

The Usefulness of Machine Learning in Predicting the Sign of Earnings Changes: Evidence from Japanese Listed Firms

Minjun Jeong

Table of Contents

1. Introduction
2. Preliminaries
3. Research Design
4. Sample and Data
5. Results
6. Conclusion

Abstract

This study examines the usefulness of machine learning for predicting the sign of earnings changes for Japanese listed firms. The results show that machine-learning predictions achieve overall accuracy comparable to management earnings forecasts, while high-confidence predictions are more accurate. Hedge portfolios constructed from the machine-learning predictions generate statistically significant abnormal returns. Variable-importance analyses indicate that management forecast information is the primary driver of the predictions, while financial-statement information provides complementary predictive content. The article comprises the main text and an online appendix; footnotes in the main text identify material discussed in greater detail in the appendix.

[DOI] <https://doi.org/10.69388/saj.63.09.89>



Minjun Jeong

Minjun Jeong is a doctoral student at the Graduate School of Business, Osaka Metropolitan University. Jeong completed the master's program at the Graduate School of Business, Osaka City University (now Osaka Metropolitan University), in 2023.

1. Introduction

The study by Ou and Penman [1989] is a pioneering application of machine learning to the prediction of future earnings-change signs. It shows that such predictions can be used to earn abnormal returns and has had a major influence on subsequent earnings-prediction research. Conventional prediction methods, however, have inherent limitations. Because of their restrictive functional forms, for example, time-series models and cross-sectional regressions may not fully capture the nonlinear and interdependent relationships among numerous accounting variables. Although analyst forecasts are generally accurate, concerns have been raised about herding and the concentration of analyst coverage among large firms (Nara and Noma [2024]). Management forecasts are an important source of forward-looking information in Japan, but their accuracy and reliability have also been questioned (Ota [2006]).

To address these limitations, researchers have increasingly applied machine learning in accounting research. Machine-learning methods can capture nonlinearities, high-dimensional relationships, and complex interactions among variables, thereby relaxing the linearity assumptions and restrictive variable-selection requirements of conventional regression models. Studies by Chen et al. [2022] and Hunt et al. [2022] report improvements in earnings prediction from machine learning, whereas evidence for Japanese firms remains limited. Ishibashi et al. [2016] examine data-mining-based variable-selection methods for constructing earnings-prediction models from financial statements but do not compare their models with management forecasts. Yakabi et al. [2024] use machine learning to predict the sign of earnings changes and assess the economic value of the predictions; they report greater predictive accuracy than conventional logistic regression models but do not directly compare their predictions with management forecasts.

This study uses machine learning to predict the sign of earnings changes for Japanese listed firms and empirically evaluates the usefulness of those predictions. Specifically, it compares the accuracy of the machine-learning predictions with that of management forecasts and assesses their economic usefulness using portfolios constructed from the predicted signs. Because machine learning is often criticized as a “black box,” the study also analyzes variable importance to improve the transparency and interpretability of the prediction process.

The remainder of the paper is organized as follows. Section 2 describes the machine-learning model, and section 3 presents the research design. Section 4 describes the sample and data, section 5 reports the results, and section 6 concludes and outlines directions for future research.

2. Preliminaries

Machine learning encompasses methods that learn patterns from data and use those patterns to generate predictions for new observations. During training, a model estimates parameters or decision rules from the data by optimizing a loss function or related criterion.

Machine-learning methods are commonly classified as supervised or unsupervised. This study uses supervised learning in which a model learns the relationship between input variables and a labeled output. For the binary classification task of predicting the sign of earnings changes from financial information, decision tree-based methods offer a useful balance between interpretability and predictive accuracy. The study therefore employs a random forest, which is an ensemble method based on decision trees.

A decision tree generates predictions by recursively partitioning the data according to a sequence of split rules. Because its structure is readily interpretable, the method has been applied in a wide range of fields. A single decision tree, however, is prone to overfitting: it may fit the training data well but generalize poorly to new data. Breiman [2001] addressed this problem by developing the random forest (RF), which combines multiple decision trees.

RF builds on bootstrap aggregating (bagging) in which multiple bootstrap samples are drawn from the original data and a separate decision tree is trained on each sample. At

each split in each tree, the algorithm considers only a randomly selected subset of the input variables. Randomizing both the observations and the candidate variables increases diversity among the trees and reduces the variance of the ensemble. Because the trees are trained on different samples and variable subsets, errors by individual trees can be offset by the remaining trees. The final classification is based on a majority vote or on averaged class probabilities across the trees.

