Measuring Firm Quality Using Machine Learning

Changyi Chen,[1] Bin Ke,[2] Qi Zhao[3]

January 22, 2023

ABSTRACT

Firm quality is a foundational construct in the fundamental analysis literature. Asness et al. (2019), a recent representative example of this literature, measures firm quality based on 19 fundamental signals guided by valuation theory (referred to as Asness' Q score). We examine whether it is possible to leverage the power of machine learning to construct a better measure of firm quality using the same 19 fundamental signals. We show that an advanced machine learning model called XGBoost based on the 19 signals can outperform a linear OLS regression model based on Asness' Q score (our benchmark) by 27%. However, we fail to find economically significant evidence that adding more raw accounting data items identified by the prior literature or commercial databases to XGBoost can generate a stronger prediction model. We show that our measure of firm quality based on XGBoost and the 19 signals can better explain contemporaneous stock prices than Asness' Q score. In addition, a value investing trading strategy using our XGBoost model outperforms the same trading strategy based on Asness' Q score by an economically significant margin.

Keywords: Quality; Fundamental analysis; Machine learning; Value investing

JEL Codes: C53; G12; M41

We thank workshop participants at National of Singapore brownbag seminar, Dongbei University of Finance and Economics, the 5th Annual China Finance and Accounting Academic Conference, 2nd Annual Intelligent Accounting Alliance Conference, School of Engineering at Westlake University, Nanjing Audit University, for helpful comments. Qi Zhao thanks the funding provided by National Natural Science Foundation of China (No. 71771091), China Scholarship Council (No. 202006150130) and Guangdong Basic and Applied Basic Research Fund (No. 2019A1515011752).

[1]Department of Accounting, NUS Business School, National University of Singapore. Email: changyic@u.nus.edu.
[2] Department of Accounting, NUS Business School, National University of Singapore, and NUS (Chongqing) Research Institute. Email: bizk@nus.edu.sg.
[3] Department of Decision Science, School of Business Administration, South China University of Technology. Email: zhaoqiscut@foxmail.com.

1. Introduction

Value investing consists of two key elements: (1) finding quality companies and (2) buying them at reasonable prices, commonly referred to as "cheapness" (Lee 2014). A stock's cheapness is relatively easy to measure because it is typically defined using a stock's current price relative to its existing capital-in-place. However, firm quality is much harder to measure because it is a forward-looking concept (Lee 2014).

Developing an accurate measurement of firm quality is important not only to value investors but also to many other stakeholders. For example, prospective employees often prefer to work for high quality firms because such firms have stronger future prospects. Corporate boards may require a forward-looking measure of firm quality in order to better evaluate and monitor management's current decisions. Financial regulators may also need a reliable measure of quality to gauge a firm's health as part of their regulatory interventions.

There has been a long stream of literature in accounting and finance (commonly referred to as fundamental analysis) that is devoted to developing observable signals from publicly available financial statements that are suggestive of firm quality (see Lee and So 2014 and Hou et al. 2022 for reviews). Many empirical studies develop proxies for firm quality in an ad hoc manner. Asness et al. (2019) is one of the most recent studies in this literature. They develop a unified theoretical framework of firm quality that can be expressed as a function of four additive components: profitability, growth, safety and payout. Asness et al. develop a comprehensive list of 19 standardized proxies for the four components based on publicly available financial statement data. They compute the value of each component based on the average of the standardized proxies and then construct an index of firm quality based on the average of the four standardized components (referred to as Asness' Q score or Q score thereafter). Asness et al. (2019) show that high-quality stocks based on the Q score not only have higher contemporaneous stock prices but also predict higher future risk-adjusted returns.

However, like many prior empirical studies, Asness et al. (2019) use a heuristic approach to construct their Q score. Recent research in accounting and finance has demonstrated the power of machine learning in many prediction tasks (e.g., Bao et al. 2020; Ding et al. 2020; Gu et al. 2020; Bertomeu et al. 2021; Chen et al. 2022). As firm quality is a forward-looking concept that requires forecasting, machine learning could be more appropriate for the measurement of firm quality. The objective of this study is to explore whether we can leverage the power of machine learning to construct a more accurate measure of firm quality. Compared with human heuristics, machine learning can accommodate more flexible functions in the mapping from model inputs to model output. In addition, many machine learning methods can readily handle missing values of model inputs, which is a common challenge for fundamental analysis based on financial statement data (Bao et al. 2020; Chen et al. 2022).

Consistent with Lee (2014) and Asness et al. (2019), we define true firm quality (i.e., the ground truth) as the sum of the present value of a firm's future *realized* residual earnings. We use the one-year expansion of the residual income valuation model to measure true firm quality, though inferences are qualitatively similar if we use a two-year or three-year expansion of the residual income valuation model. Following Asness et al. (2019), we scale true firm quality by the current book value of common equity to make our measure of true firm quality more stationary over time and in the cross section. As true firm quality, the dependent variable, is continuous, we use mean absolute error (MAE) as our primary performance evaluation criterion, which is less sensitive to the influence of outliers. However, all of our inferences are qualitatively similar if we use the mean square error (MSE) as the performance evaluation criterion.

Our sample covers all publicly listed U.S. firms that satisfy our sample selection criteria for the period 1973-2018. We use a three-year rolling window for model training and hence the test sample covers the years 1976-2018. As publicly listed U.S. firms follow different fiscal

year ends, we follow Hou et al. (2012) and Li and Mohanram (2014) by training the machine learning models using only publicly available data as of June 30 for each year.

Since Asness et al. (2019) constructed their Q score without referring to the ground truth, we develop a baseline benchmark model that uses Asness' Q score as the only model input and the linear OLS regression as the machine learning model (referred to as LR-Q).

First, we show that it pays to use disaggregated fundamental signals (i.e., the 19 proxies) rather than Asness' Q score (i.e., the aggregation of the 19 proxies based on human heuristics) to predict true firm quality. Compared with LR-Q, we find that a firm quality prediction model based on the 19 proxies and OLS regression (referred to as LR-19) significantly outperforms LR-Q. For all test years, the mean MAE of LR-19 drops by 16% relative to the mean MAE of LR-Q.

Second, we show that it pays to use an advanced machine learning method (i.e., XGBoost, one of the state-of-art machine learning methods), to construct firm quality. We build two prediction models using XGBoost: (i) use Asness' Q score as the model input (referred to as XGBoost-Q); and (ii) use the 19 proxies as model inputs (referred to as XGBoost-19). Using the Q score as the only model input, we find that the mean MAE of XGBoost-Q drops by 15% relative to the mean MAE of LR-Q. Similarly, using the 19 proxies as model inputs, we show that the mean MAE of XGBoost-19 drops by 14% relative to the mean MAE of LR-19. These results demonstrate the power of using state-of-art machine learning in constructing firm quality measures.

Combining the benefits of using both more disaggregated model inputs and a more powerful machine learning model, we find that the mean MAE of XGBoost-19 drops by 27% relative to the mean MAE of LR-Q (the baseline model). The performance of our best model XGBoost-19 is also highly stable over time: XGBoost-19 always outperforms all the other models for each of the test years.

When constructing firm quality, one empirical challenge is the presence of missing values for many raw accounting data items used to construct the 19 proxies. To maintain a relatively large sample, Asness et al. (2019) drop an individual proxy in the construction of the Q score if the individual proxy contains missing values. As the OLS regression model requires all input variables to be non-missing, we filled the missing values of raw accounting data items based on accounting knowledge and then discarded all remaining firm years with missing values on any of the 19 proxies so that we can maintain a common sample for all models discussed above. To make sure that the superior performance of XGBoost-19 relative to LR-Q is not driven by this special missing value treatment, we perform two robustness checks. First, we examine whether our manual filling of missing raw accounting data items affects our inferences. Specifically, we rerun the XGBoost-19 model by retaining the firm-years with missing values for any of the 19 proxies (referred to as XGBoost-19m). We add an "m" at the end of a model name to denote the fact that the model can accommodate missing values. The sample size for XGBoost-19m is identical to that for the aforementioned models including XGBoost-19. We find that the mean MAE is not significantly different for XGBoost-19 versus XGBoost-19m. This finding suggests that the missing value treatment is not a driver of XGBoost-19's superior performance.

Second, we examine whether the superior performance of XGBoost-19m relative to LR-Q is driven by the sample size reduction resulting from the non-filled missing values. Compared with Asness' Q score, our final sample used in the above analyses loses approximately 27.7% of the full sample before the missing value treatment. To make sure that the superior performance of XGBoost-19m relative to LR-Q is not due to the difference in sample size resulting from missing values, we rerun the LR-Q model and the XGBoost-19m model by using full sample of firm-years that contains all observations with missing values for any of the 19 proxies (referred to as LR-Q_Full and XGBoost-19m_Full, respectively). We

find that the mean MAE is still significantly larger for LR-Q_Full than for XGBoost-19m_Full. This finding suggests that the sample size reduction resulting from missing values is not a driver of XGBoost-19m's superior performance relative to LR-Q.

