Dissecting Machine Learning Return Predictability: With Classification[1]

Yang Bai

Kuntara Pukthuanthong

First Draft: May 1, 2020

This Draft: August 1, 2023

Abstract

Machine learning classifiers have proven highly effective in predicting return deciles both statistically and economically. To fully understand their performance, we analyze their prediction precision and information incompleteness and examine how a list of predictors can impact their accuracy. Machines excel at predicting returns at the center and tails of the return distribution, taking advantage of uneven return transition probabilities. Our research reveals that future returns are negatively correlated with past prediction precision and information incompleteness. Lastly, while unpredictable firms may have strong governance and stakeholder approval, they are also prone to accounting issues and SEC enforcement.

Keywords: Artificial neural network, classification, gradient boosting tree, information entropy, information uncertainty, machine learning, portfolio allocation, out-of-sample prediction, random forest.

JEL Classification: C14, C38, C55, G11, G14Introduction

[1] Yang Bai (yangbai@mail.missouri.edu) and Kuntara Pukthuanthong (pukthuanthongk@missouri.edu) are from Robert J. Trulaske College of Business, University of Missouri.

This paper was previously circulated under the title "Machine Learning Classification Methods and Portfolio Allocation". We thank Ahmed Guecioueur (discussant), Massimiliano Caporin (discussant), and seminar and conference participants at University of Missouri, AFA 2021 Ph.D. Poster Session, Crowell Prize 2020 Seminar, University of Miami Winter Conference of Machine Learning and Business 2020, and World Finance Conference 2021. All errors are our own.

The accuracy of cross-sectional return predictions using machine learning regressions has been validated by studies conducted by Gu et al. (2020) and Chen et al. (2023). However, the opacity of machine learning can make it difficult to comprehend its predictive power. To gain a deeper understanding of the predictability of companies, we have employed classification methods to forecast cross-sectional return deciles. This offers a fresh perspective on how machines make predictions and how uncertainty can affect their accuracy.

The conventional approach to modeling returns involves minimizing the difference between estimated and realized returns, but this does not shed much light on the stock allocation decision-making process. We better understand the return states and their transitions by quantifying returns as deciles and directly modeling the portfolio allocation process. This alternative approach gives us a more comprehensive understanding of the capabilities of machine learning.

This paper focuses on tree models and neural networks due to their superior performance and our objective of understanding machine learning return predictability. We argue that our predictions represent publicly available information by synthesizing information based on a comprehensive set of characteristics. Through head-to-head comparisons and grid searches with a wide range of candidate parameters, we indicate that classifiers are better at placing individual stocks into correct future deciles with a precision of over 15% compared to machine learning regressions, which achieve a precision of 12%.[2] Our analysis shows that machines achieve higher precision in center predictions and return

[2] Both our classifiers and their benchmark machine learning regressions deliver statistically meaningful prediction precision compared to the precision of the naïve classifier at 10%. The naïve classifier is the raw benchmark in machine learning that predictively assigns each observation to the majority class and provides a benchmark of 10% in our case. Models producing higher precision than the naïve classifier are considered as producing statistically significant predictability. This comparison is similar to the comparison between the numerical predictions against the historical means in Goyal and Welch (2007) or Gu et al. (2020).

distributions' tails, with a more pronounced imbalance in the lowest return decile. Aggregated predictions based on individual classifiers deliver 49% precision and resonate with the arbitrage asymmetry.[3] Additionally, using exogenous shocks, we demonstrate that a sudden increase in macroeconomic uncertainty promptly affects the machine's prediction precision at the aggregated level, causing it to instantaneously slump.[4]

Based on our examination of the transition matrix, we find that there is an imbalance in the transition of return states in deciles. When compared to a random distribution, transitions from extreme deciles to other extreme deciles and from middle deciles to other middle deciles are more certain, deviating from the random distribution by about 1%. For instance, there is a 1.8% probability of transitioning from the lowest decile to the highest decile. The classifiers utilize this nonlinearity in the transition probabilities to achieve the highest performance in both the middle and extreme deciles. We also use Shannon (1948)'s information entropy to measure the information incompleteness, which represents the expected minimum number of binary questions required to eliminate prediction uncertainty. Our findings show that the information incompleteness replicates the nonlinear structure of the transition matrix of return deciles as proposed by Shannon (1948).[5] Overall, we confirm that machines benefit the most from the least uncertain predictions.[6]

[3] For example, see Dong et al. (2021) and Stambaugh et al. (2015).

[4] We adopt shocks that are not obviously endogenous, including 9-11 attack, Hurricane Katrina, Hurricane Maria, and COVID-19 outbreak.

[5] Because it measures the expected minimum binary questions that need to be answered to completely resolve the prediction uncertainty, the unit of the information incompleteness is "bit", the standard unit of information. Firms with greater information incompleteness thus require more information to resolve return prediction uncertainty.

[6] Such heterogeneity in return predictability can signal different levels of market efficiency, i.e., market efficiency level is a function of firm characteristics.

According to our predictions, the long-short portfolios with monthly rebalancing can achieve a 5-factor adjusted alpha of 2.1%, Sharpe ratios as high as 2.72, and average monthly excess returns above 2.3%. Our models' statistical and economic performance highlights their ability to capture the market prediction accurately.[7]

We use the prediction success variable to examine which characteristics increase or decrease prediction precision. This dummy variable has a value of 1 when the realized return decile matches the predicted return decile. Variables such as change in momentum and return on assets contribute to the precision of machine predictions. Shannon's information entropy is used to measure information incompleteness, which shows that predictors like firm age and change in momentum reduce information incompleteness. However, predictors such as analyst forecast dispersion increases the information's incompleteness.

We are exploring the consequences of predictable returns from machine learning using new measures of prediction precision and information incompleteness. Our focus is on the impact on pricing and the corporate information environment. The process of predicting returns reflects uncertainty in the available information. To measure this, researchers have used variables such as firm age and analyst forecast dispersion to proxy the difficulty in forecasting firm value due to information uncertainty. The findings are mixed, with some studies showing a negative association between information uncertainty and stock returns, while others predict the opposite.

We argue that information uncertainty comprises two componenetns: information incompleteness and prediction precision and they specifically related

[7] Holding everything equal, the classifiers deliver performance that is comparable if not better than the machine learning regressions both statistically and economically.

to return predictions. We confirm the theoretical prediction that lower prediction precision or higher information uncertainty leads to higher stock returns. Our empirical evidence supports this finding, with both variables showing a negative relationship to future returns at the stock level. Specifically, a 1% increase in the past 12-month prediction precision is related to a 0.02% reduction in the stock return. We also show a single-bit increase in the information incompleteness in terms of information entropy is connected to a 4-5% reduction in the future stock return. ~~Additionally, we demonstrate that the information incompleteness based on predicted probabilities for the next period is negatively related to the prediction precision in the next period at the individual stock level.~~

Easley and O'Hara (2004) suggest that firms can endogenously choose the information environment to achieve specific goals, such as deciding the cost of capital. For example, Clement et al. (2003) show that firms can voluntarily disclose management's earnings forecast, reducing the information uncertainty and influencing pricing and analysts. In our test on the information environment, we examine the relation between our information uncertainty measures and firm-level accounting quality using firm-year data.

We show that the accuracy of predictions has little impact on accounting quality. Instead, it positively affects Altman's Z Score and improves the financial stability of a company. However, incomplete information negatively correlates with all our accounting quality measures. This includes statements that are difficult to understand, a greater likelihood of fraudulent accounting, and lower financial stability (Bonsall et al., 2017; Dechow et al., 1995; Altman, 1968). Most significantly, for every increase of one bit in information incompleteness, a company's risk for restatement increases by 5%, which is a major concern for accounting quality.

A significant issue to consider is the connection between the predictability of returns and the quality of a company. To begin, we examine firms that are difficult to predict due to incomplete information. Our findings indicate that companies that are hard to predict typically have a better governance structure and a CEO with decentralized power. These firms are typically larger and have better outcomes regarding governance. They take fewer risks, comply better with non-financial regulations, and are less prone to external attacks.[8] As a result, they tend to have better stakeholder approval, as evidenced by fewer litigation cases. Conversely, predictable firms are usually smaller and take higher risks. They also receive more disapproval from stakeholders, particularly shareholders.

How might financial regulators respond to the variation in return predictability? As we examine the information environment, we aim to uncover how regulators will react. This will enable us to gain a better understanding of the economic significance of (un)predictable returns. On one hand, the SEC's primary objective is to safeguard investors. When predictability decreases, investors' profitability suffers, and this can result in increased scrutiny. Additionally, the correlation between return predictability and accounting quality can cause the SEC to scrutinize unpredictable firms more intensely, especially given the existence of the Accounting Quality Model as an enforcement targeting filter (AQM).[9] On the other hand, the SEC also seeks to maintain efficient markets, and predictability is a sign of inefficiency.

Our study involves analyzing the SEC comment letter, SEC investigation in secrecy, and the fraud period of the SEC's Accounting and Auditing Enforcement

[8] We proxy risk-taking (non-financial regulation/external attacks on operation) by R&D expenditure (Environment Protection Agency's enforcement/cyberattack).

[9] See the speech given by Lewis, the Chief Economist and Director of the Division of Risk, Strategy, and Financial Innovation at the SEC: https://www.sec.gov/news/speech/2012-spch121312cmlhtm.

Release (AAER) actions by means of dummy variables. Our findings indicate that our first hypothesis regarding investor protection holds true. We find that companies with higher accuracy in their predictions are less likely to receive SEC comment letters for their 10K filings. On the other hand, companies with incomplete information are at a greater risk of receiving comment letters and being investigated. A one-bit increase in information incompleteness leads to a 10% rise in the chances of receiving the SEC's comment letter and a 7% increase in the risk of the SEC's private investigation, which can lead to enforcement actions.

Through thorough analysis, we create investment portfolios based on our aggregated predictions. We formed two portfolios using stocks with the top 10% prediction precision and the bottom 10% information incompleteness. By doing so, we anticipate that our conditional portfolios will perform better if our measures accurately capture return information uncertainty. The performance of our conditional portfolios exceeds our expectations. The long-short portfolios' average monthly return based on past 12-month precision is an impressive 62.3%, with a Sharpe ratio of 13.47 in an equal-weight scheme. The value-weighted portfolios have also performed comparably. Additionally, the long-short portfolios' monthly returns based on information incompleteness are as high as 7% for equal weights and 6% for value weights. These findings highlight the importance of momentum in prediction precision and information incompleteness. Stocks that are easy to predict in the past are more likely to remain predictable in the future.

This paper is organized as follows. Section 1 discusses the related literature. Section 2 describes the empirical modeling and introduces the measures. Section 3 reports the modeling performance and portfolio performance. Section 4 first analyzes the relation between prediction precision and the predictors and then discusses the relation between return and our predictability measures. We also

report the portfolio performance conditional on precision and information incompleteness. Section 5 reports the performance of the enhanced portfolios from the classification conditional on precision and information incompleteness. Section 6 analyzes the information environment of firms with respect to return predictability. Section 7 concludes the paper.

1. Related Literature

This research paper makes three contributions to the finance and accounting literature. Firstly, we present a novel approach to the cross-sectional return prediction problem by framing it as a machine-learning classification problem. This provides an alternative perspective on return predictions and contributes to the application of machine learning in asset pricing. The literature has traditionally focused on modeling returns directly, which limits our understanding of potential angles and does not reflect the portfolio allocation decision-making process. However, our research demonstrates that the direct modeling of return states is a viable strategy, despite concerns about the loss of information when modeling returns as categorical variables.

Secondly, previous literature on asset pricing has predominantly used machine learning regressions. For example, Gu et al. (2020) and Chen et al. (2023) survey popular algorithms for predicting stock returns using regression. Bali et al. (2023) and Bianchi et al. (2021) apply the same approach to stock options and the bond market. Li and Rossi (2020) use this approach for mutual fund selections, while Aubry et al. (2023) employ neural networks to predict art auction prices. However, our study is among the first to apply multi-class classification to return prediction and demonstrates its impressive performance in out-of-sample predictions.

Overall, our findings provide valuable insights into the potential of machine learning in asset pricing and contribute to the ongoing conversation in the finance and accounting literature.

Our analysis involves measuring the accuracy of predictions made by machines and gaining insights into their predictive abilities. We discover that the machines' approach closely resembles the market's underlying logic, where there are imbalanced transition probabilities from an old return decile to a new return decile at the individual stock level. We also introduce exogenous macroeconomic shocks, such as the 9/11 attack, and notice a significant decline in prediction accuracy, indicating an increased economic uncertainty. Additionally, our results indicate the existence of momentum in cross-sectional predictability from machines. By using long-short portfolios that are conditional on past prediction accuracy, we are able to achieve annualized Sharpe ratios as high as 13.

Our proposal presents a novel application of prediction results that delves into the effects of return predictability and firm-level information environment. This is a significant contribution that fills a gap in the literature of information uncertainty. Our approach provides a fresh and empirical perspective on information uncertainty by breaking it down into two distinct aspects that are optimized through the condensation of high-dimensional information. While previous studies focused mainly on the return consequences of information asymmetry through liquidity measures (Acharya and Pedersen, 2005; Amihud and Mendelson, 1987), our research sheds light on the valuation consequences of information uncertainty, which has been a largely understudied area.

In their studies on stock returns, Jiang et al. (2005) and Zhang (2006) find that information uncertainty, as measured by factors like firm age or analyst coverage, can have a significant impact on portfolio performance. This can lead to

lower future returns. However, these studies have certain limitations. Firstly, the proxies utilized by these researchers lack a clear relationship with returns through optimization, making it challenging to understand what they actually capture. Secondly, while using individual variables as proxies is simpler, investors and traders consider numerous firm characteristics simultaneously. Therefore, single-variable proxies from the literature underrepresent the overall market information set. [10]

It is worth noting that some studies fail to distinguish between incomplete information regarding future returns and prediction accuracy. However, one should bear in mind that prediction accuracy and information incompleteness are two distinct aspects. A prediction can be precise yet challenging to make, and vice versa. Confusing these aspects can lead to incomplete empirical conclusions. Nevertheless, through defining direct measures and examining the distinct features of information uncertainty, we can observe that the theories advanced by Easley and O'Hara (2004) and Merton (1987) can complement each other. This is precisely what we have found.

Based on classifiers, our prediction framework is a powerful tool for unifying the measures of prediction precision and information incompleteness in the stock market. We have successfully defined these measures empirically, capturing both the accuracy of predictions and the additional information required to value a firm accurately. We can link high-dimensional firm characteristics to future returns with predicted probabilities by utilizing machine learning and Shannon's information entropy. Our approach eliminates any confusion between these two distinct

[10] Martin and Nagel (2021) derive a result which also shows that the high-dimensional information can be hard for individuals to fully process. This finding implies that the measurement of the market aggregate of predictions needs to consider a comprehensive set of information.

aspects of information uncertainty. Our empirical findings confirm robustly two theories at the stock level, which were previously based on limited direct evidence.

Third, our paper also contributes to understanding accounting quality, governance, and litigation. Easley and O'Hara (2004) suggest that firms can endogenously choose their information environment to achieve corporate goals. Therefore, the information uncertainty measures may capture such choices and reflect the firm-level qualities. Therefore, we explore the information environment on the corporate side.

Extensive literature has documented the determinant and the proxies of accounting quality (Ahmed et al. 2012; Dechow et al. 1995, 2010, 2011; Hribar et al. 2014; Ghoul et al. 2021). Accounting quality also has profound consequences. For example, Bharath et al. (2008) show that accounting quality influences a firm's debt contracting. Biddle and Hilary (2004) show that firms with higher accounting quality invest more efficiently. Mcnichols and Stubben (2015) find that acquirers can price target firms with higher accounting quality in acquisition more efficiently. However, the literature has not directly analyzed the relation between return prediction uncertainty and accounting quality. We fill this gap by showing the differential relations between our information uncertainty measures and the accounting quality proxied by different variables for readability, earnings management, and financial stability. Our findings also emphasize the severe consequence of restatement related to return information uncertainty. Yet, our results also confirm that governance does not lead to bad accounting quality and that firms of information incompleteness have higher shareholder approval as proxied by shareholder litigation.

We add into the literature the understanding of the relation between information uncertainty and the SEC actions. Recent literature documents that the

SEC's resource constraints, political connections, and revolving doors can influence the SEC regulatory activities (Correia 2014; Heese 2019; Kedia and Rajgopal 2011). In particular, Holzman et al. (2023) show that the SEC would investigate firms in secrecy based on the likelihood of noncompliance, private sector scrutiny, and public trigger events. However, the literature does not show the SEC's reaction to return predictability. We take a step forward and provide novel evidence that the information uncertainty in both its aspects captures future accounting quality and can predict regulatory consequences, including the SEC's undisclosed investigation.

2. Empirical Methods

We provide a general description of our methods in this section. First, we explain the basics of our modeling process. We briefly introduce the machine learning classification methods and the training process. We also discuss the metrics we adopt in evaluating modeling performance. Next, we detail our data construction at the end of this section.

2.1 Introduction to Return Prediction as A Classification Problem

We frame the cross-sectional return prediction as a multi-class classification problem. Given a set of candidate outcomes, the classification selects the most promising outcome as a prediction. This is the foundation of our measures of information uncertainty. Following the convention of the asset pricing literature, we group individual stock returns into ten deciles per month and try to allocate each stock to its correct return decile.[11]

We refer to a strategy that makes the classification prediction as a classifier. A classifier takes the input variables and calibrates the parameters through the

[11] The classic asset pricing studies and the recent machine learning prediction studies often focus on the decile portfolios. For example, Fama and French (1992) sort stocks into deciles based β loading, while Gu et al. (2020) sort stocks into deciles based on predicted returns.

modeling architecture that maps the input variables to the probability space to minimize a loss function. Figure 1 illustrates the modeling process. Specifically, when we frame the cross-sectional return prediction as a classification problem, our optimization objective is to create a model such that the predicted probabilities distribute exactly like the observed probabilities. We follow the standard practice in multi-class classification problems and adopt a cross-entropy loss function to achieve this matching process. The cross-entropy function measures the difference between two probability distributions. A classifier will minimize the loss function below for the real return distribution cap P relative to the predicted distribution cap Q over a set of return deciles cap D.

$$L = -E_p\left[\log_2 q\right] = -\sum_{d_{it} \in D} P(d_{it}) \log_2 Q(d_{it}), \tag{1}$$

where $P(d_{it})$ is proxied empirically by the true outcome, i.e., return decile of a stock i at time t, with a value of 1 or 0.