3. Research Design

This section describes the machine-learning prediction procedure, including the definition of the output variable, construction of the input variables, model training and evaluation, and the methods used to interpret the model.

3.1 Output and Input Variables¹

The output variable is the sign of the next-period change in earnings (Pr). It is coded as one when next-period net income minus current-period net income is positive and as zero when the difference is zero or negative. This coding follows Ou and Penman [1989], although the present study does not apply their adjustment for mean reversion toward historical average earnings (drift).²

The model uses 120 input variables, most of which are quantitative measures derived from financial statements. The variables fall into three broad groups. First, the financial-statement group comprises 16 balance-sheet items, 15 income-statement items, and 3 cash-flow-statement items. For each item, the current-period value, one-period lag, and first difference are included, yielding 102 variables. Second, the management forecast group is based on the forecast of next-period net income disclosed in the Summary of Financial Results (*Kessan Tanshin*). Four measures are constructed: the management earnings forecast (MF), realized forecast bias (MFB), realized absolute forecast error

¹Following machine-learning terminology, this study refers to the variable being predicted as the output variable and to the variables used for prediction as input variables. In econometrics and accounting research, these are generally called the dependent variable and independent variables, respectively. In conventional regression terminology, the input variables also correspond to explanatory variables; in machine learning, they are also called features.

²There are three reasons for omitting this adjustment. First, Hunt et al. [2022] report that estimating drift requires several years of historical data and reduces the sample size while yielding almost no improvement in predictive accuracy. Second, Belesis et al. [2023] find that adding a mean-reversion measure improves the F1-score by no more than approximately one percentage point, suggesting that the mean reversion in earnings changes does not materially improve predictive power. Third, Chen et al. [2022] show that their results are robust to including or excluding the drift adjustment.

(MFAE), and ex ante forecast bias (MFS). The current-period value, one-period lag, and first difference of each measure yield 12 variables. Third, the price-based group comprises six variables: the earnings-to-price (EP) and book-to-price (BP) ratios, their one-period lags, and their first differences.³

3.2 Training the Machine Learning Model

Model training proceeds in three stages.

3.2.1 Training and Validation with Time-Series Data

Given the time-series structure of the financial data, this study uses time-series cross-validation. The model is trained on historical data and validated on the immediately following period, thereby preserving the temporal ordering of the observations. A fixed training window and an immediately subsequent fixed validation window are rolled forward to create multiple folds, and model performance is evaluated in each fold.

The procedure uses 12 folds with the training and validation windows advanced one month at a time. For example, when March 2017 is the test month, the first fold uses the 36 months from April 2012 through March 2015 for training and the 12 months from April 2015 through March 2016 for validation. After the windows are shifted forward one month at a time, the twelfth fold uses March 2013 through February 2016 for training and March 2016 through February 2017 for validation.

3.2.2 Hyperparameter Tuning

Predictive performance depends in part on hyperparameters, which are configuration settings specified before model training. These settings govern how the model is fitted and must therefore be selected separately from the estimated model parameters.

For each fold, the hyperparameters are tuned by randomized search. Sixty hyperparameter combinations are drawn from a predefined search space, and each combination is evaluated on the validation data. The combination with the highest validation accuracy is selected for that fold.⁴

³See the Online Appendix for the definitions of all variables.

⁴See the Online Appendix for the hyperparameter search space.

3.2.3 Retraining and Prediction for the Test Month

After tuning the hyperparameters within each fold, the study selects the fold with the highest validation accuracy and uses its hyperparameter configuration. The model is then retrained on the combined training and validation data for that fold (48 months in total) to generate the final prediction for the test month. This procedure uses the full 48-month estimation window associated with the selected configuration.

3.3 Management Earnings Forecasts

Management earnings forecasts (MEF) serve as the benchmark for the machine-learning predictions. The benchmark uses the initial forecast of next-period net income disclosed in the Summary of Financial Results and excludes all subsequent upward or downward revisions. Management's predicted sign of the next-period earnings change is defined as the sign of the difference between the initial forecast and current-period net income.