Even though Asness et al. (2019) define firm quality based on valuation theory, the 19 standardized proxies they use to construct firm quality are selected in an ad hoc manner due to a lack of theoretical guidance. As they utilize only a subset of publicly available financial statement raw data items, they could have missed important financial statement variables that are useful for the measurement of firm quality. To investigate this possibility, we perform three inductive analyses advocated by Karpoff and Wittry (2018). The first inductive analysis uses the 63 individual raw accounting data items that are used to construct the 19 individual firm quality proxies. As the aforementioned analyses suggest that disaggregated financial data could be more useful than aggregated ones for the purpose of predicting firm quality, we examine whether we can develop a more accurate XGBoost prediction model using the 63 individual raw accounting data items as model inputs (referred to as XGBoost-63m). We do not delete observations with missing values for any of the 63 raw accounting data items because XGBoost allows missing values. The mean MAE of XGBoost-63m drops by 1.8% relative to the mean MAE of XGBoost-19m, suggesting limited benefit of further disaggregating the 19 proxies used by Asness et al. (2019).

Second, we include an additional 24 raw accounting data items used by two recent fundamental analysis studies (Li and Mohanram 2019 and Bartram and Grinblatt 2018) but omitted by Asness et al. (2019). We examine whether XGBoost based on the 87 (63+24) raw accounting data items (referred to as XGBoost-87m) can outperform XGBoost-63m. The mean MAE of XGBoost-87m drops by 1% relative to the mean MAE of XGBoost-63m, suggesting weak evidence of further performance improvement for XGBoost-87m versus XGBoost-63m.

Our final inductive analysis examines whether it is possible to build a more powerful XGBoost model by deploying all readily available 318 raw accounting data items from Compustat (referred to XGBoost-318m). We find no statistically significant evidence that XGBoost-318m can outperform XGBoost-87m, suggesting that a brutal data mining approach devoid of any theoretical guidance does not yield a more powerful firm quality prediction model.

As noted above, reliable measures of firm quality are useful to many different stakeholders. We illustrate the usefulness of our machine learning models in the context of value investing. To benchmark with Asness' Q score, which is constructed based on 19 financial variables, we focus on XGBoost-19m for the following discussions. We conduct two complementary analyses. First, we show that firm quality based on XGBoost-19m can better explain contemporaneous stock prices than Asness' Q score. This finding suggests that our firm quality measure is closer to the firm quality measure used by stock market investors.

Second, we examine whether it is possible to use XGBoost-19m based firm quality measure to construct a value investing trading strategy that can generate abnormal returns (alpha) over a 12-month investment horizon and whether our measure can outperform Asness' Q score. We have no clear predictions on these questions because all of our firm quality measures are based on publicly available accounting data and thus, if the stock market is semi-strong efficient, we should not be able to obtain abnormal stock returns using either Asness' Q score or any of our machine learning models.

To make sure that we have adequately controlled for all known pricing factors, we use the most comprehensive asset pricing factor model available in the literature, i.e., the q^5 factor model developed by Hou et al. (2021). Hou et al. (2021) show that many documented stock pricing anomalies vanish once abnormal stock return is defined using the q^5 factor model.

Following Asness et al. (2019), we use the standard calendar time portfolio approach to assess the statistical significance of abnormal stock returns by forming equal weighted calendar time hedging portfolios that take a long position in the stocks in the top decile of a relevant sorting variable and take a short position in the stocks in the bottom decile of a relevant sorting variable. Asness et al. (2019) form their hedging portfolio using the level of their Q score. We use change rather than level of firm quality in the construction of our hedging portfolios because stock markets react to new information, which is better measured by change rather than level of firm quality (Ball and Brown 1968; Chen et al. 2022). Value investing requires purchasing high quality stocks at reasonable prices (Lee 2014). Hence, we form our hedging portfolios based on the double sorting of the following two variables: first sort all stocks into 10 deciles based on change in firm quality; second, for all stocks in each decile, we sort them into 10 deciles based on the book-to-market ratio.[1] The double sorting value investing strategy based on Asness' Q score yields a statistically significant monthly alpha of 0.84%. In contrast, the double-sort value investing strategy based on XGBoost-19m yields a much larger monthly alpha of 2.18%, which is almost 3 times as large as the monthly alpha for the trading strategy based on Asness' Q. This finding offers further support that machine learning based on disaggregated input data (i.e., XGBoost-19m) is more effective than heuristics in constructing a firm quality proxy.

We also examine whether the double-sort value investing trading strategy based on XGBoost-63m, XGBoost-87m and XGBoost-318m can outperform the trading strategy based on XGBoost-19m. Even though XGBoost-63m, XGBoost-87m and XGBoost-318m all statistically outperform XGBoost-19m in predicting firm quality, we find no evidence that the

[1] Our double sort trading strategy does not yield a significant q^5 factor model alpha if we use the level of firm quality as a sorting variable (untabulated).

double-sort trading strategy based on any of these three XGBoost models generates a larger alpha than the same trading strategy based on XGBoost-19m.

As our test sample spans a fairly long time period over 1976-2018, we also break it into 3 equal time periods for the double-sorting hedging portfolio analysis. We find that the double sort trading strategy's alpha for XGBoost-19m is statistically and economically significant for all three time periods, suggesting that our machine learning based firm quality measure provides benefits to value investing even in the most recent time period.

We make two important contributions to the existing literature. First, we contribute to the literature on measurement of firm quality. A large accounting and finance literature is devoted to developing observable signals of firm quality based on human heuristics (e.g., Piotroski 2000; Mohanram 2005; Piotroski and So 2012; Asness et al. 2019; Li and Mohanram 2019). We contribute to this literature by being one of the first studies to use machine learning to develop a more accurate measure of firm quality. Two concurrent studies by Binz et al. (2022) and Cao and You (2021) also apply machine learning to fundamental analysis. Our study differs from these two studies in two important ways. First, the dependent variables are different. Binz et al. (2022) use return on common equity (ROCE) and return on net operating assets (RNOA) and Cao and You (2021) focus on earnings. In contrast, we aim to predict firm quality, a concept that is closely related to but still distinct from earnings. Second and more importantly, we use different input variables (i.e., predictors). Even though all three studies construct their input variables based on financial statement data items, they adopt different approaches. As noted by Binz et al. (2022), Cao and You (2021) identify a comprehensive list of 60 financial statement raw data items (both levels and changes) without a framework. Binz et al. (2022) select their input variables based on Nissim and Penman's (2001) hierarchical approach to financial statement analysis (i.e., the DuPont decomposition), which implicitly assumes that lagged financial statement data items from the DuPont decomposition are useful in predicting

future firm performance. In contrast, Asness et al.'s (2019) approach used in our study is different in that the 19 firm quality predictors are selected based on a rigorous value framework, which is distinct from Nissim and Penman (2001). We show in the inductive analysis that adding additional financial statement raw data items beyond the 19 proxies used by Asness et al. does not materially improve the prediction performance of firm quality, suggesting that Asness et al.'s approach to selecting fundamental signals for measuring firm quality is relatively complete.

Second, we contribute to the asset pricing anomaly literature based on fundamental analysis. Hou et al. (2022) show that many previously documented pricing anomalies based on fundamental analysis disappears using the q^5 factor pricing model, including Asness et al.'s (2019) trading strategy based on level of firm quality. We show that a trading strategy based on change of our firm quality measure still yields an economically significant q^5 factor-adjusted alpha that is much higher than the alpha based on the change of Asness et al.'s Q score. This finding suggests that the stock market has not fully incorporated the value of fundamental analysis which can be extracted more effectively using machine learning.

The remainder of the paper is structured as follows. The next section discusses the research design, including definition and measurement of firm quality, an introduction to XGBoost, model estimation procedures, and sample selection procedures. Section 3 presents the prediction performance of different models. Section 4 shows the ability of various firm quality measures in explaining contemporaneous stock prices. Section 5 shows the abnormal returns from the trading strategy based on different firm quality proxies. Section 6 concludes.

2. Research Design

2.1. Firm quality: definition and measurement

Lee (2014) define firm quality based on the residual income valuation framework (Feltham and Ohlson 1995) below:

$$
\begin{aligned}
V_t^i = \; & B_t^i + \sum_{j=1}^{\infty} \frac{E_t[\mathrm{NI}_{t+j}^i - r_e^i B_{t+j-1}^i)]}{(1+r_e^i)^j} \\
= \; & B_t^i + \sum_{j=1}^{\infty} \frac{E_t[(\mathrm{ROE}_{t+j}^i - r_e^i) B_{t+j-1}^i]}{(1+r_e^i)^j}
\end{aligned} \tag{1}
$$

Where V_t^i is the present value of residual income of firm *i* at time *t*; B_t^i is the book value of common equity for firm *i* at time *t*; $E_t[\cdot]$ is the expectation based on information available at time *t*; NI_{t+j}^i is the net income of firm *i* for period *t*+*j*; ROE_{t+j}^i is the after-tax return on book equity of firm *i* for period *t*+*j*; r_e^i is the cost of equity capital of firm *i*. The residual income of firm *i* for period *t* in this formula, RI_t^i, is defined as $\mathrm{NI}_t^i - r_e^i B_{t-1}^i$, i.e., the period *t* earnings minus a normal rate-on-return (i.e., r_e^i) on the beginning capital base.