Then, the classifier selects the return decile with the highest predicted probability as its final prediction. In the appendix, we include the benchmark machine learning regression results, and we adopt the standard mean squared error as the loss function for these benchmark models (See Gu et al. 2020).[12]

[Include Figure 1 Here]

Our main models include the standard multilayer perceptron, i.e., Artificial Neuron Network or ANN, the random forest, and the gradient-boosting trees. Our choice of models depends on two considerations. First, we want to focus on powerful models only. Second, we do not attempt to search for models with

[12] Mean squared error loss is the loss function in ordinary least square regressions. It takes the following form: $L = \frac{1}{N}\sum_{it}(y_{it} - \hat{y}_{it})^2$ for stocks i and time t.

marginal improvement in predictive power benchmark to the existing works in the literature. Instead, we want our models to be replicable and intuitive. Thus, we focus on standard models with strong predictive power.

2.2 Artificial Neural Network

Figure 2 illustrates an example of the ANN architecture in this paper. The standard ANN processes input through backpropagation, a calibration process that adjusts parameters to minimize the loss function, in a fully connected architecture, including the input layer, the hidden layer(s), and the output layer.

[Include Figure 2 Here]

In our ANN classifiers, the input layers include the firm characteristics. Then, the firm characteristics go through the fully connected hidden layers. Each neuron in a hidden layer takes the input from the prior layer with a linear function wrapped in a nonlinear function, which is again included in another linear function (See Hastie et al. 2009). The nonlinear function is referred to as activation function. The hidden layers then feed the output to the output layer in our ANN classifiers, which includes ten neurons for return deciles. Each neuron in the output layer employs a SoftMax function that translates the output from hidden layers into probabilities.[13] In the ANN regressions, the output layer includes only a regression neuron.

More specifically, consider our ANNs with multiple hidden layers. The first hidden layer includes N^1 neurons, and the neuron i^1 includes a weight vector $w^1_{m^1 j} \in W^1_{m^1}$ for the corresponding firm characteristics $x_j \in X_J$ and a bias $b^1_{m^1}$.

$$h^1_{m^1} = \sigma\left(\sum_j w^1_{m^1 j} x_j + b^1_{m^1}\right), \tag{2}$$

[13] The softmax function is a popular scaling function in regressions to model categorical response variable. For example, multinomial regression also employs softmax function.

where σ is an activation function.[14] In this paper, we have two ANN models, including a model with rectifier activation function $\sigma(a) = \max(0, a)$ and the other model with tanh activation function $\sigma(a) = \frac{\exp(a)-\exp(-a)}{\exp(a)+\exp(-a)}$. Then, $h_1^1, \cdots, h_{m^1}^1, \cdots, h_{N^1}^1$ become the input of the second hidden layer. In general, the neuron m^l in the hidden layer $l \in [1, L]$ transforms all N^{l-1} output from hidden layer $l-1$, i.e., $h_1^{l-1}, \cdots, h_{m^{l-1}}^{l-1}, \cdots, h_{N^{l-1}}^{l-1}$ with a weight vector $w_{m^l m^{l-1}}^l \in W_{m^l}^l$ and a bias b_{m^l} as the following.

$$h_{m^l}^l = \sigma\left(\sum_{m^{l-1}} w_{m^l m^{l-1}}^l h_{m^{l-1}}^{l-1} + b_{m^l}^l\right). \tag{3}$$

The output layer takes the vector input H_L from the last hidden layer and makes the final linear transformation $f_d = \sum_{m^L} w_{dm^L} h_{m^L}^L + b_d$ for output neuron of class $d \in D$, and the calculation finishes with the SoftMax function as below. Then, the predicted probabilities are compared to the realized outcomes in the cross-entropy loss function.[15]

$$Q(d) = \frac{\exp(f_d)}{\sum_D \exp(f_u)}. \tag{4}$$

2.3 Random Forest and Gradient Boosting

We include two powerful tree models, i.e., random forest and gradient-boosting trees. Both models are developed from the simple decision tree. A classic binary decision tree finds the best splitting strategies to divide a sample into pieces based on the values of the input variables sequentially such that a loss function is minimized. For each subsample from the splitting process, the tree will assign a

[14] We do not change the activation function from layer to layer in this paper.

[15] Our benchmark machine learning regressions have a different single neuron output layer that takes a hidden layer's transformation, and the result is compared to the realized return using the mean squared error loss function.

class for the classification task and a numeric value for the regression task. In other words, the decision tree dissects the response space into subspaces conditional on the input variables and gives each subspace a value.

A random forest model builds on the decision trees with bootstrap aggregating (bagging). In each bootstrapping sample, the algorithm grows a tree by recursively sample from the input variables for splitting and picks the best split-point until the node size is reached. Then, the final prediction is made by aggregating the predictions from the trees in the random forest. Usually, an equal-weighted vote is taken as the prediction for the classification problems, while the average value is taken as the prediction for the regression problems.

Consider a decision tree $T(z;\Theta) = \sum_{j\in[1,J]} \gamma_j I(z \in R_j)$, where z is an observation, γ_j is the assigned value in the region R_j, J is the number of regions. Θ denotes the collection of parameters γ_j and R_j for all the regions, and it also includes J.

In our multi-class classification task, a boosted tree will make prediction on the probability of each of the outcome classes $d \in D$ and repetitively update the prediction until the loss function is minimized. Specifically, the algorithm initiates the prediction for class d as $f_{d0} = 0$. The following boosted tree grows.

$$f_d(z) = \sum_{b\in B} T(z;\Theta), \tag{5}$$

where B is the collection of all the bootstrapping subsamples. The output of the tree enters the SoftMax function to produce a set of probability predictions as follows.

$$p_d(z) = \frac{\exp\left[f_d(z)\right]}{\sum_{d\in D} \exp\left(f_u(z)\right)}. \tag{6}$$

The algorithm calculates pseudo residuals $r_{db} = y_d - p_d(z)$ for all regions R_{jb}. Then, it updates γ_j through loss minimization and outputs an updated boosted tree.

$$f_{db}(z) = f_{db-1}(z) + \sum_{j \in [1,J]} \gamma_{jdb}\, I\big(z \in R_j\big). \tag{7}$$

The optimization process solves the parameters in a recursive manner with the bootstrapping samples.

$$\widehat{\Theta}_b = \arg\min_{\Theta_b} \sum_{i \in [1,N]} L(y, f_{b-1}(z) + T(z; \Theta_b)), \tag{8}$$

where y is the response variable of the observation z, and L is the cross-entropy function with the probabilities as the input or the mean square loss function with the numeric prediction as the input.

2.4 Modeling Strategy: Training, Grid Search, and Aggregation

We separate historical observations into training sets, validation sets, and testing sets conditional on time window. In total, we update the models four times every ten years, and the out-of-sample prediction period starts in January 1983. Figure 3 demonstrates our modeling strategy.

[Insert Figure 3 Here]

Each update of the models includes two stages. First, we fit the individual models with different combinations of architectures and hyperparameters using the training data set. Then, we make predictions in the following validation set, including the observations the models do not see during the training period. We then select the best architecture and hyperparameter combination for each model, which is applied to make out-of-sample predictions in the corresponding testing

set. The specific windows that we adopt in this paper are detailed in Appendix Table A1.

We focus on 4 models: ANN with rectifier activation function, ANN with tanh activation function, random forest, and gradient boosting tree. The main architectural hyperparameters for ANN models are the number of hidden layers and neurons in each hidden layer. In contrast, the main architectural hyperparameter for tree models is the max number of layers that the tree models can grow. We conduct a wide range search of the architectural hyperparameters, and Table 1 reports our modeling specification.

[Insert Table 1 Here]

We build two ANN models. Each will search for 30 sub-models that take the architectural specification with a shrinkage parameter. We also build two tree models. Each will search for 4 sub-models with the specified numbers of depths. We specify the number of epochs for the ANN models to 1000 times. Comparably, we specify the number of trees in the tree models to be 1000. The details of the optimization choices can be found in Appendix Table A2.

In Section 4, we report the individual model's prediction performance. However, we report the economic analyses based on the measures formed with the aggregated predictions for brevity. We take the most straightforward route and aggregate the predictions from our four models with averaging. However, we do not average the prediction directly. Instead, we take the average of the ten predicted probabilities as the aggregated probability predictions. We make the return decile predictions by selecting the decile with the highest aggregated predicted decile probability.

$$\widehat{P_{agg}}(d_{i,t}) = \frac{1}{4} \sum_{c \in \{all\ 4\ classifiers\}} \widehat{P_c}\,(d_{i,t}), \tag{9}$$

where $\widehat{P_{agg}}(d_{i,t})$ is the aggregated predicted decile probability for stock i at time t to be in decile $d \in D$ and 4 represents the total of 4 classifiers. We then use the aggregated predictions and the aggregated probabilities to calculate our measures. We discuss the measures in Section 5.

2.5 Data

Our data universe contains 3,342,486 monthly stock observations of 26,302 distinct common stocks with current returns listed on 3 major exchanges covering 196201:202112. The lagged predictors include the return decile, 102 firm characteristics, 2-digit SIC industry indicator, and 2-digit SIC industry lagged returns.[16] Specifically, we construct the firm characteristics following Green et al. (2017) and Gu et al. (2020) based on CRSP and COMPUSTAT. We start by making the data set to be ultimately CRSP centric data with no data elimination if possible.[17] We only eliminate rows with missing current returns and not common stocks (SHRCD 10, 11 or 12) listed on the major 3 exchanges (EXCHCD 1, 2 or 3). For factor model tests and risk-free rate, we obtain Fama and French's (2015) five factors from French's website. Appendix Table A3 reports the summary statistics of our prediction sample.

We also construct a separate data set for our firm-year tests of corporate implications. Easley and O'Hara (2004) mention that firms can choose their information environment to achieve strategical goals, such as choosing their risk premium. Therefore, we test the implication and investigate whether the

[16] Following Green et al. (2017) and Gu et al. (2020), we lag the annual firm characteristics by at least 6 months, we lag the quarterly firm characteristics by at least 4 months and we lag the monthly firm characteristics by at least 1 month.

[17] Our data construction avoids the problem of fluctuating number of stocks from month to month.

information uncertainty measures in this paper predictively reflect future accounting quality and future governance outcomes. To do this, we start with the Audit Analytics database for accounting variables and litigation information, and we incorporate necessary variables from I/B/E/S, RiskMetrics, and Execucomp for analyst forecast dispersion, CEO information, and CEO salary. We get the Accounting and Auditing Enforcement Releases (AAER) from the University of Southern California (Dechow et al. 2011). In addition, we also request the SEC's records of their undisclosed investigation through the Freedom of Information Act (FOIA). In the end, this separate data set starts from 2000 primarily because most of the variables from Audit Analytics start from 2000, and we would like to make the testing sample as consistent as possible. In the end, this firm-year sample contains 56,889 distinct firm-years. We detail the variable definition in the appendix Table.[18]

3. Modeling Performance

In this section, we demonstrate our models' statistical and economic performance. Successful performance is important to our objective in this paper. Suppose our models perform well in extracting return predictability from the comprehensive public information. In that case, we can then be confident of leveraging the prediction process to study the implications of return predictability.

3.1 Prediction Precision

Panel A and B in Table 2 report the precision of the predictions from our models individually and in aggregate. The best in-sample and out-of-sample model is the random forest model, delivering prediction precisions of 17.9% and 16%, respectively. The ANN models underperform the tree models in both the

[18] Although the total number of firm-year observations is 56,889, some regressions may have less observations due to limited variable availability.

training and testing sets. The ANN models produce a prediction precision of around 15.5%. The training set precision is generally higher than the testing set precision. But the deterioration is small, except for random forests. Appendix Table A3 details the parameters selected from the in-sample training and the validation process for each model. Compared to the ANN model with the Rectifier activation function, the ANN model with the Tanh activation function tends to select small models. However, the RF model prefers complex structures.

[Insert Table 2 Here]

The naïve classifier precision is the benchmark that assigns the return decile with the largest discrete decile distribution prevalence to all the observations as the predicted decile.[19] In other words, the naïve classifier's prediction maximizes the prediction precision conditional only on the past return decile distribution. Since we balanced our in-sample data following the common practice of the classification, the naïve precision is 10% for the in-sample prediction. The out-of-sample data has slightly higher naïve precision at 10.1%. The binomial tests indicate that the precisions delivered by the machine learning classifiers are statistically meaningful.[20] In other words, all the models successfully extract information of future returns from the input variables through the modeling structure. When we aggregate the predictions, the aggregated classification achieves even higher out-of-sample prediction precision at 16.1% (see Table 2 Panel B).

[19] The comparison between the prediction precision from the classifiers and the prediction precision from the naïve classifier is similar to the comparison between the prediction precision from predictive regressions and the historical mean.

[20] The binomial test is popular in testing whether two probabilities of success is equal. Because of the success is measurable in classification practice, i.e., a correct prediction is a success, the binomial test is often applied in machine learning to test if the classifier actually learns something from the data that is meaningful.

In Table 2 Panel C, we report the performance of the benchmark regression models with the same parameter and hyperparameter specifications of their classifier counterparts such that a head-to-head comparison is possible. Specifically, these machine learning regression models predict numeric returns first, and then prediction is sorted to form the decile predictions. We compare the decile predictions from the classifiers and the regressions and conclude that the classifiers achieve higher precision in allocating the stocks into the correct future deciles.

In Table 3, we report the details of the out-of-sample prediction from the aggregated predictions in confusion matrices. Panel A reports the number of observations with the predicted decile $\widehat{d_t}$ in contrast with the realized decile d_t. For example, the first row in the first column shows that the aggregated predictions place 122,627 out-of-sample observations in the predicted decile 1, and these observations also realize in decile 1 in the next period. Panel B reports the scaled version of Panel A by the number of observations in the true class. Panel C reports the scaled version of Panel A by the number of observations in the entire sample.

[Insert Table 3 Here]

Our results show that the models, on average devote the most resources to the deciles on the two tails and around the center of the return distribution. The models also achieve the highest precisions in these deciles. For example, for the real decile 1, the models spend the most resources and made 546,858 predictions, out of which 112,627 observations realize in decile 1.[21] These 112,627 observations make up 5% in the total precision out of the 100 possible combinations between

[21] We view the number of the observations allocated into a predicted decile as the total resources the machines spend on the predictions. For example, the summation of first row in Table 3 Panel A is 546,858, indicating that the machines predict this many observations as decile 1 observations. The numbers of the observations predicted to be in decile 1 to decile 10 are the following: 546,858, 221,697, 57,031, 72,270, 126,301, 435,344, 261,155, 230,693, 137,730, and 411396. Therefore, the models spend the least resources on decile 3 and decile 4.

the predicted deciles and the realized deciles. 49% of these observations realized in decile 1 are detected correctly by the aggregated predictions from the machines. While the model also gains precision from deciles 6-8 and decile 10. Appendix Table A4 also reports the prediction precision by 2-digit SIC. Industries like Forestry, Metal, Mining, and Oil & Gas Extraction demonstrate the highest prediction precision, whereas Automotive Dealers & Service Stations, Trucking & Warehousing, and Hotels & Other Lodging Places exhibit the lowest prediction precision.

3.2 Return Decile Transitions, Prediction Precision, and Information Incompleteness

We proxy the information incompleteness on a market level for individual stocks using Shannon's information entropy based on the predicted probabilities as the following (Shannon 1948).[22]

$$E_{i,t} = -\sum_{d_{i,t} \in D} \widehat{p(d_{i,t})} \log_2 \widehat{p(d_{i,t})}, \tag{10}$$

where D includes [1:10] and $\widehat{p(d_{i,t})}$ stands for the predicted probability of the event that the stock i at time t will be in the decile $d_{i,t}$. Note that our measure of information incompleteness is concurrent because it is directly from the predictions, while the measure of precision depends on past prediction accuracy.

By definition, our information incompleteness measures the expected minimum number of binary questions that needs to be answered to make 100% precision predictions and the unit of the information incompleteness is then in bits. In other words, if a stock is associated with an entropy or information

[22] Such measure is conditional on the past public information. Since the models condense the information from a comprehensive list of predictors and deliver significant performance, we argue that this information incompleteness is representative for the best predictions based on public market information.

incompleteness of three, at least three binary questions about the decile returns must be answered so that the prediction can be made without uncertainty. This also interprets as that the prediction is at least short by three bits in the prediction information.

With the measures of prediction precision and the information incompleteness, we continue to investigate the transition probabilities and their relations with the machine learning prediction precision. Table 4 presents our analyses. Panel A reports the unconditional cross-sectional return decile transition probability during our out-of-sample period. Panel B demonstrates the prediction precision from our combined model by the transitions, while Panel C reports the information incompleteness by the transitions.

[Insert Table 4 Here]

Compared with a random distribution of return transition, which should be around 1% for each decile, the unconditional transition probabilities are distributed unbalanced. First, the center of the transition matrix highlights the certainty of the return transitions from deciles 4-7 to both the center of the distribution with a transition probability around 1.2% and the certainty of the transitions from the tail deciles. Transitions from decile 1 to deciles 1 and 10 have densities of 1.7% and 1.8%, respectively. Similarly, transitions from decile 10 to deciles 1 and 10 also have greater certainty. In Panel B, the prediction precisions for each transition suggest that the machines take advantage of the unbalanced distribution of the transition probabilities. The machines achieve the highest precision for the transitions from the center deciles to the center deciles and the transitions from the extreme deciles.