There are three reasons for using MEF as the benchmark. First, most listed firms in Japan disclose management earnings forecasts, providing forward-looking information even for firms with limited analyst coverage. Second, managers have an informational advantage because they can access internal information. Third, benchmarking the model against information available to investors facilitates an assessment of the practical usefulness of machine learning.

3.4 Evaluation of Predictive Performance

Model performance is evaluated in terms of both predictive accuracy and economic usefulness as described below.

3.4.1 Predictive Accuracy

This study evaluates predictive accuracy using accuracy and the F1-score. Both measures are derived from a confusion matrix but emphasize different aspects of classification performance. Accuracy is the proportion of correctly classified observations; however, it can be misleading when the class distribution is imbalanced. The F1-score is the harmonic mean of precision and recall. Precision is the proportion of actual earnings increases among firms predicted to have an earnings increase, whereas recall is the proportion of actual earnings increases that are correctly predicted. Because the F1-score balances precision and recall, it is the primary performance measure in this study. The area under

the receiver operating characteristic curve (AUC) is also commonly used to evaluate classification models, but it requires continuous scores or predicted probabilities. Because MEF does not provide probability information, AUC cannot be calculated on a comparable basis. Accuracy and the F1-score, therefore, permit a consistent comparison between the machine-learning model and MEF.

Because the model produces probabilities between zero and one, a cutoff is required to convert them into predicted classes. Although 0.5 is the conventional cutoff, observations with predicted probabilities between 0.4 and 0.6 constitute a low-confidence region. Excluding these observations can improve classification performance. Chen et al. [2022] reassessed the dual cutoffs of 0.4 and 0.6 using data on U.S. firms and reported improvements in both predictive accuracy and economic usefulness: AUC increased by approximately one to two percentage points, and the 12-month hedge return increased from 5.12 percent to 9.74 percent. Following Chen et al. [2022], this study classifies observations with predicted probabilities of at least 0.6 as positive and those with predicted probabilities of at most 0.4 as negative. The RF_HC (high-confidence) specification excludes observations in the low-confidence region.

3.4.2 Economic Usefulness

To assess economic usefulness, the study constructs hedge portfolios based on market-adjusted returns (MAR). For each firm-month, MAR equals the firm's monthly return minus the return on the Tokyo Stock Price Index (TOPIX). The portfolios take long positions in firms predicted to have an earnings increase ($Pr = 1$) and short positions in firms predicted to have no increase ($Pr = 0$).

The study first calculates 12-month cumulative MAR beginning three months after the fiscal year-end and assesses statistical significance using 10,000 bootstrap iterations, following Chen et al. [2022]. Specifically, hypothetical portfolios are constructed by sampling with replacement while holding the numbers of long and short positions constant. Two-tailed p -values for the observed hedge returns are calculated from the resulting bootstrap distributions.⁵

⁵Portfolio returns are measured beginning three months after fiscal year-end because the Summary of Financial Results is disclosed within 45 days of the fiscal year-end, and additional time may be required for the information to be incorporated into market prices.

Cumulative MAR is also calculated for each fiscal year over investment horizons of 3 to 24 months, and the resulting observations are pooled for Fama–MacBeth regressions. The *t*-statistics are based on the time-series mean and standard error across periods. For horizons of 15 months or longer, the standard errors are adjusted using the Newey–West procedure with two lags to account for autocorrelation caused by overlapping return windows.

3.5 Interpretability

Finally, the study examines the information on which the machine-learning model relies. Although such models can achieve high predictive accuracy, they are often criticized as “black boxes.” In financial statement analysis and investment decision-making, understanding the factors that drive a prediction is important. The input variables are therefore classified into five groups—balance-sheet items (BS), income-statement items (PL), cash-flow-statement items (CF), variables related to management forecasts (MEF), and price-based measures (P)—and the relative importance of each group is analyzed.

Variable importance is measured in two ways. First, permutation feature importance (PFI) measures the decrease in overall predictive accuracy when the values of a variable are randomly permuted. Second, the SHapley Additive exPlanations (SHAP) method uses Shapley values from game theory to estimate each variable’s contribution to an individual prediction. Together, these methods indicate, for example, whether the model assigns greater importance to BS or PL information and how strongly management forecast variables contribute to the predictions.

4. Sample and Data

The analysis covers January 2002 through December 2021. The financial and stock-price data used to construct the prediction models are organized monthly. The test period comprises the 15 fiscal years ending in March from 2007 through 2021, and predictive accuracy is evaluated only for firms with March fiscal year-ends.