Equation (1) expresses a firm's fundamental value as the sum of two components: invested capital in place B_t^i and present value of expected future residual income PVRI_t^i (i.e., the second term of Equation (1)). Lee (2014) refers to PVRI_t^i as the stock market's measure of firm quality. To make PVRI_t^i more comparable both over time and in the cross section, prior studies typically scale PVRI_t^i by the current book value of common equity (Lee 2014; Asness et al. 2019). Hence, one could define firm quality as follows:

$$
firm\ quality = E_t(Q_t^i) = \frac{\mathrm{PVRI}_t^i}{B_t^i} = \sum_{j=1}^{\infty} \frac{E_t[(\mathrm{ROE}_{t+j}^i - r_e^i) B_{t+j-1}^i]}{B_t^i (1+r_e^i)^j} \tag{2}
$$

Equation (2) does not have a closed form solution. The fundamental analysis literature typically uses an *n*-period expansion of equation (2) by assuming that the future residual income (RI) grows at a constant rate g starting from period *n*+1. Following Frankel and Lee (1998), we use a one-year expansion of equation (2) and assume $g = 0$.[2] As a result, equation (2) can be further simplified as follows:

[2] We find similar inferences if we use a two-year or three-year expansion of equation (2).

$$firm\ quality = E_t(Q_t^i) = \frac{\text{PVRI}_t^i}{B_t^i} = \frac{(E_t[\text{ROE}_{t+1}^i] - r_e^i)}{r_e^i} \tag{3}$$

By replacing $E_t[\text{ROE}_{t+1}^i]$ with ROE_{t+1}^i, which is realized in year *t*+1, we derive the following measure of a firm's *true* firm quality (i.e., the ground truth) that will be used in subsequent machine learning:[3]

$$Q_t^i = \frac{\text{PVRI}_t^i}{B_t^i} = \frac{(\text{ROE}_{t+1}^i - r_e^i)}{r_e^i} \tag{4}$$

As it is difficult to estimate a firm-specific cost of equity capital, r_e^i is proxied using the industry cost of capital as of year *t*. The industry cost of capital is defined as the sum of the industry risk premium calculated following Fama and French (1997, the last column of Table 7) and a risk-free rate equal to the average annualized 30-day T-bill rate.

This study's primary goal is to develop an empirical proxy for Q_t^i. We model Q_t^i as a function of various fundamental variables available as of current year *t*:

$$Q_t^i = f(\boldsymbol{x}_t^i) \tag{5}$$

Where $\boldsymbol{x}_t^i$ is a *P*-dimensional vector of fundamental variables. The fundamental analysis literature in accounting and finance has been devoted to identifying proxies for Q_t^i using heuristic approaches. In contrast, we use machine learning to estimate model (5). After estimating the function $f(\cdot)$, we construct a proxy for firm quality (denoted as $\hat{Q}_t^i$) for firm *i* at calendar year *t*.

Figure 1 shows the timeline for our firm quality prediction model. Following Hou et al. (2012) and Li and Mohanram (2014), we perform the prediction as of June 30 for each calendar

[3] Chen et al. (2022) use machine learning to predict the sign of annual earnings change, which is defined as (eps_{t+1} – eps_t – drift_{t+1}), where eps is the earnings-per-share before extraordinary item and drift_{t+1} is the mean eps change over the four years prior to year t+1. The Pearson correlation between Chen et al.'s earnings change sign and Q_t^i is 0.08 only, suggesting that Q_t^i is conceptually different from the dependent variable of Chen et al. (2022).

year *t*. To avoid any look-ahead biases, we assume that only financial information for firms with fiscal year ending (FYE) prior to April 1 of calendar year *t* is available on June 30 of calendar year *t*. We compute $\widehat{Q}_t^i$ on June 30 of calendar year *t* and then form trading portfolios on July 1 of calendar year *t*.

As Asness et al. (2019) is our benchmark paper, we start with their 19 variables used to construct their firm quality measure. Table A1 in the Appendix lists the definitions of the 19 variables. Asness et al. identify four components of firm quality derived from a unifying valuation framework, including profitability, growth, safety and payout. Then, they use the standardized values of the 19 variables to construct the four components of firm quality (referred to as Asness' Q score). Their Q score is the sum of these four components.

2.2. An introduction to XGBoost

Asness et al.'s (2019) construction of their Q score is based on human heuristics. To find the optimal weights of fundamental variables, we use machine learning to approximate Equation (5). The simplest form of machine learning model we consider is the linear ordinary least squares (OLS) regression. As the functional form of equation (5) could be nonlinear, we also consider nonlinear models that can allow nonlinear effects and interaction effects among the predictors. We choose XGBoost, a state-of-art machine learning method belonging to ensemble learning, which aims to improve model performance by combining single models. We also use a Pseudo Huber loss for XGBoost to further increase its robustness to the influence of outliers. In the remainder of this section, we provide a brief introduction to XGBoost.

Proposed by Chen and Guestrin (2016), XGBoost, which stands for "extreme gradient boosting", is an implementation of machine learning algorithms under the gradient boosting framework originating from Friedman (2001). XGBoost refers to the engineering goal to push the limit of computation resources for gradient boosting trees (GBT). In other words, XGBoost

is an improved GBT in terms of execution speed and model performance. In this section, we mainly introduce the GBT algorithms and simply discuss how XGBoost makes improvements for the algorithms.

As a member of ensemble learning, GBT is a tree-based method that combines multiple decision trees via gradient boosting technique. The decision tree is a common machine learning method for allowing interactions of independent variables and nonlinearity (Gu et al. 2020), which follows a divide-and-conquer procedure to make prediction for each observation. Figure 2 provides a simple example of a decision tree to show how it works in our regression task. For a given observation, the example decision tree makes a prediction based on the variables ROE and ROA of the observation: If the ROE of the observation is larger than 0.3, the example decision tree will predict its quality measure to be V_1; If not, the decision tree will further check if the observation's ROA is smaller than 0.1, and gives the prediction based on this judgement. The decision tree can grow deeper when using more independent variables to make predictions.

Using observations of the training set, the decision tree can be constructed by choosing the independent variable in each node and determining the cutoff of each independent variable. The typical algorithm tries every possible cutoff for each independent variable and chooses the variable and cutoff that minimizes the prediction error. The tree stops growing when adding a node cannot reduce the prediction error, or a tree attribute reaches a pre-set threshold. For the regression task, the prediction for a leaf node is the simple average of the true dependent variable values of the training observations whose values of independent variables satisfy all the cutoffs in the path of the leaf node.

Although easy to interpret, decision trees might not be accurate as it is overly dependent on the training data. The hierarchical nature of the tree-growing process makes the decision trees sensitive to small changes of data: For example, errors at the top node can heavily affect the rest of the tree (Murphy 2012). A popular solution is to combine the predictions of multiple

decision trees, which is what the GBT aims to do. GBT is based on the ensemble technique boosting that adds new decision trees to correct the errors of existing trees. Given a set of existing decision trees and their prediction errors on training data, a new decision tree is constructed by minimizing the prediction error via specific algorithms. Proposed by Friedman (2001), "gradient" refers to the gradient descent algorithm that is used to minimize the prediction error when adding new decision trees.

Based on GBT, XGBoost makes some key improvements for model performance, including the optimization of the loss function, introduction of regularization terms, support for sampling independent variables, as well as some engineering improvement that can speed up its execution. More importantly, XGBoost can handle missing values of independent variables, which could be severe for machine learning tasks using financial statement data.

2.3. Model Estimation Procedures

As our dependent variable is continuous, we adopt Mean Squared Error (MSE) and Mean Absolute Error (MAE) as our performance evaluation metrics. Even though inferences are qualitatively similar using both performance evaluation metrics, we focus on MAE in the following discussions because MAE is more robust to the influence of outliers.

To estimate the objective function in equation (5) that maps the independent variables $\boldsymbol{x}_t^i$ to true firm quality Q_t^i of firm *i* for each calendar year *t*, we need a training set to train the prediction model and a testing set to evaluate the out-of-sample performance of the model. To avoid any look-ahead biases, we use historical data over the past three years up to calendar year *t* (i.e., year *t*-3 to year *t*-1) as the training set. We use three years rather than one year of training data so that we have a sufficiently large training data set for model training. Figure 3 shows a graphical timeline of the training period and test period for a typical calendar year.

XGBoost requires tuning of hyperparameters. Table 1 shows the hyperparameters and the candidate values for the hyperparameter tuning process. We focus on tuning four key hyperparameters only and set the rest at fixed values to strike a balance between prediction accuracy and computational cost. Hyperparameters not mentioned in Table 1 are set at their default values. We use grid search to identify the optimal hyperparameters of XGBoost by employing a five-fold *time-series* cross-validation on the training dataset. As our prediction task has a temporal dimension, it is not appropriate to use the typical *k*-fold cross-validation for cross-section data that splits the training set into *k* groups randomly without considering the timing of the observations. Time-series cross-validation helps avoid the look-ahead bias in the ordinary *k*-fold cross-validation procedure as it does not introduce future data when splitting the training set and the testing set for validation (Hyndman and Athanasopoulos 2018). An intuitive illustration of our five-fold time-series cross-validation is shown in Figure 4. Specifically, the observations in the training dataset are first sorted in chronological order based on their fiscal year end (FYE), and then divided into six equal-size groups (labelled 1 to 6) in the chronological order. Group 1 is the earliest and group 6 the latest in calendar time. The five-fold cross validation is conducted as follows: the first fold takes group 1 as the training set and group 2 as the testing set; the second fold takes the groups 1 and 2 as the training set and group 3 as the testing set; we repeat this procedure for a total of five times. Given multiple alternative values of each hyperparameter, we can obtain a set of hyperparameter combinations in which the elements are the Cartesian product of the value set of each hyperparameter. By performing training and testing on each fold, the hyperparameter combination that achieves the best average performance on the five testing sets is selected. In our study, we use the average of the minimum absolute prediction error as our performance evaluation metric for selecting hyperparameters to avoid potential influence of outliers. After selecting the optimal

hyperparameters, we set the hyperparameters of the XGBoost algorithm as the optimal ones, and retrain the XGBoost model on the whole training dataset.