Our results in Panel C of Table 4 emphasize the machines' choice from the information perspective. The table replicates the nonlinear distribution of transition probabilities from Panel A and reflects the similar nonlinearity in the information

incompleteness. The transitions in the center and the extreme transitions clearly have smaller information incompleteness, while other transitions have greater information incompleteness.

3.3 Uncertainty Shocks and Machine Learning Return Predictability

To validate our measure, we apply exogenous shocks that represent the information uncertainty. We expect the prediction precision to drop after the information uncertainty. We apply the notable exogenous events that has profound economic impacts including 9-11, Hurricane Katrina, Hurricane Mari and COVID-19. Figure 4 presents the average prediction precision from the aggregated prediction in our out-of-sample period. The results imply the precision of machine learning prediction can be a particularly good measure for information uncertainty.

[Insert Figure 4 Here]

3.4 Variable Importance

Figure 5 reports the average variable importance across the training periods for each variable. We take the average percentage of the total sum of squared error reduction to estimate the variable importance for the tree models across all the trees and the splitting nodes related to the predictors of interest. We apply the Gedeon method to computing the variable importance in the neural networks based on the summation of the squared normalized weights related to each input predictor in all the layers (Breiman 1984, 2001; Gedeon 1997; Hastie et al. 2009).

Our results show that the models draw information from different predictors. The ANN models extract information from a wider range of predictors than the tree models. Notably, the gradient boosting tree heavily relies on idiosyncratic volatility (*idiovol*), contributing 45% of the sum of squared error reduction in the model. The ANN models rely more on past industry information (*sich2*) and return decile distribution (*label10*), contributing more than 20% and more than 6% to the

neural weights, respectively. The selection effect is also obvious. Variables such as annual income (*acc* and *absacc*), industry-adjusted percentage change in capital expenditures (*pchcapx_ia*), and analysts' mean annual earnings forecast (*sfe*) contribute the least to the machines' predictions.

[Insert Figure 5 Here]

3.5 Economic Performance

Next, we turn to economic performance. Because we have several models, we focus on the portfolios constructed with the aggregation of the predictions (see Table 2 Panel B). We form both the equal-weight and the value-weight portfolios. We also include the long-short portfolios, where we short the lowest decile portfolio (the highest precision of prediction) and hold the top highest decile portfolio (the lowest precision) and the return is also adjusted with the risk-free rate. Analogously, we sell stocks that are the most predictable and long the opposite.

Table 5 reports the portfolio performance based on the decile predictions. We report several important statistics. First, we report the average excess return across the time periods. The excess return is defined as the portfolio return minus the risk-free rate. Second, we report the cumulative return in our out-of-sample period, i.e., 198301:202112. Third, we report alphas from the standard factor models including capital asset pricing model (CAPM), Fama-French 3 factor model (FF3F), and Fama-French 5 factor model (FF5F) (Fama and French 1992, 2015).[23] The alphas are obtained from fitting the following regression.

$$R_{p,t}^{e} = \alpha_{p,t} + \boldsymbol{F}_t \mathbf{B}_p + \varepsilon_{p,t}, \tag{11}$$

[23] We report Newy-West t statistics for the alphas with a lag of 6 (Newey and West 1987).

where $\boldsymbol{F}_t$ contains the factors at time t and $\mathbf{B}_p$ is the risk loadings for the portfolio p. Lastly, we report important portfolio performance including standard deviation, annualized Sharpe ratio, turnover, maximum drawdown, and average number of stocks in each portfolio.

[Insert Table 5 Here]

We define monthly Sharpe ratio as a portfolio's excess return scaled by the standard deviation of the portfolio return, and we annualize the Sharpe ratio by multiplying the monthly Sharpe ratio with $\sqrt{12}$:

$$SR_p = \frac{E(R_p - R_f)}{\sigma(R_p)} \times \sqrt{12}. \tag{12}$$

The turnover is defined as

$$Turnover = \frac{1}{n}\sum_{i=t}^{t+n}\left(\sum_j \left| w_{j,i+1} - \frac{w_{j,i}(1+r_{j,i+1})}{\sum_k w_{k,i}(1+r_{k,i+1})} \right|\right), \tag{13}$$

where $w_{j,i}$ represents the weight of stock j during month i in a portfolio (Gu et al. 2020; Neely et al. 2014). We define the maximum drawdown according to the most recent peak of the cumulative return in our sample coverage.

$$MaxDD_{t:t+n} = \min_{t:t+n}\left(\frac{Y_{i+1} - Y_i^{peak}}{Y_i^{peak}}\right), \tag{14}$$

where i is a trading month during the investment window $t: t+n$. Y_i^{peak} is the highest cumulative return until the month i.

Table 5 Panel A reports the equal-weight portfolio performance, while Panel B reports the value-weight portfolio performance. In general, the aggregate of the algorithms is good at dissecting future returns. The portfolio returns present a linear pattern with the lowest decile delivering the lowest return and the highest

decile delivering the highest return. Our portfolios deliver average excess return as high as 2.3% (1.3%) monthly for the equal-weight (value-weight) scheme. The alphas from CAPM and factor models indicate that the portfolios including stocks of returns that are away from the market median return are not explained by the standard risk factors. The long-short portfolios deliver Sharpe ratios significantly higher than the Sharpe ratios from holding the market return. The maximum drawdowns are significantly lowered in the long-short portfolios. In Appendix Table A4, we include the performance statistics for the portfolios based on the predictions including only the stocks from the top 50% market capitalization. Our findings indicate that the performance of the strategy is robust.

Our analysis reveals that stocks with lower predictability exhibit superior performance compared to more predictable ones. This outperformance is quite substantial, with equal-weighted and value-weighted portfolios having Sharpe ratios of 2.73 and 1.01 respectively. It's worth noting that investors do not need to be correct on all return predictions to outperform the market. In fact, correctly predicting stock returns for just over 16% of the time can lead to significant market-beating returns (refer to Panel B Table 2).

4. Returns and Information Uncertainty

Section 3 discusses the modeling performance, which establishes the validity of using the predictions as our proxies for measuring information uncertainty. Specifically, we focus on the information uncertainty as reflected through the predictability and the intermediate modeling uncertainty, i.e., information incompleteness. In this section, we discuss our measures of information uncertainty and their interpretation related to the stock returns.

4.1 Prediction Success, Information Incompleteness, and Characteristics

First, we study the return predictability from the machines with respect to the predictors, assuming linear relations. Specifically, we focus on two measures, i.e., the prediction success and information incompleteness. The prediction success is defined as a dummy variable with value of 1 indicating the predicted decile is the same as the realized decile, while the information incompleteness is defined in Section 3.

We perform two Fama-MacBeth regressions and regress the prediction success and the information incompleteness on the firm characteristics included in the machine learning models such that the coefficients of the Fama-MacBeth regressions indicates the marginal contribution of additional unit of value increase from the firm characteristics to the probability of the prediction success and the information incompleteness.

$$Success_{i,t} \text{ or } Info.Incomp._{i,t} = \gamma_0 + \boldsymbol{Char}_{\cdot i,t-1}\,\boldsymbol{\Gamma} + \varepsilon_{i,t}. \qquad (15)$$

[Insert Table 6 Here]

Table 6 presents the results of our analyses, and we report the significant predictors only. Panel A shows a list of variables that are related to the prediction precision. For example, a one standard deviation increase in the change of the 6-month momentum (*chmom*) is related to a 0.4% increase in the prediction precision, while return on assets (*roa*) is related to a 0.05% decrease in the return predictions. In total, 24 (30) firm characteristics are positively (negatively) related to machine learning prediction precision.

Panel B reports the results of the analysis on information incompleteness. 49 firm characteristics are positively related to information incompleteness, including variables such as analysts earnings forecast dispersion (*disp*), return on

equity (*roeq*), earnings-to-price ratio (*ep*), and beta. In comparison, 35 predictors are negatively related to the information incompleteness, including firm age (*age*), change in 6-month momentum (*chmom*), dividend yield (*dy*), and bid-ask spread (*baspread*). Specifically, for example, a standard deviation increase in analysts earnings forecast dispersion (*disp)* is associated with the increase in the information incompleteness of 0.003 bit, while one-year increase in firm age is related to a reduction of 0.012 bit in the information incompleteness.

4.2 Stock Return and Information Uncertainty

Next, we measure the information uncertainty in two ways and study the relation between stock returns and information uncertainty. Easley and O'Hara (2004) predict that the returns are negatively related to the return prediction precision. Because our machine learning classifiers present powerful performance in deciding the future return deciles, we use the prediction precision as the proxy of the market's aggregate prediction precision for each stocks and apply this precision to investigate the theoretical prediction. Specifically, we measure the market aggregate prediction precision of stock i at time t based on the prediction from the past 12 months.

$$Precision_{i,t} = \sum_{\tau \in t-12:t-1} \frac{I(d_\tau = \widehat{d_\tau})}{12}, \tag{16}$$

where I is an indicator of value 1 or 0, $\widehat{d_\tau}$ is the predicted return decile, and d_τ is the realized decile. Our calculation uses only our out-of-sample predictions. In addition, we adopt the information incompleteness calculated as the information entropy defined in Section 4 as a measure of information uncertainty that captures the additional requirement of information to make fully correct predictions.

We focus on providing individual stock-level evidence of the relationship between returns and the two aspects of information uncertainty. We hypothesize

that both prediction precision and information incompleteness can be associated with future returns. Specifically, following Green et al. (2017), we run the following Fama-MacBeth predictive regression controlling 102 firm characteristics, 2-digit SIC fixed effects, and past return deciles.

$$R_{i,t} = \gamma_0 + \gamma_1 Precision_{i,t} + \gamma_2 Info.Incomp._{i,t} + \boldsymbol{Controls_{i,t-1}}\boldsymbol{\Gamma} + \varepsilon_{i,t}, \quad (17)$$

where all regressors are based on lagged information.

We report the regression results in Table 7 Panel A covering the monthly stock data from 198301:202112 on more than 2.5 million observations. Note that the 102 firm characteristics include almost all common proxies of uncertainty directly related to or irrelated to stock return predictions. For example, firm age, monthly average of bid-ask spread (*baspread*), standard deviation of analyst earnings forecasts (*disp*), dollar value volatility (*dolvol*), number of analyst coverage (*nanalyst*), return volatility (*retvol*), earnings supprise (*sue*), among others are all included in the regressions (Green et al. 2007; Gu et al 2020; Jiang et al. 2005; Zhang 2006). Chib et al (2022) argue that the more characteristics we include in the Fama-MacBeth regression, the coefficient of the characterstics becomes pure-play and pristine. Our measures are still significant at the 0.01 level, indicating that the relation between return and our measures is strong and robust.[24]

[Insert Table 7 Here]

Easley and O'Hara (2004) predict that the precision is negatively related to the stock return. Meanwhile, Merton (1987) points out that "the effect of

[24] We follow Green et al. (2017) and report the predictive Fama-French regression estimates and statistics. In untabulated results, we show that our results are also robust under the ordinary least square (OLS) estimates with regular clustered errors at the firm level.

incomplete information on equilibrium price is similar to applying an additional discount rate", which suggests that the limitation on information content incorporated in stock price due to information asymmetry and lack of participants with private information can lead to lower stock returns (Merton 1987, p. 493). Our results are consistent with these theoretical predictions. Standalone, the prediction precision has a marginal effect of 0.022, which means that a 1% increase in the prediction precision on the market level decreases the future return of individual stock by 0.02%. The standalone regression of information incompleteness indicates that 1 bit increase in the additional information necessary to make perfect predictions will lead to 4% decrease in the future return.

More importantly, the last column in Table 7 emphasizes that the prediction precision and the incompleteness are distinct aspects of information uncertainty about returns, and they do not subsume each other's effect. Conceptually, lower precision and higher information incompleteness can be interpreted as informationally uncertain, leading to contradictions between the empirical findings and the theory.

Specifically, using proxies, Jiang et al. (2005) and Zhang (2006) find that the information uncertainty leads to lower return, which is contradictory to the theory predictions from Easley and O'Hara (2004), who have an explicit expression indicating that the returns are a function of prediction precision and that lower precision thus higher information uncertainty should be related to higher return. Through dissecting the two distinct aspects of information uncertainty, our results show that both precision and information incompleteness leads to lower individual stock returns and thus reconsolidate the empirical findings and the theories.

Additionally, leveraging our novel measures of prediction success, a dummy variable indicating correct prediction, and the information incompleteness, we

explore the relation between prediction precision and information incompleteness. Panel B in Table 7 reports the results across the models and the aggregated prediction. The results emphasize the negative relation between the probability of successful prediction and the information incompleteness.

4.3 Portfolios Conditional on Information Uncertainty

Taking advantage of the information uncertainty measures, we sort the stocks according to the predicted deciles and the two information uncertainty measures, i.e., the past 12-month prediction precision and the information incompleteness. We first construct a set of prediction-based portfolios using only the stocks from the top precision decile. Then, we construct a second set of portfolios using only the stocks from the lowest information incompleteness. Table 8 reports the portfolio performance organized in the same way as Table 5.

[Insert Table 8 Here]

Regardless of the weighting scheme, Panel A and B in Table 8 show the conditioning information of prior precision enhances the portfolio performance tremendously. The equal-weight (value-weight) long-short portfolio can deliver excess return of 62.3% (53.3%) monthly with 87 stocks in the long leg and 116 stocks in the short leg on average. The annualized Sharpe ratios are as high as 13.5. One important measure worth noting is that the maximum drawdown which becomes 0 for the highest return decile, meaning that all stocks are perfectly predicted. The factor models cannot explain the anomaly returns from the long-short portfolios. To sum up, Table 8 confirms that the precision measure captures the prediction precision. Moreover, the prediction precision observes strong continuation. Stocks that realize high prediction precision in the past will be more accurate to predict in the future.

In Panel C and D, we investigate the information incompleteness and portfolio returns. The findings are like those of the portfolios conditional on precision, despite weaker effect. Table 8 together validates our measures through the conditional portfolios and demonstrates strong enhancing power in portfolio performance. It also emphasizes the momentum in prediction precision and the information incompleteness at the individual stock level.

5. Corporate Environment

Easley and O'Hara (2004) point out that the firms can adjust their information environment endogenously to achieve strategic goals, such as moving the investors' required rate of return. Information uncertainty can also lead to different strategical outcomes. Therefore, the information measures should reflect important firm choices. Starting from this section, we put our measures into practice and investigate the consequences of the information uncertainty from the corporate perspective.

5.1 Accounting and Financial Quality

First, we turn to accounting quality. Accounting quality is an important outcome of firms' governance and directly reflects the firms' choices of their information environment. We expect that our information uncertainty measures should predictively capture the firm's future accounting quality. This is also an important robustness test of our information uncertainty measures, i.e., prediction precision and information incompleteness.

Specifically, we adopt Bog index, M Score, and Altman's Z score to measure the three aspects of accounting quality (Altman 1968; Bonsall et al. 2017; Beneish 1999). Meanwhile, we also test if our return information uncertainty predicts restatement risk of the fiscal year. To carry out this set of tests, we aggregate our

measure to fiscal years at the stock level. We calculate the fiscal year prediction precision and the fiscal year monthly prediction information incompleteness.

Then, we regress the response variables of accounting quality individually on a set of common control variables for accounting quality, including annualized stock volatility, market adjusted return, firm size, leverage, cash holding, Tobin's Q, discretionary accrual, firm age, analyst coverage, and Fortune 500 indicator. The details of the control variables are in the appendix Table TA5.

$$Acct_{i,t} = \beta_0 + \beta_1 Precision_{i,t-1} + \beta_2 Info.Incomp._{i,t-1} + \boldsymbol{Controls}\boldsymbol{\Gamma} + \varepsilon_{i,t}, \quad (18)$$

where $Acct_{i,t}$ stands for one of our accounting quality variables in fiscal year t for stock i and our regressors are lagged by one fiscal year.[25]

We report the results in Table 9. We find that the information incompleteness significantly indicates lower statement readability, higher likelihood of earnings management, and lower financial quality. More importantly, one bit of average information incompleteness in the past period for return prediction is associated with 5% probability increase for restatement. On the other hand, a 1% increase in the past-period prediction precision is associated with a 0.05% increase in the Altman's Z score, which means that the prediction precision predictively indicates financial soundness.

[Insert Table 9 Here]

[25] Inoue and Killian (2007) show that the in-sample predictability tests are statistically more powerful than out-of-sample tests. This can be especially true in a situation with limited time length, which can be challenging for separating the data into an in-sample subset and an out-of-sample subset. Therefore, we rely on the in-sample regressions to show the predictive power of our information uncertainty measures on future accounting quality and corporate governance.

5.2 Governance and Governance Outcome

Table 9 shows that the firms with severe information incompleteness can have lower accounting quality. Therefore, an immediate question is whether firms of information incompleteness are firms of bad governance. This question is important as firms with extremely bad governance can face strong limitations in real-time trading, such as severe liquidity issues, because investors can have concerns due to the governance quality.

We concentrate on two aspects of governance, i.e., governance structure and governance outcomes. First, we analyze the relation between information uncertainty of return prediction and the governance environment. Specifically, we examine the institutional ownership, the CEO-chairman duality, the CEO salary amount, and the firm size.

[Insert Table 10 Here]

We fit the regression with equation 18 and replace the response variable with the governance environment variables. Table 10 Panel A reports our analysis result on the governance environment. Our results show that precision does predict the governance environment, i.e., they do not have significant statistical relation. However, our information incompleteness is associated with higher institutional ownership, lower likelihood of CEO-chairman duality, lower CEO salary, and larger firm size. In other words, the firms of information incompleteness are firms with more mature governance environment and lower CEO power.

Next, we investigate the governance outcomes. We explore three response variables, including R&D expenditure, the enforcement from Environment Protection Agency (EPA), and the cyber-attack risk. We adopt R&D expenditure as a proxy of risk-taking, we adopt EPA enforcement to proxy for firm-level law compliance, and we adopt cyber-attack risk to proxy of firm operating risks. Table

8 Panel B shows the regression results. We find that the prediction precision is positively related to future risk-taking. A 1% increase in the prediction precision is associated with a 90-dollar increase in R&D expenditure. On the opposite side, the information incompleteness is negatively associated with the governance outcomes. One bit increase in the information incompleteness is associated with 113,000 dollars reduction in R&D expenditure, 2.4% reduction in EPA enforcement risk, and 2% reduction in cyber-attack risk.