The initial sample includes all firms listed on the Tokyo Stock Exchange, excluding financial institutions, such as banks, insurers, and securities firms. Observations must satisfy the following criteria: (1) the fiscal period is 12 months; (2) book value of equity is positive; (3) an initial MEF is disclosed in the Summary of Financial Results; (4) the

reporting basis—consolidated or unconsolidated—is consistent between the Summary of Financial Results and the annual securities report; (5) both documents are filed within 90 days after the fiscal year-end; and (6) all variables used in the analysis are non-missing.

During preprocessing, all input variables other than the six price-based EP and BP measures are first deflated by the number of shares outstanding at the fiscal year-end. The financial variables are then winsorized annually at the 0.5th and 99.5th percentiles. MAR is winsorized monthly at the 1st and 99th percentiles to mitigate the influence of outliers.

After these criteria are applied, the final sample comprises 44,309 firm-year observations of which 23,362 are in the test period. Financial and stock-price data are obtained from NEEDS-FinancialQUEST (Nikkei Media Marketing).

5. Results

5.1 Predictive Performance

[Table 1 about here]

Table 1 reports the annual predictive accuracy for the sign of earnings changes. RF is the baseline model. RF_HC is the high-confidence specification, which classifies only observations with predicted probabilities of at least 0.6 or at most 0.4. MEF is the benchmark based on the initial management forecast disclosed in the Summary of Financial Results.

Overall performance is similar for RF and MEF: accuracy is 68.40 percent for RF and 69.40 percent for MEF, while the corresponding F1-scores are 72.50 percent and 74.70 percent. Annual performance is also broadly comparable, although neither method consistently dominates the other.

RF_HC outperforms RF on both measures, achieving accuracy of 73.80 percent and an F1-score of 77.40 percent, the highest values among the three specifications. Excluding low-confidence observations with predicted probabilities between 0.4 and 0.6 reduces the number of ambiguous classifications and improves accuracy.

MEF exhibits greater year-to-year variation than RF and RF_HC. In fiscal 2019, for example, MEF achieves accuracy of 75.04 percent but an F1-score of only 69.92 percent, suggesting that class imbalance may partly explain the high accuracy. The comparatively

strong performance of RF_HC on both measures underscores the value of evaluating accuracy together with the F1-score.

[Table 2 about here]

Table 2 reports annual abnormal returns (hedge returns) for 12-month portfolios formed from each set of predictions. RF_HC generates the highest overall return, 9.15 percent, which is statistically significant. The higher return is associated with restricting the portfolio to high-confidence classifications. The baseline RF specification generates an overall abnormal return of 5.90 percent; its fiscal 2011 return is not statistically significant, but the returns for the remaining 14 years are significant at the 1 percent level. Thus, the machine-learning predictions are associated with economically meaningful return spreads.

MEF produces the lowest overall abnormal return, 3.87 percent, and its statistical significance varies across years. In fiscal 2011 and 2012, the returns are negative, indicating that the MEF-based strategy was unprofitable in those years under the study's research design and sample.

[Table 3 about here]

Table 3 reports hedge returns over investment horizons of 3 to 24 months. RF_HC generates the highest return at every horizon. Its returns remain high over medium- and long-term horizons, including 10.30 percent at 12 months and 10.70 percent at both 15 and 18 months, indicating that the predictive signal persists over time. MEF generates lower returns at every horizon; at 24 months, its return is 1.59 percent and is not statistically significant ($t = 0.83$).

5.2 Interpretability

[Table 4 about here]

Table 4 summarizes the variable importance by group using PFI and SHAP. For each method, it reports mean importance, aggregate importance (sum), and the group's percentage share of total importance.

Management forecast variables (MEF) make the largest contribution under both measures. The PFI mean and sum are 2.42 and 29.09, respectively, while the corresponding SHAP values are 1.81 and 21.67. MEF accounts for 43.4 percent of total PFI and 48.6 percent of total SHAP, making it the most influential variable group in the machine-learning predictions.

Price-based measures (P) rank second by mean importance under both PFI and SHAP, but fourth by aggregate importance. The PFI mean is 1.27 (sum = 7.62; 11.4 percent of total), and the SHAP mean is 0.96 (sum = 5.74; 12.9 percent of total). Because the group contains only six variables, these values indicate the substantial importance per variable. The results suggest that market prices contain information that complements accounting data.