For both model training and testing, we standardize all the independent variables before they are fed into a model.[4] To avoid look-ahead biases, each independent variable in the testing set is standardized by the mean value and standard deviation of the same variable in the corresponding training set. Asness et al. (2019) first rank the raw values of each independent variable before performing the standardization. This approach could cause information loss because ranking ignores the magnitude of the distance between different observations. Hence, we perform the standardization on the raw values of each independent variable directly. As we aim to predict the magnitude of true firm quality, we prefer our standardization approach.

2.4. Data sources and sample Selection procedures

We obtain the raw accounting data and the stock return data from Compustat and CRSP, respectively. We match the observations from Compustat Fundamentals Annual whose FYEs are between April year *t*-1 and March year *t* with the monthly observations from CRSP whose data date is June year *t*. As the predictions are made on June 30 of calendar year t, and CRSP data is updated monthly, we make use of the latest stock information and accounting data when making predictions.

Table 2 shows the detailed sample selection procedures. Our sample starts from all available common stocks in the merged Compustat/CRSP database at the end of 2018. Our initial sample consists of 238,607 firm-year observations for the calendar years 1951-2018. We drop firms that are not in the 49 industries defined in Fama and French (1997) because the construction of firm quality requires non-missing industry cost of capital. Considering that

[4] Although the performance of XGBoost model is not affected by whether the independent variables are standardized, we do so for XGBoost models in order to be consistent with other models.

many accounting variables are not meaningful for financial firms, we also exclude firms in the financial service industry. We obtain a sample of 199,225 firm years for the period 1951-2018.

Many Compustat accounting variables contain missing values. As the missing value problem is very severe in the early years, we restrict the sample period to 1970-2018. As OLS regression cannot handle missing values, we further fill the missing values of each raw accounting data item to the extent possible based on accounting knowledge. Please see Table A2 of the Appendix for the details. After filling the missing values, we drop the firm years that still have missing values for any of the accounting variables, which reduces the sample to 185,449 firm years for the period 1970-2018.

As the calculation of certain independent variables requires data for at least four years, our sample is further reduced to 176,876 firm years for the period 1973-2017.

Similarly, requiring a non-missing dependent variable reduces the sample to 118,267 firm years. Following Frankel and Lee (1998), we further drop firm-years with negative or extremely small book value of common equity (referred to as abnormal dependent variable values in Table 2) for the calculation of the dependent variable. Extremely small book value of common equity is defined as firm-years whose book values are smaller than the 1% percentile of all firms in that year. This step reduces the sample to 113,336 firm years. Finally, we drop duplicate firm IDs (PERMNO) for each year due to fiscal year change.

To avoid the influence of outliers, we winsorize all model input variables in the final sample at the top and bottom percentiles for each calendar year. We perform the winsorization procedure annually to avoid any look-ahead biases.

3. Out-of-Sample Performance results

3.1. Results using Asness et al.'s (2019) 19 financial variables

3.1.1. Results using the one-year expansion of the residual income valuation model

Asness et al. (2019) construct their Q score using 19 individual financial variables based on valuation theory. Hence, we start with the same 19 financial variables to build machine learning models. Because Asness et al. use simple heuristics to construct their Q score without estimating equation (5), we first build a benchmark machine learning model, which is a linear OLS model using Asness' Q score as the only model input (denoted as LR-Q). Next, we examine whether we can build better machine learning models by (i) using more disaggregated data (i.e., the 19 financial variables) rather than the aggregated Q score and (ii) using XGBoost rather than OLS. As a result, we build the following three types of machine learning models: (1) OLS regression model using Asness et al.'s 19 financial variables as model inputs (denoted as LR-19); (2) XGBoost model using Asness' Q score as the only model input (denoted as XGBoost-Q); and (3) XGBoost model using Asness et al.'s 19 financial variables as model inputs (denoted as XGBoost-19).

Table 3 shows the summary statistics (mean) of the out-of-sample performance metrics for the three machine learning models versus the benchmark model over the test period 1976-2018. The test sample starts from 1976 because our sample starts from 1973 and model training requires a three-year rolling window (see Figure 3). For each model, Panel A of Table 3 reports the summary statistics for the two performance evaluation metrics, MSE and MAE. Panel B of Table 3 conducts formal statistical tests on the performance difference for each pair of machine learning models. As inferences are qualitatively similar using both performance evaluation metrics, we focus on MAE in the following discussions for brevity.

We first examine whether it is possible to develop a more accurate OLS prediction model by using the 19 disaggregated financial variables rather than the aggregated Q score.

The mean MAE is 2.206 for LR-19, representing a reduction of 16% relative to the mean MAE of 2.626 for the benchmark model LR-Q. This difference is both statistically and economically significant. Therefore, there is clear evidence that disaggregated financial variables are more useful than the aggregated Q score in constructing firm quality, even when one uses the linear OLS regression.

Second, we examine whether it is possible to build an even better firm quality prediction model using XGBoost. The mean MAE is 1.907 for XGBoost-19, representing a reduction of 14% relative to the mean MAE of 2.206 for LR-19. This difference is both statistically and economically significant. Thus, we find strong evidence that the more advanced machine learning model, XGBoost, can further help improve the performance of the firm quality prediction model, while holding the model inputs constant. Combining the benefits of both disaggregated data (19 variables) and a more advanced machine learning model (XGboost), our best model XGBoost-19 outperforms LR-Q by more than 27% in terms of MAE (1.907 versus 2.626).

So far we have assessed the performance of all machine learning models for all test years as a whole. To examine the stability of our best model, XGBoost-19 versus the other models, we tabulate the values of MAE for all the models in each test year in Figure 5. We find that the performance of all models suffers for the years around financial crises (e.g., 1988, 1994, 2002, 2008). More importantly, the performance of our best model XGBoost always dominates the performance of the other models. In addition, the performance of the benchmark model LR-Q is almost always the worst for all years.

3.1.2. XGBoost with missing values

Recall that we filled many raw accounting data items with missing values based on accounting knowledge. To examine whether this special treatment affects our inferences, we

re-estimate the XGBoost-19 model for the same sample without filling the missing values (referred to as XGBoost-19m). Table 3 shows the performance result of XGBoost-19m. We find no statistically significant difference in performance for XGBoost-19 versus XGBoost-19m, suggesting that the missing value treatment does not drive the superior performance of XGBoost-19 in Table 3.

3.1.3. XGBoost for the full sample without deleting observations with missing values

The sample size for Asness' Q score is much larger than our final sample used in Table 3 because Asness et al. (2019) drop any of the 19 individual proxies if it contains missing values and use the remaining non-missing proxies to construct the Q score. Our final sample used in Table 3 loses approximately 27.7% of the full sample before the missing value treatment. To examine whether this sample size difference affects our inference, we re-estimate the LR-Q model and the XGBoost-19 model for the full sample without dropping the observations whose independent variables contain any missing values (referred to as LR-Q_Full and XGBoost-19m_Full, respectively). Untabulated results show that the MAE is still significantly larger for LR-Q_Full than for XGBoost-19m_Full, suggesting that the sample size difference does not drive the superior performance of XGBoost-19m relative to LR-Q in Table 3.

3.1.4. Results using two-year and three-year expansions of the residual income valuation model

The fundamental analysis literature also defines PVRI using the two- or three-year expansion of the residual income model (Frankel and Lee 1998). As a robustness check, we use the two-year and three-year expansions of PVRI to define true firm quality Q_t as follows:

$$Q_t^i = \frac{\text{PVRI}_t^i}{B_t^i} = \frac{(\text{ROE}_{t+1}^i - r_e^i)}{(1+r_e^i)} + \frac{(\text{ROE}_{t+2}^i - r_e^i)B_{t+1}^i}{r_e^i(1+r_e^i)B_t^i} \quad (6)$$

$$Q_t^i = \frac{\text{PVRI}_t^i}{B_t^i} = \frac{(\text{ROE}_{t+1}^i - r_e^i)}{(1+r_e^i)} + \frac{(\text{ROE}_{t+2}^i - r_e^i)B_{t+1}^i}{(1+r_e^i)^2 B_t^i} + \frac{(\text{ROE}_{t+3}^i - r_e^i)B_{t+2}^i}{r_e^i(1+r_e^i)^2 B_t^i} \quad (7)$$

The multiple-year expansion of the residual income valuation model imposes more demand on data availability because a firm may not always exist for multiple years into the future. For example, the three-year expansion of the model could be missing due to missing values for ROE_{t+3}. To avoid losing observations for the three-year expansion of the residual income model, we use the two-year expansion of the model as a replacement. If the two-year expansion of the model is also missing, we use the one-year expansion as a replacement.

Table 4 shows the performance results for the two-year expansion and three-year expansion of the valuation model. The mean MAE grows as *n* increases. This finding is expected because as *n* increases, the prediction of firm quality will cover a longer time horizon in the future and hence becomes more challenging. More importantly, the pecking order of the machine learning models remains the same. XGBoost-19 remains the best prediction model for firm quality. Because inferences are qualitatively similar using one-year expansion versus two-year or three-year expansion of the valuation model, we will focus on the one-year expansion for the subsequent analyses.

3.2. XGBoost results using more input variables

Asness et al. (2019) identified the 19 financial variables based on their conceptual valuation framework and prior research. We explore the possibility of building even more powerful machine learning models by using more input variables. We perform three types of extensions in this section. As XGBoost has proved to be a more powerful prediction method in Table 3, we use XGBoost only for the following models. In addition, we allow all of the following models to allow missing values so that we maintain the same sample for all model

comparisons. Accordingly, we also use XGBoost-19m as our benchmark model. Overall, we find little economically significant differences between XGBoost-19m and the XGBoost models with more model inputs.