5.3 Stakeholder Approval

We further examine the stakeholder satisfaction attributable to information uncertainty on return predictions. Specifically, we adopt litigations as our proxies and include eight response variables representing the occurrence of eight distinct types of litigations: 1. all litigation, 2. civil rights litigation, 3. environment litigation, 4. illegal activity litigation 5. intellectual property litigation, 6. labor litigation, 7. regulatory litigation, and 8. shareholder litigation. We fit the regression in equation 16 with these litigation variables as the dependent variables. Table 11 reports the results.

[Insert Table 11 Here]

In general, Table 11 shows that the precision in return prediction leads to increased litigation risks, especially in shareholder litigations. A 1% increase in the average annual precision of return prediction is associated with a 0.04% (0.03%) increase in shareholder (all types of) litigation risk. On the other hand, information incompleteness is negatively related to the litigation risk. One bit increase in the additional information needed to resolve prediction uncertainty is associated with 12% (3%/4%/6%/4%/7%) reduction in the overall (civil rights/illegal activity/intellectual property/labor/shareholder) litigation risk.

5.4 The SEC Enforcement

Next, built on our findings in accounting qualities and governance qualities, we question what actual regulatory consequences that information uncertainty will lead to. Specifically, we focus on the SEC regulatory actions, including comment letter issuance related to 10K filings, the SEC's investigation in secrecy, and the SEC's most famous enforcement actions, i.e., Accounting Auditing Enforcement (AAER enforcement). In particular, we obtain the comment letter data from Audit Analytics. The AAER enforcement data is from University of Southern California (Dechow et al. 2011). We put Freedom of Information Act (FOIA) requests to obtain the SEC's undisclosed investigation records. We repeat the regression in Equation 16 with the SEC actions as response variables. Table 12 reports the regression results.

[Insert Table 12 Here]

The SEC's missions include both investor protection and maintaining an efficient market. Therefore, the commission is motivated to increase scrutiny for the less predictable and more uncertain firms. However, predictability can also signal inefficiency of the market. Therefore, it is also plausible to assume that the SEC will increase scrutiny intensity. Our results support the first hypothesis.

The overall return information uncertainty does not lead to AAER enforcement in the next fiscal year. However, our findings indicate that the information incompleteness will increase both the comment letter risk and the SEC investigation risk. One-bit extra information needed on average to resolve return prediction uncertainty in the past year will increase the comment letter risk by 10% and the SEC investigation risk by 7%.

Firms with predictable returns are less likely to receive comment letters from the routine review of filings by the SEC's Division of Corporation Finance, which is

supposed to be focusing on the statement quality. In other words, the return predictability and the accounting quality can be two sides of the same coin, and both variables can have a strong impact on the SEC actions.

Taking together our findings from Table 9 to Table 12, our results show that the firms of higher information uncertainty are not necessarily firms with bad governance. Moreover, the information uncertainty of return prediction has substantial disciplining effect on firm governance such that the firms with higher return prediction uncertainty tend to be more conservative and thus face less bad governance outcomes.

As such, the SEC's interest in the firms of information uncertainty is not likely driven by firm governance. In fact, firms with higher return predictability face more stakeholder litigations, while firms with greater information incompleteness are associated with lower litigation risks. These litigation risk changes highlight the stakeholder satisfaction of firms with higher information incompleteness for return predictions.

6. Conclusion

In this paper, we attempt to provide an alternative perspective of modeling returns with machine learning and offer the literature intuition on 1. the source of the machine learning return predictability, 2. the consequences of uncertainty in machine learning predictability, and 3. the information environment of the predictable firms.

Specifically, we first dissect the stock returns into deciles and construct classification models to predict probabilities of future return deciles. Our models deliver statistically meaningful performance and successfully predict 16% of the return deciles, which translates to significant economic performance. Indeed, our

classification-based long-short portfolios can achieve Fama-French 5-factor adjusted alpha of 1.1 and 2.1% monthly for the value-weight and equal weight portfolios, respectively. When conditional on the top decile of historical precision, our long-short portfolios can deliver as high as 65% monthly returns or annualized Sharpe ratios as high as 13. Therefore, we argue that our models capture the market prediction in aggregate for individual stocks. Based on the models, we measure the prediction precision and information incompleteness directly.

We document that the market transition probabilities are distributed in an unbalanced way. The transitions from the tails and the center of the distribution are more certain with probabilities deviating from the random distributed probability of 1%. We show that the machines take advantage of such unbalancedness and achieve exceptional detection rate in the transition from lowest decile to lowest decile. Our measure of information incompleteness calculated as Shannon's information entropy based on predicted decile probabilities reflects that the machines replicate the nonlinearity of the transition probability matrix. In addition, we show that the machine return predictability slumps upon the exogenous macroeconomic shocks, highlighting the importance of economic uncertainty in prediction precision.

Our results show that a wide range of firm characteristics contribute to the prediction precision and the information incompleteness. For example, a one standard deviation increase in the change of the 6-month momentum (*chmom*) is related to a 0.4% increase in the prediction precision, while return on assets (*roa*) is related to a 0.05% decrease in the return predictions. 49 firm characteristics are positively related to information incompleteness, including variables such as analysts earnings forecast dispersion (*disp*), return on equity (*roeq*), earnings-to-price ratio (*ep*), and beta. In comparison, 35 predictors are negatively related to the

information incompleteness, including firm age (*age*), change in 6-month momentum (*chmom*), dividend yield (*dy*), and bid-ask spread (*baspread*).

While proxies for information uncertainty in the literature such as firm age or analysts earnings forecast dispersion are popular in the studies of information uncertainty in inferring firm values, these proxies fall short in accounting for the comprehensive set of available information related to return predictions and do not establish direct relation to return predictions through explicit modeling process (Jiang et al. 2005; Zhang 2006). More importantly, the proxies do not separate the two distinct aspects of information uncertainty on return predictions, i.e., the precision and the information incompleteness (Easley and O'Hara 2004; Merton 1987). This paper measures the information uncertainty of return predictions through the realized prediction precision and the predicted probabilities together under a unified modeling process, i.e., classification models.

We are especially interested in the consequence of the return (un)predictability and the corporate characters reflected through the information uncertainty of return prediction. First, we examine the direct pricing consequences. The literature suggests controversial conclusions. Specifically, Easley and O'Hara (2004) show that information uncertainty measured by precision in a rational expectation model is negatively related to firm value, while the empirical evidence from Jiang et al. (2004) and Zhang (2006) shows the opposite conclusion with proxy variables. We reconsolidate the two sides of the literature and show that higher information uncertainty measured as lower return prediction precision does lead to higher stock returns while greater information incompleteness leads to lower stock returns.

Firms can endogenously choose their information environment. Therefore, return predictability as captured by the machines may reflect the information

environment difference and choices. Therefore, we question what these (un)predictable firms are. We first turn to accounting quality, since the accounting quality directly influences the return predictions. Our results show that the higher information incompleteness is related to lower accounting quality in general, while the return prediction precision is related to higher financial stability. More importantly, firms with higher information incompleteness are more likely to restate their filings.

Given these differences captured through our measures, we question whether the predictable firms are bad firms. Our findings confirm that the firms of information uncertainty may not be firms of bad governance. Instead, firms with greater information incompleteness are firms with higher percentage of institutional ownership, lower CEO power, and larger firm size. Our results also emphasize the disciplining effect of information uncertainty on return prediction, i.e., firms with higher information incompleteness are more conservative in risk-taking and are more careful in compliance. Consequently, they are also associated with lower stakeholder litigation risk, meaning that they receive higher stakeholder approval. Quite the opposite, our results indicate that the shareholders initiate more lawsuits against firms that are more predictable.

Given the heterogeneity we observe in the corporate environment related to return predictability, especially the accounting quality differences, we investigate into the regulator's reaction. The SEC is on a three-part mission, among which two are related to the return predictability, i.e., investor protection and facilitating efficient markets. For example, if a stock is very unpredictable, many investors can lose their investment, then the SEC can also step in. On the other hand, if predictability is regarded as signals of inefficient markets, the SEC may pay more attention to the firms in the market segments of predictable stocks.

So does SEC care about whether a stock is predictable? We find the answer is yes. Our results support the investor protection hypothesis that the information uncertainty for return predictions can alternate SEC enforcement risk. We show that the firms with higher historical prediction precision receive comment letters less frequently, while the information incompleteness significantly increases the comment letter frequency and the SEC private investigation risk.

References

Acharya, V., and L. Pedersen, 2005, Asset Pricing with Liquidity Risk, *Journal of Financial Economics* 77, 375–410.

Ahmed, A. S., M. Neel, and D. Wang, 2013, Does Mandatory Adoption of IFRS Improve Accounting Quality? Preliminary Evidence, *Contemporary Accounting Research* 30, 1344–1372.

Altman, Edward I., 1968, Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankcrupcy, *The Journal of Finance* 23, 589–609.

Amihud, Y., and H. Mendelson, 1986, Asset Pricing and the Bid-Ask Spread, *Journal of Financial Economics* 17, 223–249.

Aubry, M., R. Kraussl, G. Manso, and C. Spaenjers, 2023, Biased Auctioneers, *The Journal of Finance* 78, 795–833.

Bali, T. G., H. Beckmeyer, M. M., and F. Weigert, 2023, Option Return Predictability with Machine Learning and Big Data, *Review of Financial Studies*.

Beneish, M. D., 1999, The Detection of Earnings Manipulation, *Financial Analysts Journal* 55, 24–36.

Bianchi, Daniele, M. Büchner, and A. Tamoni, 2021, Bond Risk Premiums with Machine Learning, *The Review of Financial Studies* 34, 1046–1089.

Biddle, G. C., and G. Hilary, 2006, Accounting Quality and Firm-Level Capital Investment, *The Accounting Review* 81, 963–982.

Bonsall, S. B., A. J. Leone, B. P. Miller, and K. Rennekamp, 2017, A plain English measure of financial reporting readability, *Journal of Accounting and Economics* 63, 329–357.

Breiman, L., 1984, *Classification and Regression Trees* (Belmont, Calif. Wadsworth International Group).

Breiman, L., 2001, Random Forests, *Machine Learning* 45, 5–32.

Chen, L., M. Pelger, and J. Zhu, 2023, Deep Learning in Asset Pricing, *Management Science*.

Clement, M., R. Frankel, and J. Miller, 2003, Confirming Management Earnings Forecasts, Earnings Uncertainty, and Stock Returns, *Journal of Accounting Research* 41, 653–679.

Correia, M. M., 2014, Political connections and SEC enforcement, *Journal of Accounting and Economics* 57, 241–262.

Dechow, P. M., R. G. Sloan, and A. P. Sweeney, 1995, Detecting Earnings Management, *The Accounting Review* 70, 193–225.

Dechow, P., W. Ge, C. R. Larson, and R. G. Sloan, 2011, Predicting Material Accounting Misstatements, *Contemporary Accounting Research* 28, 17–82.

Dechow, P., W. Ge, and C. Schrand, 2010, Understanding earnings quality: A review of the proxies, their determinants and their consequences, *Journal of Accounting and Economics* 50, 344–401.

Dong, X., Y. Li, D. E. Rapach, and G. Zhou, 2021, Anomalies and the Expected Market Return, *The Journal of Finance* 77, 639–681.

Easley, D., and M. O'Hara, 2004, Information and the Cost of Capital, *The Journal of Finance* 59, 1553–1583.

El Ghoul, Sadok, Omrane Guedhami, Yongtae Kim, and Hyo Jin Yoon, 2020, Policy Uncertainty and Accounting Quality, *The Accounting Review* 96, 233–260.

Fama, E. F., and K. R. French, 2015, A Five-Factor Asset Pricing Model, *Journal of Financial Economics* 116, 1–22.

Fama, E. F., and K. R. French, 1992, The Cross-Section of Expected Stock Returns, *The Journal of Finance* 47, 427–465.

Fama, E. F., and J. D. MacBeth, 1973, Risk, Return, and Equilibrium: Empirical Tests, *Journal of Political Economy* 81, 607–636.

Friedman, J. H., 2001, Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics* 29, 1189–1232.

Gedeon, T. D., 1997, Data Mining of Inputs: Analysing Magnitude and Functional Measures, *International Journal of Neural Systems* 08, 209–218.

Green, J., J. R. M. Hand, and X. F. Zhang, 2017, The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns, *The Review of Financial Studies* 30, 4389–4436.

Gu, S., B. Kelly, and D. Xiu, 2020, Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies* 33.

Hastie, T., R. Tibshirani, and J. Friedman, 2009, *The Elements of Statistical Learning Springer Series in Statistics* (Springer New York, New York, NY).

Heese, J., 2019, The Political Influence of Voters' Interests on SEC Enforcement, *Contemporary Accounting Research* 36, 869–903.

Holzman, E. R., N. T. Marshall, and B. A. Schmidt, 2023, When are firms on the hot seat? An analysis of SEC investigation preferences, *Journal of Accounting and Economics*.

Hribar, P., T. Kravet, and R. Wilson, 2013, A New Measure of Accounting Quality, *Review of Accounting Studies* 19, 506–538.

Jiang, G., C. M. C. Lee, and Y. Zhang, 2005, Information Uncertainty and Expected Returns, *Review of Accounting Studies* 10, 185–221.

Kedia, S., and S. Rajgopal, 2011, Do the SEC's enforcement preferences affect corporate misconduct?, *Journal of Accounting and Economics* 51, 259–278.

Kelly, B., and A. Ljungqvist, 2012, Testing Asymmetric-Information Asset Pricing Models, *Review of Financial Studies* 25, 1366–1413.

Li, B., and A. G. Rossi, 2020, Selecting Mutual Funds from the Stocks They Hold: A Machine Learning Approach, *SSRN Electronic Journal*.

McNichols, Maureen F., and Stephen R. Stubben, 2014, The Effect of Target-Firm Accounting Quality on Valuation in Acquisitions, *Review of Accounting Studies* 20, 110–140.

Merton, R. C., 2017, A Simple Model of Capital Market Equilibrium with Incomplete Information, *The Journal of Finance* 42, 483–510.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou, 2014, Forecasting the Equity Risk Premium: The Role of Technical Indicators, *Management Science* 60, 1772–1791.

Newey, W. K., and K. D. West, 1987, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703.

Shannon, C. E., 1948, A Mathematical Theory of Communication, *Bell System Technical Journal* 27, 379–423.

Stambaugh, R. F., J. Yu, and Y. Yuan, 2015, Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle, *The Journal of Finance* 70, 1903–1948.

Welch, I., and A. Goyal, 2007, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *Review of Financial Studies* 21, 1455–1508.

Zhang, X. F., 2006, Information Uncertainty and Stock Returns, *The Journal of Finance* 61, 105–137.

Table 1 Architectural Search

The table below details the main parameter choices for our models in this paper. Panel A reports the architectural search for the hyperparameters. The hyperparameters are parameters decided through the tuning process happening in the validation data sets instead of the optimization process. For our ANN models, the main architectural choice is about the number of hidden layers and the number of neurons in each hidden layer. For our tree models, the maximum number of depths that the trees can grow is the main architectural parameter. The choice column reports this information. For the ANN models, each pair of parathesis encloses an individual model. Starting from the first hidden layer following the open parathesis until the last hidden layer before the closing parathesis, each number in the parathesis represents the number of neurons in a hidden layer. If a pair of parathesis encloses n numbers, it presents an ANN model with n hidden layers. For the tree models, each number in the search choice represents a separate search of a tree model that specifies the number as the maximum depth of the tree.

Model	Hyperparameter	Search Choice
ANN (ANN Rectifier/Tanh)	1 Layers	(8), (16), (32), (64), (128)
	2 Layers	(128,64), (64,32), (32,16), (16,8)
	3 Layers	(128,64,32), (64,32,16), (32,16,8)
	4 Layers	(128,64,32,16), (64,32,16,8)
	5 Layers	(128,64,32,16,8)
	Shrinkage	L1=0.01 or 0
Tree (RF/GBT)	Depth	2,4,6,8

Table 2 Prediction Precision

This table reports the overall in-sample performance and the overall out-of-sample performance. We pull together the training set predictions, including the predictions in the validation set, to generate the statistics for the in-sample predictions below, and we do the same for the out-of-sample predictions. Panel A reports the model performance from the classifiers, and Panel B reports the out-of-sample precision of the aggregated predictions. Panel C reports the machine learning regressions with the exact same parameters and hyperparameters for a head-to-head comparison with their counterpart classifiers. The decile predictions from the regression models are based on the decile sort of predicted returns (Gu et al. 2020). The two panels are organized in the same way. Column 1 indicates whether the performance is evaluated in the sample (IS) or out of the sample (OOS). Column 2 reports the precision of the prediction. Columns 3 and 4 report the 5% and 95% bounds of the precision. Column 5 and 6 reports the binomial test results against the naïve classifier's precision. RF indicates random forest, and GBT indicates gradient boosting tree. Aggregation indicates the aggregated predictions based on all the classifiers.