Traditional accounting variables (PL, BS, and CF) also contribute materially to the model. PL variables have a mean importance of 0.31 under PFI and 0.20 under SHAP; their aggregate importance is 14.17 (21.1 percent of total) under PFI and 9.04 (20.3 percent) under SHAP. BS variables have a mean importance of 0.27 under PFI and 0.14 under SHAP. Because BS is the largest group, its aggregate importance reaches 13.10 (19.5 percent) under PFI and 6.71 (15.1 percent) under SHAP. CF variables have a PFI mean of 0.34 and a SHAP mean of 0.15, but their aggregate importance is lower because the group contains fewer variables: 3.10 (4.6 percent) under PFI and 1.39 (3.1 percent) under SHAP.

6. Conclusion

This study uses machine learning to predict the sign of earnings changes for Japanese listed firms and evaluates the usefulness of those predictions from several perspectives. The main findings are summarized below.

First, the overall predictive accuracy of RF is comparable to that of management forecasts, while the high-confidence RF_HC specification is more accurate than MEF. This finding indicates that using the model's estimated confidence to exclude ambiguous predictions can improve reliability.

Second, hedge portfolios constructed from the machine-learning predictions generate statistically significant and economically meaningful abnormal returns. RF_HC produces significantly positive returns in 14 of the 15 fiscal years and stable performance over horizons of 3 to 24 months. These results suggest that machine learning may be useful in investment applications.

Third, PFI and SHAP analyses show that management forecast variables make the largest contribution to the model, while traditional financial-statement variables, including PL

and BS items, also have meaningful predictive importance. These results suggest that accounting information can complement management forecasts, which necessarily incorporate managerial judgment.

This study makes three academic contributions. First, it provides evidence on the usefulness of machine-learning-based earnings prediction in the Japanese market. By explicitly incorporating management forecasts, the analysis addresses an important institutional feature of that market. Second, the study evaluates not only the predictive accuracy but also the economic value of the predictions in portfolio strategies, thereby demonstrating their decision usefulness. Third, the variable-importance analysis identifies the information that drives the model's predictions and improves their interpretability.

Several avenues remain for future research. First, models could predict continuous outcomes, such as earnings levels or growth rates, rather than only the sign of earnings changes; such forecasts may provide more useful information for investment decisions. Second, hybrid decision-support models could combine human forecasts, including management and analyst forecasts, with machine-learning predictions. Integrating the strengths of human judgment and machine learning may improve both predictive accuracy and interpretability.

References

- Nara, S., and M. Noma [2024] *Gyoseki Yoso no Jissho Bunseki: Kigyo Kodo to Anarisuto o Chushin ni* [An Empirical Analysis of Earnings Forecasts: Focusing on Corporate Behavior and Analysts], Chuokeizai-sha. (in Japanese)
- Yakabi, K., H. Kuroki, and K. Nakagawa [2024] “Nihon Kigyo Deta o Mochiita Kikai Gakushu ni yoru Rieki Henka no Yosoku” [“Predicting Earnings Change Using Machine Learning with Data from Japanese Companies”], *JSAI Technical Report, Type 2 SIG*, Vol. 2024, No. FIN-033, pp. 68–75. (in Japanese)
- Belesis, N. D., G. A. Papanastasopoulos, and A. M. Vasilatos [2023] “Predicting the Profitability of Directional Changes Using Machine Learning: Evidence from European Countries,” *Journal of Risk and Financial Management* 16(12).
- Breiman, L. [2001] “Random Forests,” *Machine Learning* 45, pp. 5–32.
- Chen, X., Y. H. Cho, Y. Dou, and B. Lev [2022] “Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data,” *Journal of Accounting Research* 60(2), pp. 467–515.
- Hunt, J. O. S., J. N. Myers, and L. A. Myers [2022] “Improving Earnings Predictions and Abnormal Returns with Machine Learning,” *Accounting Horizons* 36(1), pp. 131–149.
- Ishibashi, K., T. Iwasaki, S. Otomasa and K. Yada [2016] “Model Selection for Financial Statement Analysis: Variable Selection with Data Mining Technique,” *Procedia Computer Science* 96, pp. 1681–1690.
- Ota, K. [2006] “Determinants of Bias in Management Earnings Forecasts: Empirical Evidence from Japan,” in *International Accounting: Standards, Regulations, and Financial Reporting*, edited by G. N. Gregoriou and M. Gaber, Butterworth-Heinemann, pp. 267–294, Chapter 11.
- Ou, J. A., and S. H. Penman [1989] “Financial Statement Analysis and the Prediction of Stock Returns,” *Journal of Accounting and Economics* 11(4), pp. 295–329.