3.2.1. XGBoost results using disaggregation of the 19 financial variables

Given the evidence in Table 3 on the benefit of using disaggregated data for prediction, we examine whether we can build a more accurate prediction model by using the annual raw accounting data items that are used to construct the 19 financial variables. This treatment leads to a total of 63 input variables (see Table A3 in the Appendix for the details).[5] We refer to this model as XGBoost-63m.

Table 5 shows the performance results of XGBoost-63m. The mean MAE is 1.872 for XGBoost-63m, representing a reduction of 1.8% relative to the mean MAE of 1.908 for the benchmark model XGBoost-19m. This difference is statistically significant, but it does not seem economically significant.

3.2.2. XGBoost results using more fundamental signals beyond Asness et al. (2019)

Even though Asness et al. (2019) identify a comprehensive list of 19 fundamental signals, their list is not exhaustive. Hence, we next explore whether we can construct a better firm quality prediction model by incorporating more fundamental signals from other studies. For this purpose, we identify two recent studies in the fundamental analysis literature: Li and Mohanram (2019) and Bartram and Grinblatt (2018). Li and Mohanram (2019) combines the FSCORE of Piotroski (2000) and GSCORE of Mohanram (2005) to measure firm quality. Piotroski (2000) chooses nine fundamental signals to measure three areas of a firm's financial

[5] The exception is that we still keep the variables *beta* and *EVOL* in their original forms as they are from stock market data and quarterly financial data instead of annual financial data.

condition: profitability, financial leverage/liquidity, and operating efficiency. Mohanram (2005) classifies his GSCORE into three categories: Earnings and cash flow profitability, naive extrapolation and accounting conservatism. Bartram and Grinblatt (2018) simply choose 28 most common raw accounting data items from the three financial statements as predictors. The fundamental signals used by Asness et al. (2019), Li and Mohanram (2019), and Bartram and Grinblatt (2018) are derived from 87 raw accounting data items. See Table A3 in the Appendix for the detailed list. As we have shown the benefit of using disaggregated accounting data items for prediction, we use the 87 raw data items to construct the firm quality prediction model (referred to as XGBoost-87m).

Table 5 shows the performance results of XGBoost-87m. The mean MAE is 1.853 for XGBoost-87m, representing a reduction of 2.9% relative to the mean MAE of 1.908 for XGBoost-19m. This difference is statistically significant, but it does not seem economically significant. In addition, the incremental reduction in MAE for XGBoost-87m relative to XGBoost-63m is only 1%.

3.2.3. XGBoost results using more raw accounting data from Compustat

Compustat industrial annual database contains a total of 318 raw accounting data items, including those included in XGBoost-87m. Our final exploratory analysis builds an XGBoost model using the 318 raw accounting data items (referred to as XGBoost-318m). As shown in Table 5, the mean MAE is 1.858 for XGBoost-318m, not statistically different from the mean MAE for XGBoost-87m. This finding suggests no evidence that using all available raw accounting data items can build a better firm quality prediction model than XGBoost-87m.

4. Contemporaneous stock pricing of firm quality

The empirical analyses in Section 3 show that we can build a more powerful firm quality prediction model using machine learning and disaggregated accounting data (i.e., XGBoost-19m) than an OLS model based on Asness et al.'s (2019) Q score (i.e., LR-Q). In this and the next section we show the value of our machine learning model-based firm quality measure relative to Asness' Q score within the context of value investing. However, we wish to emphasize that our machine learning based measure of firm quality can be used in a variety of contexts beyond value investing. In this section, we first examine whether firm quality based on XGBoost-19m can better explain contemporaneous stock pricing than Asness' Q score. To test this hypothesis, we follow Asness et al. (2019) by running the following Fama and MacBeth (1973) cross-sectional regression:

$$P_t^i = a + bQuality_t^i + Controls + \varepsilon_t^i \tag{8}$$

where P_t^i is firm *i*'s market-to-book ratio in natural logarithm at the end of June of calendar year *t*. $Quality_t^i$ is a proxy of firm quality. We consider three measures of firm quality: (i) Asness' Q score; (ii) predicted firm quality from XGBoost-19m; (iii) true firm quality based on future realized abnormal income (i.e., Q_t^i in equation (4)). Even though investors cannot directly observeQ_t^i, we include it as a benchmark. Following Asness et al. (2019), *Controls* include *Firm_Size*, *One_Year_Return*, *Firm_Age*, *Profit_Uncertainty*, *Dividend_Payer*, *Profit_Uncertainty_by_Dividend_Payer*, and industry fixed effects. See Table A4 in the Appendix for the definitions of all control variables. Following Asness et al (2019), all the explanatory variables (except for the dummies) are measured as the z-score of their cross-sectional rank.

Table 6 reports the time series averages of the regression coefficients of model (8). We adjust the standard errors for heteroskedasticity and autocorrelation of five lags (Newey and

West 1987). Not surprisingly, the coefficient on the true firm quality measure Q_t^i in column (1) is significantly positive, consistent with the forward-looking nature of the stock market. The average adjusted R^2 for the model in column (1) is 43%. More importantly, the coefficients on both Asness' Q score and firm quality based on XGBoost-19m are significantly positive, suggesting that they are both reasonable proxies for firm quality. However, the average adjusted R^2 for the model using XGBoost-19m is 42% while the average adjusted R^2 for the model using Asness' Q score is 40%. This finding suggests that our firm quality measure based on XGBoost-19m does a better job in explaining contemporaneous stock prices than Asness' Q score. We view this as another piece of complementary evidence to Table 3 that our firm quality based on XGBoost-19m is of higher quality than Asness' Q score.

5. Stock return prediction using firm quality based on XGBoost-19m

Our final empirical analysis investigates whether it is possible to earn abnormal returns by building long-short hedging investment portfolios based on firm quality derived from XGBoost-19m. We use Asness' Q score as our benchmark. As noted in the Introduction section, the answer to this question is ambiguous because the XGBoost-19m proxy is based on publicly available information only. If one assumes that the stock market is fully efficient in the semi-strong form, we should not be able to earn abnormal stock returns from our hedging portfolios. Hence, our hedging portfolio analysis is a joint hypothesis of market efficiency and informativeness of our firm quality proxy.

Consistent with the value investing philosophy of buying high quality stocks at reasonable prices, we construct our hedging portfolios using double sorting for each calendar year: first sort all stocks into 10 deciles based on change in firm quality; second, for all stocks in each decile, we sort them into 10 deciles based on the book-to-market ratio. We focus on the returns from the two most extreme portfolios in the double sort: (i) long on the stocks that are

in the top decile of the book-to-market ratio within the top decile of firm quality change; and (2) short on the stocks that are in the bottom decile of the book-to-market ratio within the bottom decile of firm quality change. We use change rather than level of firm quality in the construction of our hedging portfolios because stock markets react to new information, which is better measured by change of firm quality (Ball and Brown 1968; Chen et al. 2022).

We use the standard procedures to construct our hedging portfolios. Figure 6 shows the timeline of the portfolio analysis. Our firm quality proxies are developed using publicly available information prior to July 1 of calendar year *t*. Hence, there is no look-ahead biases in the abnormal returns generated from our trading strategy. Following Asness et al. (2019), we hold each hedging portfolio for one year, rebalanced every month using equal weighting. Each portfolio's alpha is computed using a standard time series factor pricing model (see below).

We compute abnormal stock returns using the q^5 factor pricing model proposed by Hou et al. (2021). This model is comprised of the four factors in Hou et al. (2015), including the market factor MKT, the size factor ME, the return on equity factor ROE (i.e., the profitability factor), and the investment factor IA and the expected investment growth factor EG in Hou et al. (2021). Asness et al. (2019) find that their Q score can generate future abnormal stock returns using the four-factor model that includes the three factors in Fama and French (1993) and the momentum factor, but Hou et al. (2022) argue that many market anomalies including Asness et al.'s abnormal returns disappear once they use the q^5 factor model.

Table 7 shows the monthly alpha for our hedging portfolios. Panel A uses change in Asness' Q score (i.e., Q score in year *t* minus Q score in year *t*-1) as our proxy for the change in firm quality. Panel B defines change in firm quality as the change in predicted firm quality based on XGBoost-19m relative to the realized firm quality in year *t*-1 (i.e., $\hat{Q}_t^i - Q_{t-1}^i$). Using Asness' Q score as a proxy for firm quality, the monthly alpha for the long minus short hedging portfolio is 0.84% and statistically significant. More importantly, using firm quality based on

XGBoost-19m, we find that the monthly alpha for the long minus short hedging portfolio is an even larger 2.18% and statistically significant. It is also worth noting that most of the long-short hedging portfolio's alpha is driven by the long position for both firm quality proxies. Overall, these results suggest that the stock market is not fully semi-strong efficient; moreover, our machine learning based firm quality measure can be used to build a more profitable stock trading strategy than Asness et al.'s (2019) Q score based on human heuristics.

The abnormal return analysis in Table 7 covers a fairly long time period of more than 40 years. As many things have changed over the period, including the availability of information technologies (e.g., algorithm trading) that could have significantly changed the ways people make investment decisions and the efficiency of the stock market. Hence, one may wonder whether the results show in Table 7 hold for the more recent time period. To address this question, we redo Panel B of Table 7 by dividing our sample period into three equal sub-periods. The results are reported in Table 8. The monthly alpha from the value investing strategy is significantly positive and economically significant for all three sub-periods. In particular, the monthly alpha is 1.81% for the most recent period 2004-2018, suggesting that our value investing trading strategy based on XGBoost-19m continues to work even in the most recent time period.