Panel A: Classification Prediction Precision						
	(1)	(2)	(3)	(4)	(5)	(6)
	Data Set	Precision	5% Bound	95% Bound	Naïve Classifier Precision	Binomial Test P Value
ANN Rectifier	IS	0.157	0.156	0.157	0.100	0.000
	OOS	0.155	0.155	0.155	0.101	0.000
ANN Tanh	IS	0.154	0.154	0.154	0.100	0.000
	OOS	0.154	0.154	0.155	0.101	0.000
RF	IS	0.179	0.179	0.179	0.100	0.000
	OOS	0.160	0.159	0.160	0.101	0.000
GBT	IS	0.172	0.172	0.172	0.100	0.000
	OOS	0.159	0.159	0.159	0.101	0.000

Panel B: Out-of-Sample Aggregated Prediction Precision						
Aggregated	OOS	0.161	0.161	0.162	0.101	0.000

Panel C: Out-of-Sample Regression Prediction Precision						
	(1)	(2)	(3)	(4)	(5)	(6)
Model	Data	Precision	5% Bound	95% Bound	Naïve Classifier Precision	Binomial Test P Value
ANN Rectifier	OOS	0.126	0.126	0.127	0.101	0.000
ANN Tanh	OOS	0.129	0.128	0.129	0.101	0.000
RF	OOS	0.124	0.123	0.124	0.101	0.000
GBT	OOS	0.120	0.120	0.121	0.101	0.000

Table 3 Out-of-Sample Prediction Confusion Matrices

This table reports the out-of-sample prediction confusion matrix. Panel reports the machines' allocation of number of observations based on the aggregated predictions from all classifiers. The first column indicates the predicted decile, while the first row indicates the realized decile. For example, in the table cell of predicted decile 1 and realized decile 1, the aggregated predictions include 122,627 observations. The row summation of these numbers reflects the resources spent on the deciles by the machines. Panel B reports the scaled version of Panel A by the number of observations in the true class, while Panel C reports the scaled version of Panel A by the number of observations in the entire out-of-sample testing period. The colored blocks indicate the correct predictions. For example, in Panel B, the number 12% on the diagonal means there are 12% of real decile 2 observations are detected. In Panel C, the 5% means that out of the entire sample, 5% of the observations from real decile 1 are detected. The summation of the diagonal percentages in Panel C sum up to the total precision of the aggregated predictions ensembled from the individual classifiers.

Panel A: Out-of-sample Prediction Confusion Matrix

	d_t									
$\widehat{d_t}$	1	2	3	4	5	6	7	8	9	10
1	122627	73891	49420	35848	32758	29953	31127	36050	47802	87382
2	22132	28897	25348	21123	19137	18465	18843	21039	24204	22509
3	4115	6348	6601	6050	5797	5747	5807	5966	6011	4589
4	3526	7059	8525	8481	8137	8271	8137	8117	7472	4545
5	4010	9549	13739	15869	17885	17476	16173	14466	11189	5945
6	7641	25723	44113	56564	63520	66562	63255	54990	38509	14467
7	6507	17053	25970	31158	34405	36774	36581	33948	26666	12093
8	9745	19485	23467	24557	25248	26727	28121	29081	27974	16348
9	10044	14235	14036	13144	12413	12818	13569	15572	17757	14142
10	59041	47375	37785	30769	31508	29023	30236	34215	42804	68640

Panel B: Out-of-sample Prediction Confusion Matrix (Scaled by Total Number of Observations in the True Class)

	d_t									
$\widehat{d_t}$	1	2	3	4	5	6	7	8	9	10
1	49%	30%	20%	15%	13%	12%	12%	14%	19%	35%
2	9%	12%	10%	9%	8%	7%	7%	8%	10%	9%
3	2%	3%	3%	2%	2%	2%	2%	2%	2%	2%
4	1%	3%	3%	3%	3%	3%	3%	3%	3%	2%
5	2%	4%	6%	7%	7%	7%	6%	6%	4%	2%
6	3%	10%	18%	23%	25%	26%	25%	22%	15%	6%
7	3%	7%	10%	13%	14%	15%	15%	13%	11%	5%
8	4%	8%	9%	10%	10%	11%	11%	11%	11%	7%
9	4%	6%	6%	5%	5%	5%	5%	6%	7%	6%
10	24%	19%	15%	13%	13%	12%	12%	14%	17%	27%

Panel C: Out-of-sample Prediction Confusion Matrix (Scaled by Total Number of Observations in the Entire Sample)

	d_t									
$\widehat{d_t}$	1	2	3	4	5	6	7	8	9	10
1	5%	3%	2%	1%	1%	1%	1%	1%	2%	3%
2	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
3	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5	0%	0%	1%	1%	1%	1%	1%	1%	0%	0%
6	0%	1%	2%	2%	3%	3%	3%	2%	2%	1%

7	0%	1%	1%	1%	1%	1%	1%	1%	1%	0%
8	0%	1%	1%	1%	1%	1%	1%	1%	1%	1%
9	0%	1%	1%	1%	0%	1%	1%	1%	1%	1%
10	2%	2%	2%	1%	1%	1%	1%	1%	2%	3%

Table 4 Out-of-Sample Prediction Precision by Return Decile Transition

This table reports our analysis of machine learning return predictability during our out-of-sample period by transitions. Panel A reports the unconditional transition probabilities. Probabilities deviating from the random distribution probability 1%, regardless of the direction, indicate that the transition has higher certainty. Panel B reports the prediction precision from the aggregated model by return decile transitions. For example, our prediction managed to achieve a precision of 37.4% for the return transition from decile 1 to decile 1. Panel C reports the information incompleteness created based on the predicted probabilities by return decile transitions.

Panel A: Transition Matrix

	d_t									
d_{t-1}	1	2	3	4	5	6	7	8	9	10
1	1.7%	1.1%	0.8%	0.7%	0.7%	0.7%	0.7%	0.8%	1.0%	1.8%
2	1.1%	1.1%	1.0%	0.9%	0.9%	0.9%	0.9%	1.0%	1.1%	1.2%
3	0.9%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	0.9%
4	0.7%	0.9%	1.0%	1.1%	1.1%	1.1%	1.1%	1.1%	1.0%	0.8%
5	0.7%	0.9%	1.0%	1.1%	1.1%	1.2%	1.2%	1.1%	1.0%	0.8%
6	0.7%	0.9%	1.0%	1.1%	1.2%	1.2%	1.2%	1.1%	1.0%	0.8%
7	0.7%	0.9%	1.0%	1.1%	1.1%	1.2%	1.2%	1.1%	1.0%	0.8%
8	0.8%	1.0%	1.1%	1.1%	1.1%	1.1%	1.1%	1.1%	1.0%	0.8%
9	1.0%	1.1%	1.1%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%	0.9%
10	1.7%	1.2%	1.0%	0.8%	0.8%	0.7%	0.7%	0.8%	0.9%	1.3%

Panel B: Precision

	d_t									
d_{t-1}	1	2	3	4	5	6	7	8	9	10
1	37.4%	1.3%	0.1%	0.1%	0.8%	1.9%	4.2%	7.0%	7.8%	65.2%
2	44.4%	5.3%	0.8%	0.8%	4.0%	9.6%	12.8%	16.8%	13.2%	37.6%
3	41.0%	9.3%	1.8%	1.9%	6.2%	20.8%	18.3%	17.1%	10.4%	26.1%
4	39.2%	10.3%	2.2%	2.4%	8.2%	31.0%	18.6%	15.7%	8.6%	20.1%
5	41.0%	10.4%	2.4%	3.1%	8.1%	35.6%	18.3%	12.4%	6.6%	19.1%
6	42.4%	11.4%	2.6%	3.4%	8.3%	35.1%	18.4%	12.0%	5.7%	18.0%
7	40.7%	13.8%	4.2%	4.3%	8.6%	36.7%	18.2%	11.4%	5.1%	15.4%
8	44.5%	15.8%	4.4%	5.0%	9.1%	35.9%	12.9%	9.6%	5.0%	13.6%
9	53.3%	22.2%	4.5%	7.6%	9.4%	25.0%	10.0%	7.0%	4.6%	12.6%
10	82.5%	15.6%	2.7%	4.7%	4.4%	8.6%	4.6%	3.0%	2.9%	7.5%

Panel C: Information Incompleteness

	d_t									
d_{t-1}	1	2	3	4	5	6	7	8	9	10
1	3.16	3.20	3.22	3.22	3.22	3.23	3.23	3.23	3.22	3.17
2	3.23	3.25	3.26	3.26	3.25	3.25	3.26	3.26	3.26	3.25
3	3.24	3.26	3.25	3.25	3.24	3.24	3.24	3.25	3.26	3.26
4	3.24	3.26	3.25	3.24	3.23	3.23	3.23	3.24	3.25	3.26
5	3.24	3.25	3.24	3.23	3.23	3.23	3.23	3.23	3.25	3.25
6	3.24	3.25	3.24	3.23	3.23	3.23	3.23	3.23	3.25	3.25
7	3.24	3.25	3.24	3.23	3.23	3.23	3.23	3.24	3.25	3.25
8	3.23	3.25	3.25	3.24	3.24	3.24	3.24	3.24	3.25	3.25
9	3.22	3.25	3.25	3.25	3.25	3.25	3.25	3.26	3.26	3.24
10	3.11	3.19	3.22	3.23	3.23	3.23	3.23	3.22	3.21	3.16

Table 5 Portfolio Performance

This table reports the economic performance of the portfolios constructed based on the aggregated predictions from the individual classifiers. The statistics are calculated based on the out-of-sample period covering 198301:202112. The decile portfolios are sorted based on the predicted deciles monthly, which are the deciles with the highest predicted probabilities. The column "market" reports the performance of the buy-and-hold strategy using all common stocks in the three major exchanges. The cumulative returns are in decimal unit representing gross returns in the sample period. $\alpha's$ are for the corresponding factor models, e.g., CAPM or Fama-French 3 Factor model. The t statistics for the $\alpha's$ are Newey-West t statistics of lag 6. The performance statistics are based on excess return adjusted with risk-free rate, i.e., 30-day US treasury bill. We report annualized Sharpe ratios. Turnover is defined as the average total percentage of holding changes in absolute value. Max drawdown is defined as the max difference between current price and the most recent price peak in percentage across all months in our sample period. Panel A reports the equal-weight portfolio performance, while Panel B reports the value-weight portfolio performance. A robustness check of the portfolio performance using only the stocks above the median market capitalization of the market is included in the Appendix Table A7.

Panel A: Equal-Weight Decile Portfolios												
Statistic	Market	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Mean Excess Return	0.009	-0.003	0.002	0.002	0.005	0.004	0.008	0.011	0.013	0.015	0.023	0.023
Cumulative Return	24.199	-0.963	0.127	0.371	3.620	3.785	32.084	128.769	231.499	486.085	12669.407	40473.880
CAPM Alpha	0.000	-0.014	-0.007	-0.006	-0.003	-0.001	0.003	0.005	0.005	0.006	0.014	0.025
	(0.049)	(-4.229)	(-3.969)	(-2.911)	(-1.495)	(-0.535)	(1.645)	(3.271)	(2.677)	(3.013)	(4.258)	(10.406)
FF3F Alpha	0.000	-0.013	-0.007	-0.006	-0.004	-0.002	0.002	0.004	0.004	0.006	0.014	0.024
	(0.340)	(-5.333)	(-7.168)	(-4.445)	(-3.458)	(-1.295)	(1.882)	(6.342)	(5.111)	(6.524)	(6.418)	(12.036)
FF5F Alpha	0.002	-0.007	-0.006	-0.006	-0.005	-0.003	0.000	0.003	0.003	0.006	0.017	0.021
	(1.513)	(-3.350)	(-5.439)	(-3.902)	(-4.064)	(-1.818)	(0.170)	(5.254)	(3.877)	(6.274)	(6.643)	(12.483)
Standard Deviation	0.058	0.092	0.066	0.057	0.052	0.036	0.037	0.042	0.053	0.063	0.080	0.029
Sharpe Ratio	0.515	-0.102	0.131	0.144	0.310	0.385	0.766	0.940	0.860	0.848	1.023	2.729
Turnover	0.105	0.163	0.102	0.086	0.074	0.063	0.053	0.060	0.074	0.093	0.142	0.153
Max Drawdown	-0.607	-0.904	-0.666	-0.666	-0.676	-0.663	-0.468	-0.476	-0.539	-0.543	-0.579	-0.154
Mean N	5342	1168	474	122	154	270	930	558	493	294	879	2047

Table 5 (Continues)

Panel B: Value-Weight Decile Portfolios												
Statistic	Market	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Mean Excess Return	0.008	-0.002	0.003	0.007	0.005	0.005	0.007	0.009	0.009	0.015	0.014	0.013
Cumulative Return	19.73	-0.96	-0.04	6.331	3.045	5.462	17.575	37.591	40.521	246.082	120.288	274.392
CAPM Alpha	0.00	-0.02	-0.01	-0.003	-0.004	-0.002	0.001	0.002	0.001	0.004	0.002	0.014
	(-1.67)	(-5.02)	(-4.17)	(-1.767)	(-1.803)	(-0.928)	(0.595)	(2.834)	(0.743)	(1.792)	(0.699)	(5.701)
FF3F Alpha	0.000	-0.01	-0.01	-0.003	-0.005	-0.003	0.000	0.001	0.001	0.005	0.004	0.014
	(-1.74	(-5.8)	(-4.54)	(-1.609)	(-3.008)	(-2.180)	(-0.276)	(2.412)	(1.030)	(3.059)	(1.661)	(6.492)
FF5F Alpha	0.000	-0.006	-0.004	-0.001	-0.006	-0.004	-0.002	0.001	0.001	0.006	0.008	0.011
	(-1.003)	(-3.381)	(-2.995)	(-0.741)	(-3.732)	(-3.327)	(-3.576)	(1.126)	(0.733)	(4.043)	(3.380)	(4.809)
Standard Deviation	0.045	0.099	0.078	0.070	0.059	0.047	0.041	0.044	0.054	0.073	0.088	0.044
Sharpe Ratio	0.583	-0.058	0.135	0.336	0.280	0.379	0.602	0.692	0.607	0.691	0.558	1.011
Turnover	0.057	0.132	0.096	0.070	0.064	0.051	0.047	0.048	0.065	0.087	0.118	0.125
Max Drawdown	-0.527	-0.958	-0.824	-0.753	-0.720	-0.625	-0.502	-0.509	-0.616	-0.559	-0.702	-0.414
Mean N	5342	1168	474	122	154	270	930	558	493	294	879	2047

Table 6 Prediction Precision, Information Incompleteness, and Predictors

This table reports the Fama-MacBeth regression results in the investigation of the relation between prediction precision and firm characteristics and between information incompleteness and firm characteristics. Panel A reports the results for the prediction precision, where the prediction precision is based on the aggregated predictions from the individual classifiers. Panel B reports the results for the information incompleteness, where the measure is also based on the aggregated probability predictions from the individual classifiers. We report for only variables that are statistically significant in the linear regressions, and we split the table in to the positive column and the negative column, where the positive column reports results for variables that are positively related to the prediction precision and the negative column reports for the variables that are negatively related to the prediction precision. "FM t" represents Fama-MacBeth t statistics.

Panel A: Precision vs Firm Characteristics					
Positive Relation			Negative Relation		
	Coefficient	FM t		Coefficient	FM t
chmom	0.007	10.564	pchsale_pchinvt	-0.001	-1.962
baspread	0.021	8.839	depr	-0.001	-2.108
mve_ia	0.004	8.261	cfp	-0.001	-2.411
age	0.003	8.109	roic	-0.001	-2.527
turn	0.008	7.218	sue	-0.001	-2.553
betasq	0.015	7.001	currat	-0.003	-2.806
idiovol	0.008	6.852	cashdebt	-0.001	-2.933
mom12m	0.005	6.390	gma	-0.002	-3.102
ms	0.003	6.099	securedind1	-0.002	-3.234
dy	0.003	4.963	salecash	-0.001	-3.449
pctacc	0.001	4.690	rd_mve	-0.002	-3.518
retvol	0.011	4.615	secured	-0.002	-3.636
mom1m	0.005	4.461	divi0	-0.018	-3.973
nincr	0.001	4.328	roeq	-0.002	-4.242
agr	0.002	4.248	nanalyst	-0.003	-4.369
rd0	0.003	3.395	divi1	-0.022	-4.518
chtx	0.001	2.742	mom36m	-0.002	-4.592
absacc	0.001	2.468	rsup	-0.001	-4.717
pchcapx_ia	0.001	2.369	fgr5yr	-0.003	-4.934
sgr	0.001	2.311	sp	-0.002	-5.305
pchdepr	0.001	1.757	ep	-0.004	-5.461
lev	0.001	1.691	disp	-0.003	-5.878
saleinv	0.001	1.687	zerotrade	-0.003	-5.933
ill	0.001	1.683	std_turn	-0.004	-6.393
			cash	-0.004	-7.224
			sfe	-0.004	-7.435
			bm	-0.003	-7.991
			beta	-0.017	-9.232
			roaq	-0.007	-9.760
			mom6m	-0.011	-11.793
			Constant	0.175	11.962
			102 Characteristics	Y	

Industry FE	Y
Past Return Decile	Y
Mean NOBS	5362
Mean Adj. R^2	0.028

Panel B: Information Incompleteness vs Firm Characteristics

Positive Relation			Negative Relation		
	Coefficient	FM t		Coefficient	FM t
disp	0.006	26.790	dolvol	-0.002	-1.656
cashdebt	0.004	16.989	pchsale_pchxsga	-0.001	-1.807
sp	0.007	16.308	chempia	-0.001	-1.845
roic	0.005	16.176	invest	0.000	-1.904
fgr5yr	0.008	15.185	quick	-0.001	-1.916
roeq	0.002	13.048	lev	-0.001	-2.270
roaq	0.014	12.932	rd_sale	0.000	-2.332
gma	0.004	10.898	sgr	0.000	-2.375
hire	0.002	10.182	operprof	0.000	-2.558
mom6m	0.019	10.061	turn	-0.003	-2.792
bm	0.005	9.776	std_dolvol	-0.002	-3.929
egr	0.002	9.758	absacc	-0.002	-4.738
cash	0.007	9.657	betasq	-0.023	-5.107
pricedelay	0.002	9.588	mom1m	-0.015	-5.186
cfp	0.004	9.347	saleinv	-0.001	-5.206
securedind1	0.003	9.296	maxret	-0.004	-5.260
secured	0.005	9.161	pctacc	-0.002	-5.516
rsup	0.003	9.141	mom12m	-0.005	-6.721
divi1	0.031	8.520	pchdepr	-0.001	-6.801
lgr	0.001	8.451	rd0	-0.004	-7.313
sue	0.002	8.306	mve_ia	-0.008	-7.339
ipo1	0.024	8.038	chtx	-0.001	-7.969
currat	0.004	7.508	chfeps	-0.001	-8.650
divi0	0.026	7.179	baspread	-0.018	-9.178
ep	0.006	6.878	mve	-0.012	-9.222
beta	0.032	6.591	nincr	-0.002	-10.078
rd_mve	0.002	6.169	dy	-0.008	-10.279
salerec	0.001	6.044	ps	-0.002	-10.558
chcsho	0.001	5.805	retvol	-0.019	-11.338
convind1	0.005	5.447	agr	-0.004	-11.847
std_turn	0.002	5.227	divo0	-0.007	-12.840
nanalyst	0.003	5.036	idiovol	-0.015	-13.423
mom36m	0.004	4.958	ms	-0.006	-13.767
salecash	0.001	4.845	chmom	-0.013	-13.848
tang	0.002	4.728	age	-0.012	-23.631
aeavol	0.001	4.557			
sfe	0.003	4.276			
depr	0.001	4.212			
zerotrade	0.002	3.845			

pchgm_pchsale	0.001	3.684			
chinv	0.001	3.618			
chnanalyst	0.000	3.141			
sin1	0.005	3.112			
grcapx	0.000	2.702	Constant	3.17	717.03
ear	0.000	2.551	102 Characteristics	Y	
roavol	0.001	2.418	Industry FE	Y	
cinvest	0.000	2.203	Past Return Decile	Y	
cashpr	0.001	1.910	Mean_NOBS	5362	
herf	0.007	1.778	Mean Adj. R^2	0.589	

Table 7 Stock Return, Precision, and Information Incompleteness

This table reports for the relations between the return and information uncertainty measures and the relations between the return prediction precision and the information incompleteness. Specifically, Panel A reports for the Fama-MacBeth regression results investigating the relation between future monthly stock returns and the information uncertainty measures including the past 12-month prediction precisions and the information incompleteness computed based on the predicted probabilities. Panel B reports the results from the Fama-MacBeth regression examining the relation between the prediction precision and the information incompleteness for all individual classifiers and the aggregated predictions. In the regressions, we control for the 102 firm characteristics, industry fixed effects, and past return decile. The t statistics are Fama-MacBeth t statistics.