This article was accepted following a review by two anonymous referees.

An online appendix to this article is available on the website of the Securities Analysts Association of Japan.

Table 1. Comparison of Predictive Accuracy

Fiscal year	N	RF			RF_HC			MEF		
		<i>Pr</i> = 1	Accuracy	F1-score	<i>Pr</i> = 1	Accuracy	F1-score	<i>Pr</i> = 1	Accuracy	F1-score
2007	1,617	1,027	64.32**	67.64**	774	70.11**	74.24**	1,137	62.21**	67.72**
2008	1,688	833	63.45**	50.04**	518	68.94**	55.70**	1,123	52.07**	46.95**
2009	1,688	660	66.47**	67.17**	404	70.61**	67.96**	1,140	75.71**	81.40**
2010	1,684	933	70.84**	75.41**	621	77.45**	80.83**	1,244	74.82**	81.63**
2011	1,552	718	66.49**	67.09**	456	73.30**	72.35**	1,014	69.59**	74.84**
2012	1,548	947	69.57**	75.30**	646	75.00**	79.81**	1,132	71.32**	78.78**
2013	1,697	1,030	69.95**	77.09**	687	75.53**	81.29**	1,209	73.19**	81.08**
2014	1,711	1,057	67.56**	73.61**	808	72.92**	78.67**	1,079	67.45**	73.79**
2015	1,712	1,210	71.20**	78.21**	975	75.81**	82.59**	1,177	71.14**	77.85**
2016	1,720	1,107	71.69**	78.15**	828	77.96**	83.30**	1,098	69.65**	76.49**
2017	1,745	1,157	71.63**	78.09**	895	77.53**	83.13**	1,102	71.00**	77.04**
2018	1,760	1,182	68.18**	73.66**	910	74.17**	79.83**	1,112	68.75**	73.25**
2019	1,727	1,012	65.55**	66.02**	730	70.55**	71.21**	694	75.04**	69.92**
2020	751	277	67.38**	66.30**	200	68.26**	64.33**	297	68.44**	68.27**
2021	762	469	72.83**	78.73**	377	77.27**	82.89**	434	69.55**	75.27**
Total	23,362	13,619	68.40**	72.50**	9,829	73.80**	77.40**	14,992	69.40**	74.70**

Note: *Pr* = 1 denotes observations for which the earnings-change indicator *Pr* equals one. Accuracy is the proportion of correctly classified observations, whereas the F1-score is the harmonic mean of precision and recall. Both measures are reported as percentages. ** and * denote significance at the 1 percent level ($p < 0.01$) and 5 percent level ($p < 0.05$), respectively. Significance is assessed using a randomization procedure. For each fiscal year, actual outcomes are held fixed, and predicted labels are randomly generated 10,000 times to construct a null distribution. Two-tailed *p*-values are calculated according to how extreme each observed measure is relative to this distribution.

Source: Prepared by the author.