Table 5 shows that XGBoost-63m, XGBoost-87m and XGBoost-138m all outperform XGBoost-19m in predicting firm quality statistically but not economically. Table 9 examines whether the double-sort value investing trading strategy based on XGBoost-63m, XGBoost-87m or XGBoost-138m can beat the same trading strategy based on XGBoost-19m. The monthly alphas for XGBoost-63m, XGBoost-87m and XGBoost-138m are also smaller than the monthly alpha for XGBoost-19m. This finding may not be too surprising given that these models do not show economically significant improvements in predicting firm quality relative to XGBoost-19m.

6. Conclusion

Firm quality is of paramount importance to not only value investors but also a variety of other stakeholders, including corporate managers, governance activists and government regulators. The fundamental analysis literature relies on simple heuristics to measure firm quality based on publicly observable fundamental signals. The objective of this study is to examine whether we can use machine learning to construct a better proxy for firm quality. As firm quality is a forward-looking concept, we argue that machine learning should be more powerful than human heuristics in measuring firm quality.

We use the human-built firm quality proxy (referred to as Asness' Q score) by Asness et al. (2019) as our benchmark. Asness et al. (2019) is one of the most recent studies in the fundamental analysis literature. They construct their Q score based on a comprehensive list of 19 standardized fundamental signals derived from valuation theory and publicly available financial statement data. Since Asness et al. (2019) constructed their Q score without referring to the ground truth (i.e., future realized firm quality), we develop a baseline benchmark model that uses Asness' Q score as the only model input and the linear OLS regression as the machine learning model (referred to as LR-Q). We construct our firm quality proxy based on the same 19 standardized fundamental signals but we use XGBoost, one of the state-of-art machine learning methods that can accommodate regression variables with missing values (referred to as XGBoost-19m). We show that XGBoost-19m easily beats LR-Q by a significant margin in out-of-sample performance. We also explore the possibility of building a better firm quality prediction model than XGBoost-19m and find limited success, suggesting the importance of theoretical guidance in model input selection.

We illustrate the usefulness of our machine learning models in the context of value investing. We show that firm quality based on XGBoost-19m can better explain

contemporaneous stock prices than Asness' Q score. In addition, we find that a hedging portfolio based on the double sorting of the change in firm quality and the book-to-market ratio (a proxy for a stock's cheapness) yields an economically larger abnormal return over the 12-month investment horizon for XGBoost-19m than for Asness' Q score.

The findings of our study raise many exciting new research opportunities for the fundamental analysis literature. We identify a few that are worth exploring. First, what are the additional fundamental signals one could consider to build an even more powerful machine learning prediction model of firm quality? Second, are there more advanced machine learning models (e.g., deep learning) that are more powerful than XGBoost in constructing firm quality? Third, how can the availability of our more powerful firm quality proxy based on XGBoost aid future research that requires a measure of firm quality?

References

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2019). Quality minus Junk. *Review of Accounting Studies*, 24(1), 34-112.

Ball, R., & Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*, 6(2), 159-178.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, 58(1), 199-235.

Bartram, S. M., & Grinblatt, M. (2018). Agnostic Fundamental Analysis Works. *Journal of Financial Economics*, 128(1), 125-147.

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using Machine Learning to Detect Misstatements. *Review of Accounting Studies*, 26(2), 468-519.

Binz, O., K. Schipper, K. Standridge. 2022. What Can Analysts Learn from Artificial Intelligence about Fundamental Analysis? Working paper.

Cao, K., and H.F. You. 2021. Fundamental Analysis via Machine Learning. Working paper.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Chen, X., Cho, Y. H., Dou, Y., & Lev, B. (2022). Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data. *Journal of Accounting Research*, 60(2), 467-515.

Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. (2020). Machine Learning Improves Accounting Estimates: Evidence from Insurance Payments. *Review of Accounting Studies*, 25(3), 1098-1134.

Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3-56.

Fama, E. F., & French, K. R. (1997). Industry Costs of Equity. *Journal of Financial Economics*, 43, 153-193.

Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81(3), 607-636.

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

Frankel, R., & Lee, C. M. C. (1998). Accounting Valuation, Market Expectation, and Cross-Sectional Stock Returns. *Journal of Accounting and Economics*, 25(3), 283-319.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189-1232.

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

Hou, K., Mo, H., Xue, C., & Zhang, L. (2022). The Economics of Security Analysis. *Management Science*, Forthcoming.

Hou, K., Mo, H., Xue, C., & Zhang, L. (2021). An Augmented q -Factor Model with Expected Growth*. *Review of Finance*, 25(1), 1-41.

Hou, K., van Dijk, M. A., & Zhang, Y. (2012). The Implied Cost of Capital: A New Approach. *Journal of Accounting and Economics*, 53(3), 504-526.

Hou, K., Xue, C., & Zhang, L. (2015). Digesting Anomalies: An Investment Approach. *Review of Financial Studies*, 28(3), 650-705.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Karpoff, J.M., and M.D. Wittry. 2018. Institutional and Legal Context inmNatural Experiments: The Case of State Antitakeover Laws. *Journal of Finance* 73(2), 657-714.

Lee, C. M. C. (2014). Performance Measurement: An Investor's Perspective. *Accounting and Business Research*, *44*(4), 383-406.

Lee, C. M. C. and E. C. So. 2014. Alphanomics: The Informational Underpinnings of Market Efficiency. *Foundations and Trends R in Accounting*, 9 (2–3), 59–258.

Li, K. K., & Mohanram, P. (2014). Evaluating Cross-Sectional Forecasting Models for Implied Cost of Capital. *Review of Accounting Studies*, 19(3), 1152-1185.

Li, K., & Mohanram, P. (2019). Fundamental Analysis: Combining the Search for Quality with the Search for Value. *Contemporary Accounting Research*, 36(3), 1263-1298.

Mohanram, P. S. (2005). Separating Winners from Losers among Low Book-to-Market Stocks using Financial Statement Analysis. *Review of Accounting Studies*, 10, 133-170.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703-708.

Nissim, D., and S. H. Penman. 2001. Ratio Analysis and Equity Valuation: From Research to Practice. *Review of Accounting Studies* 6 (1):109-154.

Piotroski, J. D. (2000). Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers. *Journal of Accounting Research*, 38, 1-41.

Piotroski, J. D., & So, E. C. (2012). Identifying Expectation Errors in Value/Glamour Strategies: A Fundamental Analysis Approach. *Review of Financial Studies*, 25(9), 2841-2875.

Appendix

Table A1. Comparison of Independent Variables with Asness et al. (2019)

	Name	Definition	Whether the definition is the same as that of Asness et al. (2019)	Definitions in Asness et al. (2019)	Modified definitions
Profitability	**GPOA**	Gross profits over assets	Yes	$(REVT - COGS)/AT$	-
	ROE	Return on equity	Yes	IB/BE	-
	ROA	Return on assets	Yes	IB/AT	-
	CFOA	Cash flow over assets	Yes	$(NI + DP - \Delta WC - CAPX)/AT$	-
	GMAR	Gross margin	Yes	$(REVT - COGS)/SALE$	-
	ACC	Low accruals	Yes	$-(\Delta WC - DP)/AT$	-
Growth	**ΔGPOA**	Three-year growth in residual gross profits over assets	No	$[(gp_t - r^f at_{t-1}) - (gp_{t-5} - r^f at_{t-6})]/at_{t-5}$	$[(gp_t - r^f at_{t-1}) - (gp_{t-2} - r^f at_{t-3})]/at_{t-2}$
	ΔROE	Three-year growth in residual return on equity	No	$[(ib_t - r^f be_{t-1}) - (ib_{t-5} - r^f be_{t-6})]/be_{t-5}$	$[(ib_t - r^f be_{t-1}) - (ib_{t-2} - r^f be_{t-3})]/be_{t-2}$
	ΔROA	Three-year growth in residual return over assets	No	$[(ib_t - r^f at_{t-1}) - (ib_{t-5} - r^f at_{t-6})]/at_{t-5}$	$[(ib_t - r^f at_{t-1}) - (ib_{t-2} - r^f at_{t-3})]/at_{t-2}$
	ΔCFOA	Three-year growth in residual cash flow over assets	No	$[(cf_t - r^f at_{t-1}) - (cf_{t-5} - r^f at_{t-6})]/at_{t-5}$	$[(cf_t - r^f at_{t-1}) - (cf_{t-2} - r^f at_{t-3})]/at_{t-3}$
	ΔGMAR	Three-year growth in gross margin	No	$(gp_t - gp_{t-5})/sale_{t-5}$	$(gp_t - gp_{t-2})/sale_{t-2}$
Safety	**BAB**	Low market beta	No	-Beta (following Frazzini and Pedersen 2014)	-Beta (obtained directly from CRSP)
	LEV	Low leverage	Yes	$-(DLTT + DLC + MIBT + PSTK)/AT$	-
	Ohlson's O	Low bankruptcy risk (Ohlson 1980)	Yes	$-(-1.32 - 0.407\log(ADJASSET/CPI) + 6.03TLTA - 1.43WCTA + 0.076CLCA -$	-