Panel A: Return vs Prediction Precision and Information Incompleteness			
Dependent Variable		Return	
	(1)	(2)	(3)
Precision	-0.022		-0.022
	(-5.420)		(-5.530)
Info. Incomp.		-0.039	-0.047
		(-4.743)	(-6.459)
Constant	0.063	0.182	0.213
	(1.356)	(3.346)	(4.084)
102 Characteristics	Y	Y	Y
Industry FE	Y	Y	Y
Past Return Decile	Y	Y	Y
Mean N	5362	5362	5362
Mean Adj. R^2	0.100	0.094	0.101

Panel B: Relation between Precision and Information Incompleteness					
Dependent Variable	Prediction Success				
	(1)	(2)	(3)	(4)	(5)
Models	Aggregate	Ann Tanh	Ann Rectifier	GBT	RF
Info. Incomp.	-0.440	-0.274	-0.334	-0.324	-0.588
	(-24.905)	(-37.920)	(-20.461)	(-31.996)	(-15.175)
Constant	1.571	1.087	1.221	1.255	2.072
	(28.300)	(22.682)	(23.742)	(21.267)	(16.778)
102 Characteristics	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y
Past Return Decile	Y	Y	Y	Y	Y
Mean N	5362	5362	5362	5362	5362
Mean Adj. R^2	0.032	0.034	0.034	0.030	0.033

Table 8 Portfolio Performance Conditional on Precision and Information Incompleteness

This table reports the economic performance of the ***conditional*** portfolios constructed based on the aggregated predictions from the individual classifiers using only the stocks in the highest decile of past 12-month precision and the stocks in the lowest decile of the information incompleteness. We report the results of the portfolios conditional on prediction precision (information incompleteness) in Panel A and B (C and D). The statistics are calculated based on the out-of-sample period covering 198301:202112. The decile portfolios are sorted based on the predicted deciles monthly, which are the deciles with the highest predicted probabilities. The column "market" reports the performance of the buy-and-hold strategy using all common stocks in the three major exchanges. The cumulative returns are in decimal unit representing gross returns in the sample period. $\alpha's$ are for the corresponding factor models, e.g., CAPM or Fama-French 3 Factor model. The t statistics for the $\alpha's$ are Newey-West t statistics of lag 6. The performance statistics are based on excess return adjusted with risk-free rate, i.e., 30-day US treasury bill. We report annualized Sharpe ratios. Turnover is defined as the average total percentage of holding changes in absolute value. Max drawdown is defined as the max difference between current price and the most recent price peak in percentage across all months in our sample period. Panel A and C report the equal-weight portfolio performance, while Panel B and D report the value-weight portfolio performance.

Panel A: Equal-Weight Decile Portfolios Conditional on Precision

tatistic	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Iean Excess Return	-0.261	-0.114	-0.066	-0.036	-0.013	0.010	0.035	0.067	0.118	0.365	0.623
Cumulative Return	-1.00e+00	-1.00e+00	-1.00e+00	-1.00e+00	-9.99e-01	5.43e+01	7.04e+06	8.67e+12	1.36e+22	1.03e+62	3.51e+97
CAPM Alpha	-0.270	-0.123	-0.074	-0.044	-0.020	0.003	0.028	0.059	0.108	0.351	0.618
	(-39.424)	(-30.663)	(-28.059)	(-21.338)	(-13.471)	(1.877)	(14.822)	(19.723)	(20.576)	(23.863)	(29.096)
FF3F Alpha	-0.270	-0.123	-0.074	-0.044	-0.020	0.002	0.028	0.059	0.108	0.354	0.620
	(-38.273)	(-28.938)	(-26.051)	(-21.448)	(-16.107)	(3.021)	(24.337)	(27.074)	(24.885)	(25.407)	(29.719)
FF5F Alpha	-0.266	-0.120	-0.072	-0.043	-0.020	0.002	0.028	0.059	0.109	0.355	0.530
	(-40.880)	(-29.175)	(-24.615)	(-20.785)	(-14.244)	(2.418)	(24.275)	(27.356)	(24.983)	(24.968)	(29.086)
tandard Deviation	0.074	0.066	0.057	0.052	0.045	0.045	0.048	0.057	0.077	0.157	0.160
harpe Ratio	-12.199	-6.031	-3.980	-2.435	-0.982	0.743	2.540	4.071	5.311	8.040	13.474
urnover	0.107	0.021	0.016	0.013	0.008	0.007	0.007	0.011	0.020	0.123	0.115
Iax Drawdown	-1.000	-1.000	-0.998	-0.980	-0.841	-0.487	-0.240	-0.140	-0.131	0.000	0.000
Iean N	116	47	12	15	27	93	55	49	29	87	204

Panel B: Value-Weight Decile Portfolios Conditional on Precision

tatistic	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Iean Excess Return	-0.246	-0.114	-0.065	-0.036	-0.013	0.010	0.035	0.068	0.118	0.290	0.533
Cumulative Return	-1.00e+00	-1.00e+00	-1.00e+00	-1.00e+00	-9.99e-01	5.85e+01	7.25e+06	1.06e+13	1.45e+22	6.48e+50	1.33e+86
CAPM Alpha	-0.255	-0.123	-0.073	-0.044	-0.020	0.003	0.028	0.059	0.108	0.278	0.530
	(-34.791)	(-30.709)	(-27.065)	(-20.764)	(-13.267)	(1.955)	(14.958)	(20.538)	(20.779)	(23.901)	(28.249)
FF3F Alpha	-0.255	-0.123	-0.073	-0.044	-0.020	0.003	0.028	0.059	0.109	0.279	0.532
	(-33.617)	(-29.456)	(-25.371)	(-20.578)	(-15.921)	(3.152)	(24.693)	(27.613)	(24.991)	(26.041)	(28.923)
FF5F Alpha	-0.251	-0.120	-0.072	-0.043	-0.020	0.002	0.028	0.059	0.109	0.282	0.530
	(-35.883)	(-29.636)	(-23.925)	(-19.771)	(-14.388)	(2.564)	(23.832)	(28.264)	(25.152)	(25.277)	(29.086)
tandard Deviation	0.079	0.067	0.058	0.053	0.046	0.046	0.048	0.056	0.077	0.128	0.136
harpe Ratio	-10.778	-5.910	-3.907	-2.388	-0.977	0.749	2.545	4.201	5.271	7.847	13.582
urnover	0.089	0.016	0.006	0.004	0.004	0.006	0.006	0.008	0.013	0.071	0.080
Iax Drawdown	-1.000	-1.000	-0.998	-0.982	-0.849	-0.490	-0.245	-0.149	-0.139	0.000	0.000
Iean N	116	47	12	15	27	93	55	49	29	87	204

Table 8 (Continues)

Panel C: Equal-Weight Decile Portfolios Conditional on the Lowest Decile of Information Incompleteness											
Statistic	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Mean Excess Return	-0.014	-0.002	0.005	0.005	0.005	0.010	0.012	0.014	0.019	0.067	0.078
Cumulative Return	-1.00E+00	-9.22E-01	1.75E+00	3.83E+00	6.45E+00	6.46E+01	1.59E+02	4.66E+02	1.97E+03	1.85E+11	3.81E+14
CAPM Alpha	-0.026	-0.013	-0.004	-0.003	0.001	0.005	0.007	0.008	0.010	0.054	0.078
	(-4.542)	(-4.311)	(-1.600)	(-1.102)	(0.425)	(3.354)	(4.370)	(3.561)	(3.232)	(7.111)	(12.474)
FF3F Alpha	-0.024	-0.012	-0.004	-0.004	0.000	0.004	0.005	0.006	0.010	0.057	0.078
	(-5.190)	(-5.225)	(-2.025)	(-2.020)	(-0.134)	(3.792)	(5.125)	(4.651)	(3.886)	(7.955)	(12.334)
FF5F Alpha	-0.015	-0.011	-0.003	-0.005	-0.001	0.003	0.004	0.004	0.009	0.064	0.077
	(-3.164)	(-4.117)	(-1.751)	(-2.339)	(-0.758)	(2.721)	(4.337)	(3.838)	(3.411)	(7.561)	(10.886)
Standard Deviation	0.133	0.077	0.067	0.058	0.035	0.033	0.038	0.049	0.067	0.153	0.090
Sharpe Ratio	-0.370	-0.111	0.232	0.303	0.485	1.010	1.054	1.020	0.967	1.512	2.980
Turnover	0.239	0.111	0.067	0.055	0.042	0.038	0.039	0.055	0.080	0.239	0.239
Max Drawdown	-0.967	-0.840	-0.579	-0.751	-0.614	-0.384	-0.374	-0.516	-0.637	-0.581	-0.295
Mean N	117	48	13	16	27	93	56	50	30	88	206

Table 8 (Continues)

Panel D: Value-Weight Decile Portfolios Conditional on the Lowest Decile of Information Incompleteness											
Statistic	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Mean Excess Return	-0.026	-0.003	0.004	0.005	0.007	0.008	0.009	0.010	0.018	0.037	0.060
Cumulative Return	-1.00E+00	-9.6E-01	6.1E-01	2.7E+00	1.2E+01	2.5E+01	5.0E+01	4.8E+01	9.1E+02	1.2E+05	3.81E+14
CAPM Alpha	-0.040	-0.014	-0.005	-0.003	0.001	0.003	0.004	0.003	0.009	0.022	0.059
	(-6.697)	(-4.415)	(-1.533)	(-1.444)	(0.573)	(1.810)	(2.847)	(1.407)	(2.211)	(3.408)	(8.917)
FF3F Alpha	-0.038	-0.013	-0.005	-0.004	0.000	0.002	0.003	0.002	0.009	0.024	0.059
	(-7.742)	(-4.606)	(-1.572)	(-2.234)	(0.041)	(1.416)	(2.464)	(1.145)	(2.375)	(4.266)	(9.133)
FF5F Alpha	-0.026	-0.011	-0.003	-0.005	-0.002	0.000	0.002	0.000	0.007	0.032	0.055
	(-4.693)	(-3.702)	(-0.895)	(-2.522)	(-0.896)	(0.256)	(2.091)	(0.074)	(2.115)	(4.539)	(7.666)
Standard Deviation	0.150	0.091	0.079	0.062	0.048	0.037	0.041	0.052	0.078	0.166	0.131
Sharpe Ratio	-0.601	-0.097	0.192	0.266	0.484	0.712	0.784	0.645	0.784	0.774	1.582
Turnover	0.205	0.098	0.053	0.049	0.034	0.036	0.036	0.048	0.069	0.207	0.206
Max Drawdown	-0.997	-0.881	-0.662	-0.705	-0.611	-0.374	-0.409	-0.693	-0.630	-0.666	-0.878
Mean N	117	48	13	16	27	93	56	50	30	88	206

Table 9 Information Uncertainty and Accounting Quality

This table reports the results from the regression with ordinary least square estimation using the firm-year data sample covering 2000:2021. The prediction precision is defined as the 12-month prediction precision in the lagged fiscal year, while the information incompleteness is defined in Section 5 as the average information incompleteness in the lagged fiscal year. Appendix Table A5 details the definition of the dependent and the control variables. The dependent variables are not included in the return prediction practice. The *t* statistics are robust statistics with error double clustered using industry and year.

	(1)	(2)	(3)	(4)
Variable	Bog Index	M Score	Altman Z Score	Restatement
Precision	0.264	0.006	0.047	0.001
	(1.558)	(0.976)	(4.553)	(0.083)
Info. Incomp.	1.234	0.095	-0.044	0.050
	(3.038)	(6.668)	(-1.795)	(2.498)
Volatility	0.738	0.006	0.074	-0.001
	(7.167)	(1.889)	(13.285)	(-0.176)
Market Adj. Return	-0.035	0.001	-0.015	0.001
	(-1.332)	(1.422)	(-11.986)	(0.741)
Log(Sale)	0.308	0.009	-0.031	0.004
	(8.475)	(7.480)	(-14.807)	(2.632)
Leverage	0.857	0.044	0.198	0.007
	(5.969)	(8.861)	(23.334)	(1.038)
Cash	-0.385	-0.016	-0.142	-0.013
	(-2.005)	(-2.469)	(-12.671)	(-1.451)
Tobin's Q	-0.048	0.000	-0.007	-0.001
	(-3.142)	(0.924)	(-9.361)	(-1.474)
Discretionary Accrual	0.289	0.036	0.019	0.004
	(1.436)	(6.097)	(1.832)	(0.438)
Log(Firm Age)	-0.487	0.010	0.084	-0.012
	(-4.372)	(2.893)	(13.545)	(-2.351)
Log(Analyst)	0.339	0.002	-0.017	0.005
	(6.633)	(1.205)	(-5.475)	(2.199)
Fortune 500	-0.042	-0.006	-0.018	0.002
	(-0.316)	(-1.225)	(-2.172)	(0.337)
Industry FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Error Cluster	Double Clustering	Double Clustering	Double Clustering	Double Clustering
N	45248	56889	56889	56889
Adj. R^2	0.260	-0.109	0.058	-0.126

Table 10 Information Uncertainty and Governance

This table reports the results from ordinary least square regressions. Panel A reports the results on the relation between governance and information uncertainty measures, including prediction precision and information incompleteness defined in Section 4 and 5. Panel B reports results on the relation between the outcome of governance and information uncertainty measures. The prediction precision is defined as the 12-month prediction precision in the lagged fiscal year, while the information incompleteness is defined as the average information incompleteness in the lagged fiscal year. Appendix Table A5 details the definition of the dependent and the control variables. The dependent variables are not included in the return prediction practice. The t statistics are robust statistics with error double clustered using industry and year.

