1}{N_{it}^h}\Sigma_h\right)
\end{aligned}$$

Now consider the final term in (23). By the assumption of independence between N_{it}^h and z_{it}^h in Proposi-

tion D3 (the restrictive assumption required for this identification strategy), this term is zero. Thus, we've shown

$$var\left(F_t^* EPS_{it+h}\right) = \alpha_h^2 \Theta_h + E\left(\frac{1}{N_{it}^h}\Sigma_h\right). \tag{24}$$

Next, consider an additional moment, $cov\left(F_t^* EPS_{it+h}^2, N_{it}\right)$. By direct calculation, we obtain

$$cov\left(F_t^* EPS_{it+h}^2, N_{it}\right) = \alpha_h^2 cov\left[\left(z_{it}^h\right)^2, N_{it}^h\right] + cov\left[\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h}\eta_{it}^{hj}\right)^2, N_{it}^h\right] + 2\alpha_h cov\left(\frac{z_{it}^h}{N_{it}^h}\sum_{j=1}^{N_{it}^h}\eta_{it}^{hj}\right). \tag{25}$$

The first term in this (25) is again zero by our independence assumption. The final term is also zero by an application of Wald's identity with the independence assumption. The second term can be simplified using the two facts above:

$$\begin{aligned} cov\left[\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h}\eta_{it}^{hj}\right)^2, N_{it}^h\right] &= E\left[\frac{1}{N_{it}^h}\left(\sum_{j=1}^{N_{it}^h}\eta_{it}^{hj}\right)^2\right] - E(N_{it}^h)E\left[\left(\frac{1}{N_{it}^h}\sum_{j=1}^{N_{it}^h}\eta_{it}^{hj}\right)^2\right] \\ &= E\left[E\left[\frac{1}{N}\left(\sum_{j=1}^{N}\eta_{it}^{hj}\right)^2 \mid N_{it}^h = N\right]\right] - E(N_{it}^h)E\left(\frac{1}{N_{it}^h}\right)\Sigma_h \\ &= E\left[\frac{1}{N}N\Sigma_h\right] - E(N_{it}^h)E\left(\frac{1}{N_{it}^h}\right)\Sigma_h. \end{aligned}$$

Thus, we can re-write (25) as:

$$cov\left(F_t^* EPS_{it+h}^2, N_{it}\right) = \Sigma_h\left[1 - E\left(\frac{1}{N_{it}^h}\right)E(N_{it}^h)\right],$$

which simplifies to give the first expression in Proposition D3. Note that by Jensen's inequality, the expression shows Σ_h will be positive (as expected).

The remaining expressions in Proposition D3 follow from combining (24) with the results from Proposition D2. □

Proposition E4. The formula for noise follows directly from the expression for $\text{var}\left(\eta_{it}^h\right)$. Bias is given by:

$$\begin{aligned} \Delta_h &\equiv E\left[\left(E\left(EPS_{it+h} \mid \mu, \rho, x_{it-1}\right) - E\left(F_t EPS_{it+h} \mid \mu, \rho, x_{it-1}\right)\right)^2\right], \\ &= E\left[\left(E\left(x_{it+h-1} \mid \mu, \rho, x_{it-1}\right) - E\left(F_t x_{it+h-1} \mid \mu, \rho, x_{it-1}\right)\right)^2\right], \\ &= \left[(1-\rho^h)\frac{\tau}{\tau^*(h)}\right]^2 E\left[(\mu - x_{it-1})^2\right], \\ &= \left[(1-\rho^h)\frac{\tau}{\tau^*(h)}\right]^2 \frac{\sigma_u^2}{1-\rho^2}. \end{aligned}$$

□

Appendix G. Robustness to Choice of Features in Machine Learning Estimation

In this section, we show our conclusions from the relative accuracy of our three forecasts (the analyst, machine, and analyst + machine) based on Table 1 are robust to using an alternative source of variables in x_{it}. Contemporaneous work by van Binsbergen et al. (2020), Cao and You (2020), and Hansen and Thimsen (2020) also form forecasts of earnings-per-share using machine learning, but they use a different set of variables from those used in Table A1. Although one might be concerned that are results are sensitive to our choice of variables, in this appendix we provide evidence that this is likely not the case.

Before discussing an alternative set of conditioning variables, we would like to be clear on how we chose the variables in Table A1. Unlike van Binsbergen et al. (2020), Cao and You (2020), and Hansen and Thimsen (2020), we use raw Compustat data rather than ratios that have been identified by prior literature as good predictors of earnings. We chose to do this in order to reduce any look-ahead bias that this would introduce into our forecasts: one could have not built a machine learning forecast in real-time using these ratios, given these ratios we're not identified as good predictors until later in the sample. By using variables directly from Compustat (and their interactions in our penalized linear estimators), we ensure the accuracy of our machine learning forecasts are comparable to what could have been achieved in real-time. In principle, we would have liked to use all variables in Compustat, but this is not feasible since many are frequently missing. The variables in Table A1 represent the largest set of variables that had a non-trivial amount of non-missing observations.

The results in Table A10 show we find almost identical forecasting accuracy to that in Table 1 Panel A. This suggests our conclusions are sensitive to the particular transformations of the features we used in the main text.

Table A10. Forecast Mean Squared Errors: Robustness Using Ratios

This table contains the mean squared error of various forecasts over our entire sample, normalized by the mean squared EPS. Machine and machine + analyst forecasts are formed using the procedure described in Section 1 and Appendix A, using the ratios described in Appendix G instead of the variables in Table A1.The mean squared errors displayed in the table are calculated by first calculating the mean squared error within each year and normalizing by the mean squared realized earnings-per-share in that year, which represents the percentage utility loss relative to having perfect foresight in the interpretive model presented in each year (see Section 2 for model). We then take a weighted average across all years with weights proportional to the number of observations in each year (results are insensitive to this weighting) to arrive at the numbers presented in this table, which represent the the average losses to the investor considered in Section 2. This table presents the results for forecasts at the one-year horizon, as described in Section 1.

	Analyst	Machine	Machine + Analyst
Lasso	-	15.34%	9.56%
Random Forest	-	15.52%	10.05%
Gradient-Boosted Trees	-	15.66%	10.01%