				1.72OENEG - 2.37NITA - 1.83FUTL + 0.285INTWO - 0.521CHIN)	
	Altman's Z	Low bankruptcy risk (Altman 1968)	Yes	*(1.2WC + 1.4RE + 3.3EBIT + 0.6ME + SALE) / AT*	-
	EVOL	Low earnings volatility	No	standard deviation of IBQ/BEQ over the past 60 quarters or IB/BE over the past 5 years (Taken as missing values if there are no more than 12 non-missing quarters or five non-missing fiscal years)	(-1)×standard deviation of IBQ/BEQ over the past 60 quarters or IB/BE over the past 5 years (Taken as missing values if there are no more than 12 non-missing quarters or three non-missing fiscal years)
Payout	**EISS**	Net equity issuance	Yes	$- \log(SHROUT\_ADJ_t / SHROUT\_ADJ_{t-1})$	-
	DISS	Net debt issuance	No	$- \log (TOTD_t / TOTD_{t-1})$	$- \text{sgn}(TOTD_t - TOTD_{t-1}) * \log(1+ \lvert TOTD_t - TOTD_{t-1} \rvert)$; sgn(x) = 1 if x>0, else sgn(x) = -1
	NPOP	Total net payout over profits	No	$\Sigma_5 (IB - \Delta BE) / \Sigma_5 (REVT - COGS)$	$\Sigma_3 (IB - \Delta BE) / \Sigma_3 (REVT - COGS)$

This table describes the definition of the 19 variables from Asness et al. (2019). Some of the variables are defined differently in our analysis to minimize losing too many observations due to the missing value problem.

Table A2. Methods for Filling Missing Data Items

Variable Name	Variable Description	Data Source	Our Way to Fill Missing Values	Rationale for the filling method	Number of Filled Missing Values / Number of Total Missing Values
act	ACT -- Current Assets - Total	Compustat	Use lagged value	ACT is a stock variable, so it is reasonable to use lagged value to fill missing observations	297/4550
at	AT -- Assets - Total	Compustat	Use lagged value	AT is a stock variable	76/244
capx	CAPX -- Capital Expenditures	Compustat	Use 0	Missing values in CAPX may indicate CAPX is not material, so filling with 0 is appropriate	2673/2673
ceq	CEQ -- Common/Ordinary Equity - Total	Compustat	Use lagged value	CEQ is a stock variable	85/408
che	CHE -- Cash and Short-Term Investments	Compustat	Use 0	Lagged value may not be appropriate	757/757
cogs	COGS -- Cost of Goods Sold	Compustat	Use (1-lagged gross profit margin)*current sales	Usually a firm's gross profit margin should not change significantly over time; gross profit margin = (1 – cogs/sale)	5/512
csho	CSHO -- Common Shares Outstanding	Compustat	Use lagged value	CSHO usually does not change too much within one year	107/288
dlc	DLC -- Debt in Current Liabilities - Total	Compustat	Use 0	DLC is a stock variable	565/565
dltt	DLTT -- Long-Term Debt - Total	Compustat	Use lagged value	DLTT is a stock variable	218/618
dp	DP -- Depreciation and Amortization	Compustat	Use 0	Missing values in DP may indicate DP is not material, so filling with 0 is appropriate	840/840
ebit	EBIT -- Earnings Before Interest and Taxes	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	2589/2589
gp	GP -- Gross Profit (Loss)	Compustat	Use (1 – lagged cogs/current sale)	Similar to cogs	5/512
ib	IB -- Income Before Extraordinary Items	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	502/502
lct	LCT -- Current Liabilities - Total	Compustat	Use lagged value	LCT is a stock variable	311/3563

lt	LT -- Liabilities - Total	Compustat	Use lagged value	LT is a stock variable	192/614
mib	MIB -- Minority Interest (Balance Sheet)	Compustat	Use lagged value if it exists; Else 0	MIB is a stock variable; In most case the missing values can be regarded as 0 since some firms do not have this item and do not report it	11836/11836
mibt	MIBT -- Noncontrolling Interests - Total - Balance Sheet	Compustat	Use lagged value if it exists; Else 0	MIBT is a stock variable; In most case the missing values can be regarded as 0 since some firms do not have this item and do not report it	12819/12819
ni	NI -- Net Income (Loss)	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	510/510
pi	PI -- Pretax Income	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	507/507
pstkl	PSTKL -- Preferred Stock Liquidating Value	Compustat	Use 0	Many firms may not have perferred stocks, so filling with 0 is appropriate	333/333
pstkrv	PSTKRV -- Preferred Stock Redemption Value	Compustat	Use 0	Many firms may not have perferred stocks, so filling with 0 is appropriate	359/359
re	RE -- Retained Earnings	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	671/671
revt	REVT -- Revenue - Total	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	499/499
sale	SALE -- Sales/Turnover (Net)	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	499/499
seq	SEQ -- Stockholders' Equity - Total	Compustat	Use lagged value	SEQ is a stock variable	113/1172
txp	TXP -- Income Taxes Payable	Compustat	Use 0	Flow variable; Lagged value may not be appropriate	3155/3155
prcc_f	PRCC_F -- Price Close - Annual - Fiscal	Compustat	Use the most recent available price in the last one month in crsp	0 may not be appropriate	1349/2462
SHROUT	SHROUT -- Number of Shares Outstanding	CRSP	Use lagged value	SHROUT usually does not change too much within one year	24/24
CFACSHR	CFACSHR --Cumulative Factor to Adjust Shares	CRSP	Use lagged value	0 may not be appropriate	24/24

pstk	PSTK -- Preferred/Preference Stock (Capital) - Total	Compustat	Use 0	Many firms may not have perferred stocks, so filling with 0 is appropriate	293/293
Beta		CRSP	Use lagged value	Use the data of the latest month; The interval between the latest month and the current month should be less than 1 year	1165/3331

This table describes how we fill missing values of the accounting variables that are used in the dependent and independent variables of our model. Based on knowledge of the accounting meaning of the variables, we design different methods for filling the missing values. Specifically, according to the different types of accounting variables (like stock variable versus flow variable), we fill the missing values by lagged value or zero. If the missing values cannot be filled (e. g., the lagged values are also missing), we simply let them remain missing. The last column shows the number of filled missing values and the total number of missing values for the 185449 observations in 1970-2018 reported in the fourth step of Table 2's sample selection.

Table A3. Input variables for XGBoost-63m and XGBoost-87m

Panel A: Input variables for XGBoost-63m			
	Definition	**Period**	**Related financial ratios**
act	Current Assets - Total	*t*, *t*-1, *t*-2, *t*-3	CFOA, ACC, ΔCFOA, Ohlson's O, Altman's Z
at	Assets - Total	*t*, *t*-1, *t*-2, *t*-3	GPOA, ROA, CFOA, ACC, ΔGPOA, ΔROA, ΔCFOA, LEV, Ohlson's O, Altman's Z
capx	Capital Expenditures	*t*, *t*-2	CFOA, ΔCFOA
che	Cash and Short-Term Investments	*t*, *t*-1, *t*-2, *t*-3	CFOA, ACC, ΔCFOA, Altman's Z
cogs	Cost of Goods Sold	*t*, *t*-1, *t*-2	GPOA, GMAR, ΔGPOA, ΔGMAR, NPOP
dlc	Debt in Current Liabilities - Total	*t*, *t*-1, *t*-2, *t*-3	CFOA, ACC, ΔCFOA, LEV, Ohlson's O, Altman's Z, DISS
dltt	Long-Term Debt - Total	*t*, *t*-1	LEV, Ohlson's O, DISS
dp	Depreciation and Amortization	*t*, *t*-2	CFOA, ACC, ΔCFOA
ib	Income Before Extraordinary Items	*t*, *t*-1, *t*-2	ROA, ROE, ΔROA, ΔCFOA, ΔROE, Ohlson's O, EVOL, NPOP
ebit	Earnings Before Interest and Taxes	*t*	Altman's Z
lct	Current Liabilities - Total	*t*, *t*-1, *t*-2, *t*-3	CFOA, ACC, ΔCFOA, Ohlson's O, Altman's Z
lt	Liabilities - Total	*t*	ROE, ΔROE, Ohlson's O, EVOL, NPOP
mibt	Noncontrolling Interests - Total - Balance Sheet	*t*, *t*-1	LEV, DISS
ni	Net Income (Loss)	*t*	CFOA
pi	Pretax Income	*t*	Ohlson's O
pstk*	prefered stock value (PSTKRV, PSTKL, or PSTK depending on availability)	*t*, *t*-1, *t*-2, *t*-3	ROE, ΔROE, LEV, Ohlson's O, EVOL, TOTD, NPOP
re	Retained Earnings	*t*	Altman's Z
revt	Revenue - Total	*t*, *t*-1, *t*-2	GPOA, GMAR, ΔGPOA, ΔGMAR, NPOP
sale	Sales/Turnover (Net)	*t*, *t*-2	GMAR, ΔGMAR, Altman's Z
seq*	shareholders' equity - seq, ceq+pstk, at-(lt+mib) depending on availability	*t*, *t*-1, *t*-2, *t*-3	ROE, ΔROE, Ohlson's O, EVOL, NPOP
txp	Income Taxes Payable	*t*, *t*-1, *t*-2, *t*-3	CFOA, ACC, ΔCFOA, Altman's Z
shrout_adj	split-adjusted shares - product of shrout and cfacshr	*t*, *t*-1, *t*-2, *t*-3	ΔGPOA, ΔROE, ΔROA, ΔCFOA, ΔGMAR
beta	beta provided by CRSP	*t*	beta
me	market equity (product of shrout and price)	*t*	Altman's Z
EVOL	minus standard deviation of quarterly ROE over the past 60 quarters or annual ROE over the past 5 years depending on availability)	*t*	EVOL
Panel B: Input variables for XGBoost-87m beyond those for XGBoost-63m			
oancf	Operating Activities Net Cash Flow	*t*	CFO
sstk	Sale of Common and Preferred Stock	*t*	EQ_OFFER
VARROA	standard deviation of quarterly ROA (ibq/atq) over the past two years	*t*	VARROA
VARSGR	standard deviation of quarterly sales growth rate ((saleqt/saleqt-1)-1)) over the past two years	*t*	VARSGR
xrd	Research and Development Expense	*t*	RDINT