Panel A: Governance				
Variable	Institutional Ownership	CEO Chairman	Salary	Size
Precision	0.021	-0.012	-17.053	-0.041
	(0.363)	(-0.524)	(-1.325)	(-2.421)
Info. Incomp.	0.579	-0.271	-294.903	0.627
	(4.293)	(-3.821)	(-8.656)	(15.836)
Volatility	-0.056	0.027	-21.507	-0.120
	(-1.749)	(1.316)	(-2.433)	(-13.204)
Market Adj. Return	0.021	0.002	1.011	0.026
	(2.444)	(0.675)	(0.591)	(12.967)
Log(Sale)	0.035	0.063	92.064	0.644
	(2.944)	(8.290)	(27.099)	(185.922)
Leverage	-0.025	0.010	-8.523	0.067
	(-0.533)	(0.411)	(-0.729)	(4.828)
Cash	0.353	0.018	19.836	-0.321
	(5.497)	(0.596)	(1.257)	(-17.711)
Tobin's Q	0.004	0.001	2.030	0.022
	(0.811)	(0.664)	(2.088)	(19.250)
Discretionary Accrual	-0.069	-0.052	-42.666	0.054
	(-1.057)	(-1.517)	(-2.338)	(3.237)
Log(Firm Age)	0.084	-0.033	56.375	0.043
	(2.341)	(-1.914)	(6.525)	(4.261)
Log(Analyst)	0.065	-0.001	12.133	0.110
	(3.805)	(-0.153)	(2.938)	(22.315)
Fortune 500	-0.038	-0.024	6.246	0.138
	(-0.875)	(-1.604)	(0.750)	(10.387)
Industry FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Error Cluster	Double Clustering	Double Clustering	Double Clustering	Double Clustering
N	50964	23651	32053	56663
Adj. R^2	-0.125	0.106	0.125	0.574

Table 10 (Continues)

Panel B: Governance Outcome			
	(1)	(2)	(3)
Variable	R&D Expenditure	EPA Enforcement	Cyber Attack
Precision	0.009	0.000	-0.001
	(2.162)	(0.065)	(-0.342)
Info. Incomp.	-0.113	-0.024	-0.020
	(-11.109)	(-1.787)	(-1.992)
Volatility	-0.009	-0.002	-0.003
	(-3.659)	(-0.650)	(-1.256)
Market Adj. Return	-0.003	0.000	0.000
	(-4.373)	(-0.643)	(0.258)
Log(Sale)	-0.012	0.005	0.003
	(-13.916)	(4.085)	(4.172)
Leverage	-0.004	-0.003	0.006
	(-1.090)	(-0.610)	(1.781)
Cash	0.025	-0.003	0.002
	(5.079)	(-0.503)	(0.474)
Tobin's Q	0.000	0.000	0.000
	(0.636)	(1.167)	(-0.806)
Discretionary Accrual	-0.027	0.002	0.000
	(-5.455)	(0.431)	(0.018)
Log(Firm Age)	0.002	0.040	-0.008
	(0.847)	(11.442)	(-3.358)
Log(Analyst)	0.000	-0.001	-0.003
	(0.061)	(-0.531)	(-2.218)
Fortune 500	0.005	-0.008	0.010
	(1.593)	(-1.771)	(3.121)
Industry FE	Y	Y	Y
Year FE	Y	Y	Y
Error Cluster	Double Clustering	Double Clustering	Double Clustering
N	50963	56889	56889
Adj. R^2	-0.115	-0.120	-0.120

Table 11 Information Uncertainty and Stakeholder Approval

This table reports the results on the relation between return prediction information uncertainty measures and the stakeholder approval proxied with litigation cases from ordinary least square regressions. The dependent variables are dummies of value 1 and 0 representing whether a firm is involved in a specified type of litigation case. The prediction precision is defined as the 12-month prediction precision in the lagged fiscal year, while the information incompleteness is defined as the average information incompleteness in the lagged fiscal year. The information incompleteness is defined in Section 5. Appendix Table A5 details the definition of the dependent and the control variables. The dependent variables are not included in the return prediction practice. The t statistics are robust statistics with error double clustered using industry and year.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variable	All Litigation	Civil Rights	Environment	Illegal	IP	Labor	Regulatory	Shareholder
Precision	0.034	0.001	-0.001	-0.010	0.002	0.013	-0.003	0.043
	(2.709)	(0.348)	(-0.242)	(-1.395)	(0.219)	(1.630)	(-1.401)	(3.878)
Info. Incomp.	-0.124	-0.035	-0.006	-0.041	-0.064	-0.045	-0.005	-0.070
	(-4.234)	(-4.152)	(-0.475)	(-2.477)	(-2.515)	(-2.347)	(-0.892)	(-2.663)
Volatility	0.096	0.000	0.007	0.008	-0.016	0.029	0.001	0.104
	(14.381)	(-0.012)	(2.369)	(2.126)	(-2.740)	(6.468)	(0.540)	(17.275)
Market Adj. Return	-0.006	0.000	0.000	0.001	-0.001	0.000	0.000	-0.008
	(-3.953)	(1.006)	(0.019)	(0.641)	(-0.649)	(-0.043)	(0.921)	(-5.975)
Log(Sale)	0.023	0.002	-0.003	0.009	0.019	0.005	0.001	0.023
	(9.395)	(2.320)	(-2.492)	(6.460)	(8.880)	(3.242)	(1.269)	(10.436)
Leverage	0.000	-0.004	-0.002	0.006	0.009	-0.001	0.000	0.017
	(0.004)	(-1.473)	(-0.434)	(0.994)	(1.039)	(-0.084)	(0.172)	(1.898)
Cash	0.010	-0.001	0.003	-0.002	-0.006	0.009	0.000	0.007
	(0.730)	(-0.290)	(0.472)	(-0.301)	(-0.544)	(0.976)	(0.170)	(0.615)
Tobin's Q	-0.003	0.000	0.000	-0.001	-0.004	-0.002	0.000	-0.004
	(-3.823)	(-0.794)	(0.943)	(-2.073)	(-5.384)	(-3.545)	(-0.574)	(-5.117)
Discretionary Accrual	0.009	0.001	-0.001	-0.003	0.013	0.006	0.001	0.018
	(0.767)	(0.231)	(-0.257)	(-0.400)	(1.246)	(0.708)	(0.543)	(1.687)
Log(Firm Age)	-0.038	0.008	-0.022	-0.001	0.083	0.013	0.003	-0.036
	(-5.073)	(3.595)	(-6.354)	(-0.235)	(12.846)	(2.564)	(1.855)	(-5.314)
Log(Analyst)	0.048	0.000	0.002	0.001	0.003	0.005	-0.001	0.049
	(13.238)	(-0.356)	(1.227)	(0.550)	(0.888)	(1.940)	(-1.526)	(15.074)
Fortune 500	0.045	0.008	0.020	0.025	0.016	0.031	0.003	0.030
	(4.519)	(2.814)	(4.481)	(4.548)	(1.816)	(4.793)	(1.466)	(3.355)
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Error Cluster	Double Clustering	Double Clustering	Double Clustering	Double Clustering	Double Clustering	Double Clustering	Double Clustering	Double Clustering
N	56889	56889	56889	56889	56889	56889	56889	56889
Adj. R^2	-0.104	-0.126	-0.125	-0.123	-0.090	-0.117	-0.135	-0.105

Table 12 Information Uncertainty and the SEC Enforcement

This table reports the results on the relation between return prediction information uncertainty measures and the SEC enforcement actions. The dependent variables are dummies of value 1 and 0 representing whether a firm is involved in a specified type of the SEC enforcement actions. The prediction precision is defined as the 12-month prediction precision in the lagged fiscal year, while the information incompleteness is defined as the average information incompleteness in the lagged fiscal year. The information incompleteness is defined in Section 5. Appendix Table A5 details the definition of the dependent and the control variables. The dependent variables are not included in the return prediction practice. The *t* statistics are robust statistics with error double clustered using industry and year.

	(1)	(2)	(3)
Variable	10K Comment Letter	SEC Investigation	AAER
Precision	-0.029	-0.003	0.001
	(-1.904)	(-0.378)	(0.465)
Info. Incomp.	0.095	0.067	-0.001
	(2.640)	(3.787)	(-0.095)
Volatility	0.011	-0.006	0.002
	(1.285)	(-1.446)	(2.008)
Market Adj. Alpha	-0.001	0.001	-0.001
	(-0.619)	(1.063)	(-1.990)
Log(Sale)	0.017	0.003	0.001
	(5.577)	(1.661)	(1.379)
Leverage	-0.003	0.002	-0.001
	(-0.243)	(0.348)	(-0.626)
Cash	0.037	-0.014	-0.001
	(2.277)	(-1.681)	(-0.499)
Tobin's Q	0.001	0.001	0.000
	(0.812)	(1.393)	(-0.579)
Discretionary Accrual	-0.001	0.010	-0.002
	(-0.082)	(1.392)	(-0.704)
Log(Firm Age)	-0.044	0.009	0.004
	(-4.820)	(2.037)	(3.006)
Log(Analyst)	0.001	0.006	0.000
	(0.174)	(2.715)	(-0.048)
Fortune 500	0.013	0.004	-0.002
	(1.068)	(0.713)	(-0.923)
Industry FE	Y	Y	Y
Year FE	Y	Y	Y
Error Cluster	Double Clustering	Double Clustering	Double Clustering
N	56889	56889	56889
Adj. R^2	-0.039	-0.122	-0.135

Appendix

Table A1 Modeling Windows

This table reports the specification of the modeling windows. The models are updated every ten years in this paper. The starting date of the training process is January 1962. Every update will train the model using the training data set for in-sample fitting. The fitted models will make predictions for the following validation set and the best combination of architecture and hyperparameters are chosen to make the out-of-sample predictions in the testing periods.

Window	Train Start	Train End	Validation End	Test End
1	01/31/1962	12/31/1977	12/31/1982	12/31/1992
2	01/31/1962	12/31/1987	12/31/1992	12/31/2002
3	01/31/1962	12/31/1997	12/31/2002	12/31/2012
4	01/31/1962	12/31/2007	12/31/2012	12/31/2021

Table A2 Additional Optimization Choices

We conduct a grid search for the best parameters and hyperparameters in training and validation data sets. We train all the sub-models first in the training data set. Then, we select the best performing model in the validation data set for the hyperparameter values. Panel A details the additional optimization choice of our grid search. Panel B reports the selected modeling parameters and hyperparameters after training and hyperparameter tuning.

Model	Parameter	Choice
ANN (ANN Rectifier/Tanh)	Loss Function	Cross entropy for classification/mean squared error for regression
	Learning Rate	Adadelta with rho=0.99 and epsilon=1e-8
	Activation	Rectifier or Tanh for two ANN models separately
	# Epochs	1000
GBT	Loss Function	Cross entropy for classification/mean squared error for regression
	# Trees	1000
	Learning Rate	0.1
RF	Loss Function	Cross entropy for classification/mean squared error for regression
	# Trees	1000

Table A3 Prediction Sample Firm Characteristics Summary Statistics

The follow table reports the summary statistics of the firm characteristics following Green et al. 2017. We construct the sample such that the data is CRSP centric, and we attempt to include as many common share stocks listed on three major exchanges (NYSE, AMEX, and NASDAQ) as possible. However, we do not include other securities such as REITS. Our data construction avoids issues, including high volatility in the number of stocks from month to month. In our models, we normalize these following predictors monthly.

Variable	Mean	Standard Deviation	Min	Median	Max
absacc	0.098	0.114	0.000	0.066	1.086
acc	-0.023	0.142	-1.039	-0.019	0.582
aeavol	0.853	2.051	-1.000	0.290	21.222
age	15.076	12.893	1.000	11.000	71.000
agr	0.283	1.105	-0.693	0.083	35.398
baspread	0.055	0.069	-0.430	0.036	0.985
beta	1.083	0.651	-1.489	1.014	3.910
betasq	1.602	1.810	0.000	1.032	15.291
bm	0.755	0.726	-2.581	0.585	7.894
bm_ia	23.174	691.727	-2360.690	0.021	16500.928
cash	0.170	0.217	-0.143	0.076	0.980
cashdebt	-0.045	1.670	-382.788	0.127	2.851
cashpr	-0.570	55.119	-656.405	-0.510	594.905
cfp	0.019	0.312	-4.130	0.042	7.626
cfp_ia	12.595	303.092	-310.191	0.016	6795.637
chatoia	-0.005	0.243	-1.380	0.003	1.306
chcsho	0.221	1.005	-0.892	0.008	28.089
chempia	-0.101	0.651	-24.055	-0.061	3.647
chfeps	0.003	0.603	-19.140	0.000	20.950
chinv	0.015	0.059	-0.287	0.001	0.426
chmom	-0.001	0.567	-8.455	-0.006	7.783
chnanalyst	0.026	1.571	-42.000	0.000	38.000
chpmia	0.305	7.505	-93.863	-0.004	111.909
chtx	0.001	0.013	-0.121	0.000	0.145
cinvest	-0.027	6.895	-157.600	-0.002	3390.067
convind	1.130	0.336	1.000	1.000	2.000
currat	3.381	5.994	0.102	1.971	105.898
depr	0.269	0.440	-0.984	0.152	8.147
disp	0.171	0.465	0.000	0.044	12.500
divi	2.006	0.263	1.000	2.000	3.000
divo	1.998	0.246	1.000	2.000	3.000
dolvol	11.129	3.048	-3.060	10.982	19.490
dy	0.018	0.035	-6.122	0.001	0.556
ear	0.003	0.083	-0.458	0.001	0.504
egr	0.215	1.942	-38.569	0.082	43.328
ep	-0.026	0.364	-8.012	0.048	0.683
fgr5yr	16.814	11.617	-74.000	14.830	208.830
gma	0.376	0.389	-1.520	0.313	2.977
grcapx	1.270	4.806	-18.500	0.177	67.915
grltnoa	0.096	0.172	-0.917	0.060	1.256
herf	0.067	0.081	0.003	0.043	1.000
hire	0.091	0.339	-0.700	0.008	3.917
idiovol	0.065	0.037	0.000	0.055	0.266
ill	0.000	0.000	0.000	0.000	0.001
indmom	0.142	0.300	-0.757	0.116	3.102

invest	0.100	0.235	-0.562	0.046	2.990
ipo	1.058	0.234	1.000	1.000	2.000
lev	2.191	4.712	0.000	0.668	73.048
lgr	0.309	1.060	-0.792	0.080	15.515
maxret	0.075	0.072	0.000	0.053	0.846
mom12m	0.129	0.595	-0.972	0.051	11.365

Table A3 (Continues)

Variable	Mean	SD	Min	Median	Max
mom1m	0.010	0.155	-0.728	0.000	2.000
mom36m	0.315	0.937	-0.986	0.141	14.514
mom6m	0.054	0.368	-0.911	0.020	7.533
ms	3.609	1.688	0.000	4.000	8.000
mve	11.734	2.252	2.357	11.579	18.588
mve_ia	-189.253	7566.268	-26395.790	-364.757	142031.617
nanalyst	4.884	6.657	0.000	2.000	57.000
nincr	0.945	1.299	0.000	1.000	8.000
operprof	0.831	1.603	-10.005	0.615	18.265
orgcap	0.144	0.485	-0.702	0.015	8.223
pchcapx_ia	3.754	54.529	-890.899	-0.561	939.472
pchcurrat	0.194	1.229	-0.915	-0.004	23.397
pchdepr	0.106	0.565	-0.961	0.023	7.789
pchgm_pchsale	-0.096	1.144	-20.502	-0.002	6.174
pchquick	0.243	1.464	-0.938	-0.002	29.768
pchsale_pchinvt	-0.065	0.862	-10.579	0.013	4.163
pchsale_pchrect	-0.061	0.771	-10.015	-0.001	5.431
pchsale_pchxsga	0.029	0.427	-2.897	-0.001	6.642
pchsaleinv	0.154	1.035	-121.036	0.010	30.974
pctacc	-0.647	5.934	-63.600	-0.258	65.444
pricedelay	0.143	0.999	-16.494	0.062	13.838
ps	4.089	1.762	0.000	4.000	9.000
quick	2.667	5.466	0.061	1.294	98.567
rd	2.077	0.367	1.000	2.000	3.000
rd_mve	0.065	0.112	-0.034	0.028	2.228
rd_sale	0.825	6.751	-218.737	0.031	210.899
realestate	0.266	0.200	0.000	0.231	1.000
retvol	0.033	0.026	0.000	0.026	0.262
roaq	-0.009	0.070	-1.047	0.006	0.219
roavol	0.032	0.069	0.000	0.013	1.238
roeq	-0.007	0.196	-4.833	0.022	2.773
roic	-0.128	1.152	-20.737	0.066	1.266
rsup	-0.048	3.987	-2580.272	0.013	6.239
salecash	50.266	161.272	-1230.906	9.833	2942.250
saleinv	26.255	71.165	-106.622	7.549	1203.586
salerec	11.789	50.632	-21796.000	5.918	276.499
secured	0.571	0.517	0.000	0.585	4.013
securedind	1.387	0.487	1.000	1.000	2.000
sfe	-0.596	7.512	-326.471	0.043	4.062
sgr	0.239	0.789	-0.984	0.100	13.743
sin	1.007	0.085	1.000	1.000	2.000
sp	2.222	3.651	-35.942	1.028	55.651
std_dolvol	0.862	0.410	0.000	0.794	3.332

std_turn	4.587	13.885	0.000	1.914	625.712
stdacc	9.588	60.087	0.000	0.141	1138.612
stdcf	17.605	119.120	0.000	0.156	2723.991
sue	-0.006	0.190	-11.824	0.000	3.305
tang	0.541	0.157	0.000	0.550	0.984
tb	-0.118	1.532	-25.942	-0.072	12.172
turn	1.103	2.197	0.000	0.531	76.062
zerotrade	1.369	3.366	0.000	0.000	20.046

Table A4 Selected Models after Training and Hyperparameter Tuning

This table reports the selected hyperparameters for each combination of models and training and validation period. Column "Classification" reports the parameters for the classification models of the corresponding modeling architecture, while column "Regression" reports the parameter for the regression models of the corresponding modeling architecture.

Model	Training Window		Validation Window		Classification		Regression	
	Train Start	Train End	Validation Start	Validation End	Hidden	l1	Hidden	l1
ANN (Tanh)	01/31/1962	12/31/1977	01/01/1978	12/31/1982	16	0	(64, 32, 16)	0
	01/31/1962	12/31/1987	01/01/1988	12/31/1992	16	0	(64, 32, 16, 8)	0
	01/31/1962	12/31/1997	01/01/1998	12/31/2002	(32, 16)	0	(16, 8)	0
	01/31/1962	12/31/2007	01/01/2008	12/31/2012	16	0	8	0
	Train Start	Train End	Validation Start	Validation End	Hidden	l1	Hidden	l1
ANN (Rectifier)	01/31/1962	12/31/1977	01/01/1978	12/31/1982	(128, 64, 32, 16, 8)	0	(32, 16)	0
	01/31/1962	12/31/1987	01/01/1988	12/31/1992	(128, 64, 32)	0	8	0
	01/31/1962	12/31/1997	01/01/1998	12/31/2002	(128, 64, 32)	0	(128, 64, 32)	0
	01/31/1962	12/31/2007	01/01/2008	12/31/2012	(128, 64, 32)	0	(64, 32, 16, 8)	0
	Train Start	Train End	Validation Start	Validation End	Max Depth		Max Depth	
GBT	01/31/1962	12/31/1977	01/01/1978	12/31/1982	2		4	
	01/31/1962	12/31/1987	01/01/1988	12/31/1992	4		4	
	01/31/1962	12/31/1997	01/01/1998	12/31/2002	4		2	
	01/31/1962	12/31/2007	01/01/2008	12/31/2012	4		4	
	Train Start	Train End	Validation Start	Validation End	Max Depth		Max Depth	
RF	01/31/1962	12/31/1977	01/01/1978	12/31/1982	8		8	
	01/31/1962	12/31/1987	01/01/1988	12/31/1992	8		8	
	01/31/1962	12/31/1997	01/01/1998	12/31/2002	8		8	
	01/31/1962	12/31/2007	01/01/2008	12/31/2012	8		8	

Table A5 Variable Definition for the Information Environment Tests

The table below first describes the main variable definitions we use in this paper, including their possible values and calculations in Panel A and then details the calculation of discretionary accrual in this paper using the Compustat database following the Jones model as modified by Dechow et al. (1995) in Panel B.