Table 2. Hedge Returns and Statistical Significance

Fiscal year	N	RF			RF_HC			MEF		
		<i>Pr</i> = 1	AR (%)	<i>p</i> -value	<i>Pr</i> = 1	AR (%)	<i>p</i> -value	<i>Pr</i> = 1	AR (%)	<i>p</i> -value
2007	1,617	1,027	9.85	0.000	774	12.41	0.000	1,137	2.06	0.142
2008	1,688	833	9.12	0.000	518	12.51	0.000	1,123	4.41	0.001
2009	1,688	660	5.39	0.001	404	8.81	0.000	1,140	8.21	0.000
2010	1,684	933	5.28	0.000	621	7.15	0.000	1,244	4.64	0.004
2011	1,552	718	0.91	0.528	456	1.27	0.460	1,014	-1.42	0.343
2012	1,548	947	5.26	0.001	646	9.18	0.000	1,132	-1.11	0.492
2013	1,697	1,030	7.21	0.000	687	9.15	0.000	1,209	2.12	0.153
2014	1,711	1,057	6.38	0.000	808	11.17	0.000	1,079	2.76	0.058
2015	1,712	1,210	4.54	0.001	975	8.72	0.000	1,177	2.04	0.127
2016	1,720	1,107	10.93	0.000	828	17.33	0.000	1,098	5.22	0.001
2017	1,745	1,157	4.49	0.003	895	7.00	0.000	1,102	1.27	0.383
2018	1,760	1,182	10.03	0.000	910	15.16	0.000	1,112	6.86	0.000
2019	1,727	1,012	7.89	0.000	730	10.01	0.000	694	3.98	0.000
2020	751	277	17.32	0.000	200	21.47	0.000	297	11.95	0.000
2021	762	469	7.53	0.000	377	10.68	0.000	434	6.77	0.001
Total	23,362	13,619	5.90	0.000	9,829	9.15	0.000	14,992	3.87	0.000

Note: AR is the 12-month market-adjusted return on the long portfolio minus that on the short portfolio. The *p*-value reports statistical significance. For each year, a long portfolio is formed by sampling from the full sample the same number of observations as the observed *Pr* = 1 group, and a short portfolio is formed analogously using the observed *Pr* = 0 group size. Sampling is conducted with replacement, and the procedure is repeated 10,000 times to construct the bootstrap distribution of AR. Two-tailed *p*-values are calculated according to how extreme the observed AR is relative to this distribution.

Source: Prepared by the author.

Table 3. Hedge Returns across Investment Horizons

Model		3 months	6 months	9 months	12 months	15 months	18 months	21 months	24 months
RF	(%)	6.14	7.30	7.71	7.48	8.21	7.94	7.42	4.96
		[8.90]	[9.37]	[8.26]	[7.66]	[7.05]	[4.91]	[7.47]	[2.05]
RF_HC	(%)	6.57	8.50	9.58	10.30	10.70	10.70	9.96	7.50
		[8.66]	[10.80]	[10.50]	[9.29]	[8.41]	[5.87]	[9.29]	[2.27]
MEF	(%)	5.10	5.34	4.93	3.98	4.84	4.19	4.13	1.59
		[10.90]	[7.64]	[5.59]	[4.37]	[4.66]	[2.64]	[4.14]	[0.83]

Note: Statistical significance at each horizon is evaluated using annual Fama–MacBeth regressions; *t*-statistics are reported in brackets. For horizons of 15 months or longer, the standard errors are adjusted using the Newey–West procedure with two lags to account for autocorrelation arising from overlapping investment horizons.

Source: Prepared by the author.

Table 4. Mean and Aggregate Importance by Variable Group

Group	Number of variables	PFI			SHAP		
		Mean	Sum	% of sum	Mean	Sum	% of sum
MEF	12	2.42	29.09	43.4	1.81	21.67	48.6
P	6	1.27	7.62	11.4	0.96	5.74	12.9
PL	45	0.31	14.17	21.1	0.20	9.04	20.3
CF	9	0.34	3.10	4.6	0.15	1.39	3.1
BS	48	0.27	13.10	19.5	0.14	6.71	15.1

Note: “Mean” is the average annual importance of a variable over the 15 fiscal years from 2007 through 2021. “Sum” is the sum of these variable-level means within each group, and “% of sum” is the group’s share of the aggregate sum across all groups. Because the groups contain different numbers of variables, the mean facilitates a comparison of importance per variable, whereas the sum reflects each group’s aggregate contribution.

Source: Prepared by the author.