Panel A: Information Environment Test Variable Definition		
Variables	Description	Variable Value and Calculation
10K Comment Letter	This variable is the SEC comment letter record from Audit Analytics. The SEC publicizes the comment letter records since 2005.	1 if the firm-year is associated with a comment letter related to 10K form, 0 otherwise.
AAER	This variable is the SEC AAER obtained from University of South California (see Dechow et al. 2011).	1 if the firm-year is in the SEC's accounting and auditing enforcement records, 0 otherwise.
Annual Volatility	This variable is the stock return volatility aggregated to the fiscal year calculated with CRSP database.	Aggregated 12-month return volatility at the fiscal year level.
Cash	This variable is the cash holding amount estimated with Compustat database.	Ratio between cash holding (Compustat item che) and total assets (Compustat item at).
Discretionary Accrual	This variable is the discretionary accruals estimated using the modified Jones method (see Dechow et al. 1995).	The calculation is detailed in Table A4 Panel B.
Fortune 500	This variable is a dummy of whether a firm is a Fortune 500 firm from the Compustat database.	1 if the firm is a Fortune 500 company, 0 otherwise.
Restatement	This variable is a dummy of restatements with negative changes on earnings from Audit Analytics.	1 if the fiscal period is associated with an income-reducing restatement, 0 otherwise.
Leverage	This variable is the leverage level estimated with Compustat database.	Ratio of long-term debt (Compustat item ltd) to total assets (Compustat item at).
Litigation	This variable is a dummy variable summarizing the specified type of litigation from Audit Analytics database.	1 if the firm-year is associated with the specified type of litigation, which can be shareholder, environmental, civil rights, regulatory, labor, intellectual property, illegal activities, or all of these types together, 0 otherwise.
Ln (Analyst)	This variable is the number of distinct analysts who follow the firm from the I/B/E/S database.	Log of the number of equity research analysts covering the firm.
Ln (Firm Age)	This variable is the number of years since a firm appeared in the Compustat database.	Log of firm age in years.
Size	This variable is the sales volume as a size control variable from Compustat database.	Log of sales (Compustat item sale)

Table A5 (Continued)

Variables	Description	Variable Value and Calculation
Market Adjusted Return	This variable is the market-adjusted return for the fiscal year.	Difference between fiscal period holding return and the CRSP value-weight market holding return calculated using CRSP database.
Power	This variable is the CEO's pay slice out of the top five executives (see Bebchuk et al. 2011).	CEO total compensation divided by the total compensation of the rest of the top five executives.
SEC Investigation	This variable is the SEC undisclosed investigation record obtained through Freedom of Information Act (FOIA) request.	1 if the firm-year is under the SEC's undisclosed investigation, 0 otherwise.
Tobin's Q	This variable is Tobin's Q, estimated with the Compustat database.	Tobin's Q from the last fiscal year is calculated as $Q = \frac{Total\ Assets + Market\ Equity - Book\ Equity}{Total\ Assets}$, where book equity is the shareholders' equity (Compustat item seq) adjusted with deferred taxes (Compustat item txdb) and preferred shares, and market equity is the fiscal year end stock price (Compustat item prcc_f) multiplied by outstanding common shares (Compustat item csho).

Table A5 (Continues)

Panel B: Discretionary Accrual Calculation	
Step 1 Data Preparation	The discretionary accrual is calculated following the modified Jones model. To obtain robust estimation of the average two-digit SIC code–level discretionary accrual, three types of firms are excluded from the estimation process: 1. Firms with total assets and lag 1 total asset smaller than 1 million USD, 2. Firms associated with a 2-digit SIC code of less than ten fiscal year observations in the Compustat database, and 3. Firms with missing values in total assets, lag 1 total assets, sales, lag 1 sales, income, operating net cash flow, and plant, property, and equipment.
Step 2 Raw Input Preparation	$\%Total\ Accrual_t = \frac{Total\ Income_t - Operating\ Net\ Cash\ Flow_t}{Total\ Assets_{t-1}}$ $Revenue\ Growth_t = Sales_t - Sales_{t-1}$ $Receivable\ Change_t = Total\ Receivable_t - Total\ Receivable_{t-1}$ $\%Sales_t = \frac{Revenue\ Growth_t - Receivable\ Change_t}{Total\ Assets_{t-1}}$ $\%Property\ Plant\ and\ Equipment_t = \frac{Property\ Plant\ and\ Equipment_t}{Total\ Asset_{t-1}}$
Step 3 Missing Value Substitution in %Total Accruals	If total accrual calculation in step 2 is not viable, $\%Total\ Accrual_t = [(Total\ Current\ Assets_t - Total\ Current\ Assets_{t-1}) - (Total\ Current\ Liability_t - Total\ Current\ Liability_{t-1}) + (Debt\ in\ Current\ Liability_t - Debt\ in\ Current\ Liability_{t-1}) - Depreciation_t]/Total\ Assets_t$
Step 4 Winsorization	To ensure the robustness of the industry-year benchmark, the input variables are Winsorized in fiscal-year groups at the 1% and 99% levels.
Step 5 Calculation	For each two-digit SIC code and fiscal year combination, we conduct the following regression: $\%Total\ Accrual_{it} = \beta_0 + \beta_1 \%Sales_{it} + \beta_2 \frac{1}{Total\ Assets_{it-1}} + \beta_3 \%Property\ Plant\ and\ Equipment_t + Discretionary\ Accrual_{it}$, where the residual term $Discretionary\ Accrual_{it}$ is taken as the discretionary accrual for firm *i* in fiscal year *t*. In the empirical analysis, the absolute value is adopted to focus the analysis only on the magnitude of the discretionary accrual without consideration of the direction.

Table A6 Average Precision and Average Information Incompleteness by Industry

This table reports the industry averages of prediction precision and information incompleteness across the out-of-sample period computed with all common stocks in the three major exchanges (NYSE, AMEX, and NASDAQ). Panel A reports the averages for the prediction precision, while Panel B reports averages for the information incompleteness. The prediction precision is defined as the ratio between number of successful predictions and the total number of predictions. The information incompleteness is defined based on the aggregated predicted decile probabilities $E_{i,t} = -\sum_{d_{i,t} \in D} p_{agg}\widehat{(d_{\iota,t})} \log_2 p_{agg}\widehat{(d_{\iota,t})}$. The information incompleteness measures the minimum number of binary questions that need to be answered to completely eliminate the return prediction uncertainty. In other words, it measures the shortage of information.

Panel A: Industry Level Prediction Precision

2-digit SIC Industry	Code	Precision	2-digit SIC Industry	Code	Precision
Forestry	8	0.263	Apparel & Other Textile Products	23	0.154
Membership Organizations	86	0.241	Real Estate	65	0.153
Services, Not Elsewhere Classified	89	0.192	Rubber & Miscellaneous Plastics Products	30	0.152
Metal, Mining	10	0.186	Wholesale Trade – Nondurable Goods	51	0.152
Motion Pictures	78	0.184	Educational Services	82	0.151
Agricultural Production – Livestock	2	0.184	Fabricated Metal Products	34	0.151
Non-Classifiable Establishments	99	0.183	Insurance Carriers	63	0.149
Chemical & Allied Products	28	0.179	Heavy Construction, Except Building	16	0.149
Legal Services	81	0.178	Personal Services	72	0.148
Coal Mining	12	0.176	General Building Contractors	15	0.148
Oil & Gas Extraction	13	0.173	Security & Commodity Brokers	62	0.147
Business Services	73	0.172	Transportation Equipment	37	0.147
Local & Interurban Passenger Transit	41	0.169	Eating & Drinking Places	58	0.147
Holding & Other Investment Offices	67	0.169	Transportation Services	47	0.147
Instruments & Related Products	38	0.166	Auto Repair, Services, & Parking	75	0.147
Electronic & Other Electric Equipment	36	0.166	Apparel & Accessory Stores	56	0.147
Water Transportation	44	0.166	Furniture & Fixtures	25	0.146
Nonmetallic Minerals, Except Fuels	14	0.166	Building Materials & Gardening Supplies	52	0.146
Electric, Gas, & Sanitary Services	49	0.166	Lumber & Wood Products	24	0.146
Communications	48	0.164	Food & Kindred Products	20	0.146
Health Services	80	0.163	General Merchandise Stores	53	0.145
Special Trade Contractors	17	0.163	Paper & Allied Products	26	0.144
Miscellaneous Manufacturing Industries	39	0.163	Tobacco Products	21	0.144
Engineering & Management Services	87	0.162	Petroleum & Coal Products	29	0.143
Industrial Machinery & Equipment	35	0.161	Stone, Clay, & Glass Products	32	0.143
Amusement & Recreation Services	79	0.161	Transportation by Air	45	0.142
Insurance Agents, Brokers, & Service	64	0.160	Primary Metal Industries	33	0.142
Furniture & Home furnishings Stores	57	0.158	Railroad Transportation	40	0.141
Depository Institutions	60	0.157	Social Services	83	0.140
Nondepository Institutions	61	0.157	Hotels & Other Lodging Places	70	0.140
Miscellaneous Retail	59	0.157	Trucking & Warehousing	42	0.139
Wholesale Trade – Durable Goods	50	0.157	Food Stores	54	0.139
Pipelines, Except Natural Gas	46	0.156	Textile Mill Products	22	0.139
Printing & Publishing	27	0.155	Automotive Dealers & Service Stations	55	0.136
Agricultural Production – Crops	1	0.154	Leather & Leather Products	31	0.136
Agricultural Services	7	0.154	Miscellaneous Repair Services	76	0.133

Table A6 (Continues)

Panel B: Industry Level Prediction Uncertainty					
2-digit SIC Industry	Code	Info. Incomp.	2-digit SIC Industry	Code	Info. Incomp.
Electric, Gas, & Sanitary Services	49	3.190	Miscellaneous Manufacturing Industries	39	3.247
Depository Institutions	60	3.196	Nondepository Institutions	61	3.247
Tobacco Products	21	3.197	Health Services	80	3.248
Non-Classifiable Establishments	99	3.198	Furniture & Fixtures	25	3.248
Chemical & Allied Products	28	3.204	Industrial Machinery & Equipment	35	3.248
Pipelines, Except Natural Gas	46	3.211	Wholesale Trade – Nondurable Goods	51	3.248
Railroad Transportation	40	3.212	Stone, Clay, & Glass Products	32	3.249
Forestry	8	3.213	Real Estate	65	3.249
Membership Organizations	86	3.218	General Merchandise Stores	53	3.250
Metal, Mining	10	3.221	Auto Repair, Services, & Parking	75	3.250
Insurance Carriers	63	3.223	Apparel & Other Textile Products	23	3.250
Petroleum & Coal Products	29	3.226	Rubber & Miscellaneous Plastics Products	30	3.250
Motion Pictures	78	3.231	Special Trade Contractors	17	3.251
Business Services	73	3.231	Wholesale Trade – Durable Goods	50	3.251
Nonmetallic Minerals, Except Fuels	14	3.232	Hotels & Other Lodging Places	70	3.251
Insurance Agents, Brokers, & Service	64	3.232	Services, Not Elsewhere Classified	89	3.252
Holding & Other Investment Offices	67	3.232	Miscellaneous Retail	59	3.252
Communications	48	3.233	Agricultural Services	7	3.252
Paper & Allied Products	26	3.233	Local & Interurban Passenger Transit	41	3.253
Food & Kindred Products	20	3.233	Educational Services	82	3.254
Printing & Publishing	27	3.234	Eating & Drinking Places	58	3.255
Oil & Gas Extraction	13	3.235	Building Materials & Gardening Supplies	52	3.257
Water Transportation	44	3.237	Lumber & Wood Products	24	3.257
Engineering & Management Services	87	3.237	Trucking & Warehousing	42	3.257
Instruments & Related Products	38	3.238	Social Services	83	3.260
Personal Services	72	3.240	General Building Contractors	15	3.261
Security & Commodity Brokers	62	3.241	Legal Services	81	3.262
Electronic & Other Electric Equipment	36	3.242	Leather & Leather Products	31	3.262
Amusement & Recreation Services	79	3.244	Heavy Construction, Except Building	16	3.263
Fabricated Metal Products	34	3.244	Primary Metal Industries	33	3.263
Coal Mining	12	3.244	Automotive Dealers & Service Stations	55	3.263
Agricultural Production – Crops	1	3.245	Apparel & Accessory Stores	56	3.263
Food Stores	54	3.245	Furniture & Home furnishings Stores	57	3.264
Agricultural Production – Livestock	2	3.245	Transportation by Air	45	3.266
Transportation Equipment	37	3.245	Textile Mill Products	22	3.268
Transportation Services	47	3.246	Miscellaneous Repair Services	76	3.285

Table A7 Performance of Portfolios Including Stocks with Top 50% Market Capitalization

This table reports the economic performance of the portfolios *using only the stocks with above median capitalization* constructed based on the aggregated predictions from the individual classifiers. The statistics are calculated based on the out-of-sample period covering 198301:202112. The decile portfolios are sorted based on the predicted deciles monthly, which are the deciles with the highest predicted probabilities. The column "market" reports the performance of the buy-and-hold strategy using all common stocks in the three major exchanges. The cumulative returns are in decimal unit representing gross returns in the sample period. $\alpha's$ are for the corresponding factor models, e.g., CAPM or Fama-French 3 Factor model. The t statistics for the $\alpha's$ are Newey-West t statistics of lag 6. The performance statistics are based on excess return adjusted with risk-free rate, i.e., 30-day US treasury bill. We report annualized Sharpe ratios. Turnover is defined as the average total percentage of holding changes in absolute value. Max drawdown is defined as the max difference between current price and the most recent price peak in percentage across all months in our sample period. Panel A reports the equal-weight portfolio performance, while Panel B reports the value-weight portfolio performance. A robustness check of the portfolio performance using only the stocks above the median market capitalization of the market is included in the Appendix Table A7.

Panel A: Equal-Weight Decile Portfolios

Statistic	Market	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Mean Excess Return	0.009	-0.002	0.003	0.004	0.004	0.004	0.008	0.010	0.012	0.015	0.014	0.013
Cumulative Return	24.199	-0.962	0.129	1.628	2.438	3.488	27.891	77.556	135.289	310.412	120.312	341.751
CAPM Alpha	0.000	-0.016	-0.008	-0.005	-0.004	-0.002	0.002	0.003	0.004	0.004	0.002	0.015
	0.049	-5.166	-4.224	-2.636	-2.137	-1.277	1.270	2.360	1.975	2.296	0.721	5.920
FF3F Alpha	0.000	-0.013	-0.007	-0.005	-0.005	-0.003	0.001	0.003	0.003	0.005	0.003	0.014
	0.340	-7.339	-6.773	-3.821	-4.378	-2.668	1.172	4.022	3.492	4.662	2.031	7.075
FF5F Alpha	0.002	-0.007	-0.006	-0.005	-0.006	-0.005	-0.001	0.001	0.001	0.005	0.006	0.010
	1.513	-4.781	-5.426	-3.625	-4.683	-3.618	-1.114	2.063	1.959	4.758	3.551	5.727
Standard Deviation	0.058	0.098	0.073	0.062	0.057	0.043	0.040	0.045	0.055	0.067	0.085	0.036
Sharpe Ratio	0.515	-0.073	0.143	0.227	0.263	0.334	0.694	0.807	0.766	0.758	0.569	1.267
Turnover	0.105	0.134	0.100	0.082	0.071	0.059	0.052	0.058	0.073	0.092	0.120	0.127
Max Drawdown	-0.607	-0.936	-0.649	-0.637	-0.667	-0.667	-0.450	-0.480	-0.541	-0.558	-0.659	-0.430
Mean N	5342	283	252	66	99	126	692	407	387	202	158	441

Table A7 (Continues)

Panel B: Value-Weight Decile Portfolios												
Statistic	Market	lo	2	3	4	5	6	7	8	9	hi	hi-lo
Mean Excess Return	0.008	0.000	0.003	0.007	0.005	0.005	0.007	0.009	0.009	0.015	0.013	0.011
Cumulative Return	19.733	-0.934	-0.019	6.261	3.084	5.432	17.497	36.998	39.210	236.704	66.740	91.874
CAPM Alpha	0.000	-0.015	-0.009	-0.003	-0.004	-0.002	0.001	0.002	0.001	0.004	0.001	0.012
	-1.674	-4.644	-4.058	-1.659	-1.785	-0.931	0.576	2.753	0.660	1.707	0.183	4.795
FF3F Alpha	0.000	-0.012	-0.007	-0.003	-0.005	-0.003	0.000	0.001	0.001	0.005	0.003	0.012
	-1.738	-5.239	-4.304	-1.488	-2.939	-2.134	-0.304	2.336	0.944	2.931	1.102	5.328
FF5F Alpha	0.000	-0.005	-0.004	-0.001	-0.006	-0.004	-0.002	0.001	0.001	0.006	0.007	0.009
	-1.003	-2.612	-2.745	-0.710	-3.643	-3.208	-3.622	1.070	0.646	3.888	2.658	3.624
Standard Deviation	0.045	0.102	0.079	0.072	0.060	0.048	0.041	0.044	0.054	0.074	0.093	0.050
Sharpe Ratio	0.583	-0.015	0.140	0.333	0.281	0.374	0.600	0.689	0.602	0.681	0.498	0.753
Turnover	0.057	0.125	0.095	0.068	0.064	0.050	0.047	0.048	0.064	0.087	0.113	0.119
Max Drawdown	-0.527	-0.960	-0.829	-0.753	-0.721	-0.637	-0.502	-0.512	-0.617	-0.559	-0.723	-0.510
Mean N	5342	283	252	66	99	126	692	407	387	202	158	441