Online Appendix

Appendix A. Variable Definitions

No.	Variable	Definition	NEEDS code / formula
Category I: Income Statement Items			
1	E_t	Net income	D01110
2	$SALE_t$	Sales and operating revenue	D01021
3	$COGS_t$	Cost of goods sold	D01024
4	GP_t	Gross profit	D01026
5	$XSGA_t$	Selling, general, and administrative expenses	D01027
6	DP_t	Depreciation expense	F01022
7	OI_t	Operating income	D01029
8	$XINT_t$	Interest and discount expense	D01047
9	$NOPI_t$	Non-operating income (loss)	D01030 - D01046
10	SPI_t	Special items	D01067 - D01081
11	PI_t	Ordinary income	D01097
12	TXT_t	Income taxes	D01101
13	XAD_t	Advertising expense	K01069
14	XRD_t	Research and development expense	H01033
15	DVC_t	Total dividends on common stock	A01093 + A01114
Category II: Balance Sheet Items			
16	AT_t	Total assets	B01110
17	ATC_t	Current assets	B01021
18	LCT_t	Current liabilities	C01021
19	LT_t	Total liabilities	C01082
20	CEQ_t	Common equity	C01106
21	CHE_t	Cash and short-term financial assets	B01022 + B01033
22	$INVT_t$	Inventories	B01034
23	$RECT_t$	Accounts receivable	B01023
24	$PPENT_t$	Property, plant, and equipment	B01063
25	IVA_t	Investments and other assets	B01084
26	$INTAN_t$	Intangible assets	B01076
27	AP_t	Accounts payable and notes payable	C01022
28	DLC_t	Short-term borrowings and bonds, total	C01026
29	TXP_t	Income taxes payable	C01040
30	$DLTT_t$	Long-term borrowings, bonds, and convertible bonds	C01058
31	RE_t	Retained earnings	C01092
Category III: Cash Flow Statement Items			
32	OCF_t	Cash flow from operating activities	F01065
33	ICF_t	Cash flow from investing activities	F01087
34	FCF_t	Cash flow from financing activities	F01106
Category IV: Measures of Management Forecast Accuracy (Forecast Errors and Biases)			
35	MF_t	Management earnings forecast issued in period t	lead(CE006)
36	MFB_t	Realized forecast bias	CE006 - D01110
37	$MFAE_t$	Realized absolute forecast error	abs(CE006 - D01110)
38	MFS_t	Ex ante forecast bias	lead(CE006) - D01110
Category V: Price-Based Ratio Measures			
39	EP_t	Net income divided by market capitalization three months after fiscal year-end	D01110 / STOCKMCMKTV
40	BP_t	Common equity divided by market capitalization three months after fiscal year-end	C01106 / STOCKMCMKTV
First Differences (dif) and Lags of the Above Variables			
41–80	$E_{dif,t} - BP_{dif,t}$	First differences of each variable	—
81–120	$E_{t-1} - BP_{t-1}$	Lags of each variable	—

Source: Prepared by the author.

Appendix B. Hyperparameter Search Space

Hyperparameter	Candidate values	Description
n_estimators	500, 600, ..., 1500	Number of decision trees
max_samples	0.5	Fraction of observations sampled for each tree
max_features	0.5	Fraction of input variables considered at each split
max_depth	3, 7, 11, 15	Maximum decision-tree depth
min_samples_leaf	4, 7, 10, 13	Minimum number of observations in a leaf node
class_weight	"balanced_subsample"	Class-weighting scheme
min_samples_split	5, 7, 9, 11	Minimum number of observations required to split a node

Note: The table reports candidate settings for RandomForestClassifier in scikit-learn, a Python machine-learning library.

Source: Prepared by the author.

Appendix C. PFI and SHAP for Individual Variables

PFI				SHAP			
Rank	Variable	Mean	Std. dev.	Rank	Variable	Mean	Std. dev.
1	MFS	0.150	0.027	1	MFS	0.102	0.047
2	MF_dif	0.064	0.013	2	MF_dif	0.051	0.023
3	MFS_dif	0.035	0.010	3	MFS_dif	0.037	0.018
4	EP	0.028	0.009	4	EP	0.026	0.019
5	EP_dif	0.028	0.017	5	EP_dif	0.019	0.016
6	OI_dif	0.012	0.007	6	SPI	0.010	0.010
7	SPI	0.012	0.005	7	OI_dif	0.008	0.007
8	INVT_dif	0.008	0.005	8	E	0.007	0.010
9	MFB	0.007	0.002	9	RE_dif	0.006	0.007
10	BP	0.007	0.003	10	BP	0.005	0.006
11	MFB_lag	0.007	0.003	11	MFS_lag	0.005	0.004
12	TXP_dif	0.006	0.003	12	INVT_dif	0.004	0.004
13	MFS_lag	0.006	0.002	13	PI	0.004	0.007
14	MFAE_dif	0.006	0.002	14	MFB_lag	0.004	0.004
15	SPI_dif	0.006	0.003	15	NOPI_dif	0.004	0.004

Source: Prepared by the author.