xad	Advertising Expense	*t*	ADINT
DVP	Dividends - Preferred/Preference	*t*	-
XIDO	Extraordinary Items and Discontinued Operations	*t*	-
IBADJ	Income Before Extraordinary Items Adjusted for Common Stock Equivalents	*t*	-
IBCOM	Income Before Extraordinary Items Available for Common	*t*	-
ICAPT	Invested Capital - Total	*t*	-
TEQ	Stockholders Equity - Total	*t*	-
PSTKR	Preferred/Preference Stock - Redeemable	*t*	-
PPENT	Property, Plant and Equipment - Total (Net)	*t*	-
CEQ	Common/Ordinary Equity - Total	*t*	-
TXT	Income Taxes - Total	*t*	-
NOPI	Nonoperating Income (Expense)	*t*	-
AO	Assets - Other	*t*	-
DO	Discontinued Operations	*t*	-
LO	Liabilities - Other - Total	*t*	-
ACO	Current Assets Other Total	*t*	-
DV	Cash Dividends (Cash Flow)	*t*	-
LCO	Current Liabilities Other Total	*t*	-
AP	Accounts Payable - Trade	*t*	-

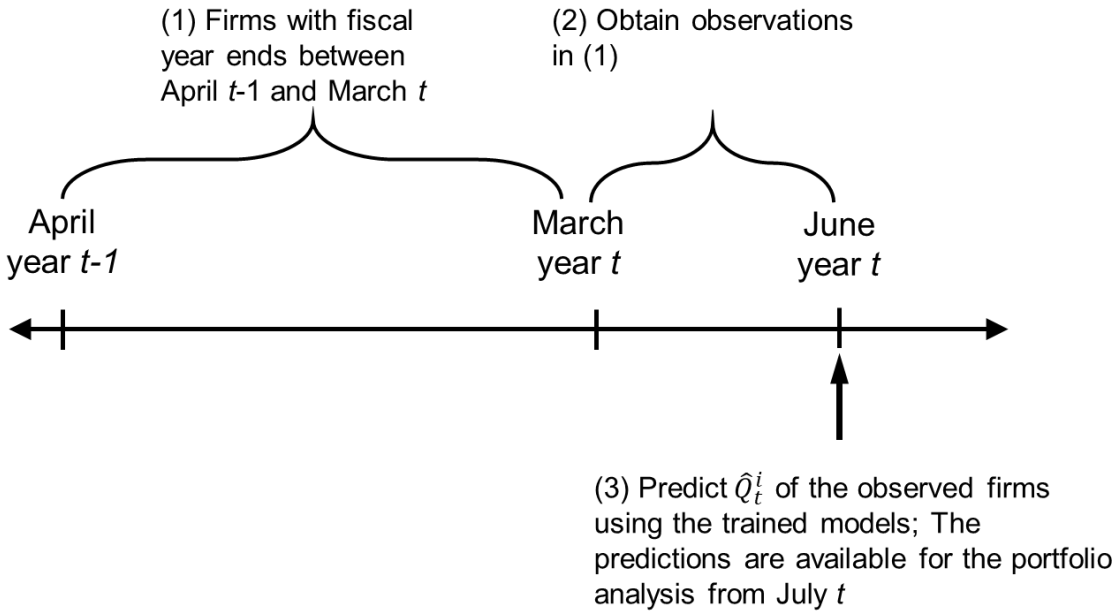
This table describes the definitions of the input variables in our XGBoost-63m model and XGBoost-87m model. Panel A shows the 63 input variables of XGBoost-63m and Panel B shows the 24 input variables for XGBoost-87m in addition to those for XGBoost-63m. Bartram and Grinblatt (2018) use quarterly data or annual data depending on which is the most recently reported but we use annual data only. As they use raw financial data items, the 24 variables in Panel B do not have related financial ratios.

Table A4. Definitions of Controls in Equation (8)

Variables	Definition
Firm_Size	The log of the firm's market capitalization *in June of year t*.
One_Year_Return	The firm's stock return over the prior year (*from June of year t-1 to June of year t*)
Firm_Age	*In June of year t*, the cumulative number of years since the first occurrence of a stock. Specifically, we look for the first occurrence of a valid stock price on CRSP, as well as the first occurrence of the valid market value in the CRSP/COMPUSTAT database, and take the earlier of the two.
Profit_Uncertainty	The standard deviation of the residuals of an AR(1) model for each firm's ROE, using the longest continuous series of a firm's valid annual ROE *up to June of year t*. We require a minimum of 5 years of non-missing ROEs.
Dividend_Payer	A dummy equal to one if the firm paid any dividends (CRSP data field DIVAMT>0) over the prior year (*from June of year t-1 to June of year t*)
Profit_Uncertainty_by_Dividend_Payer	*Profit_Uncertainty* × *Dividend_Payer*

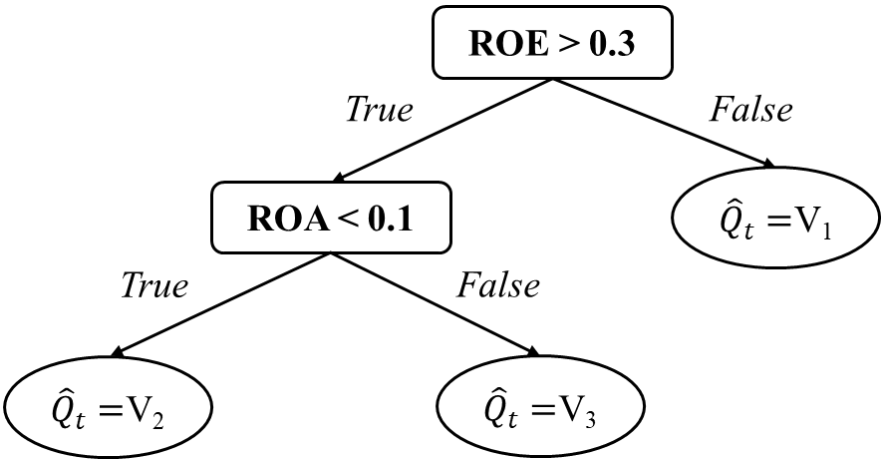
This table shows the definitions of control variables in Equation (8).

Figure 1. Timeline of the Prediction Model



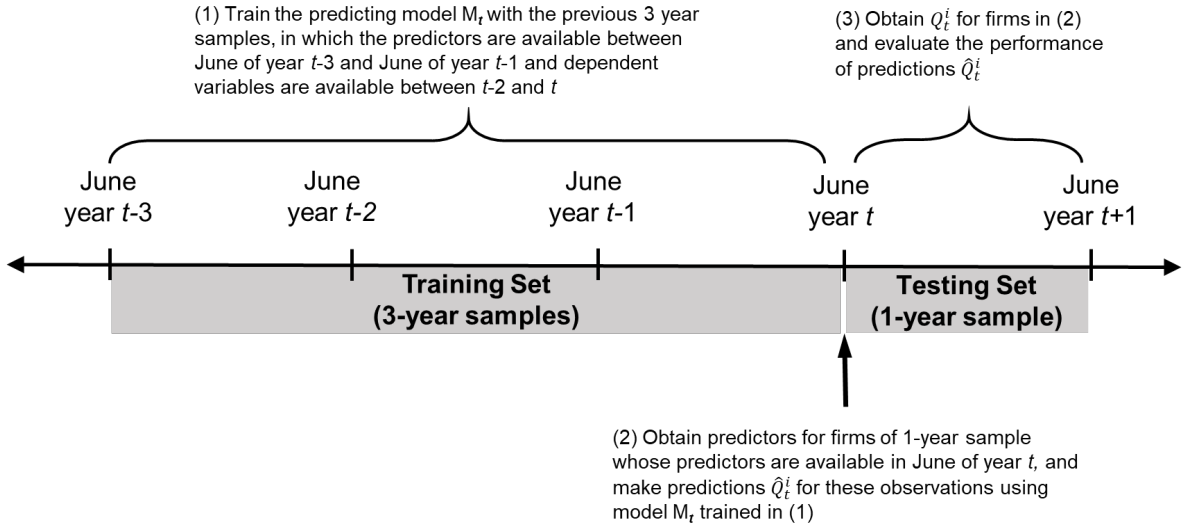
This figure shows the timeline for our firm quality prediction model. Following Hou et al. (2012) and Li and Mohanram (2014), we perform the prediction as of June 30 for each calendar year *t*. To avoid any look-ahead biases, we assume that only financial information for firms with fiscal year ending (FYE) prior to April 1 of calendar year *t* is available on June 30 of calendar year t. We compute $\hat{Q}_t^i$ on June 30 of calendar year *t* and then form trading portfolios on July 1 of calendar year *t*.

Figure 2. An Example of a Decision Tree for Predicting Firm Quality



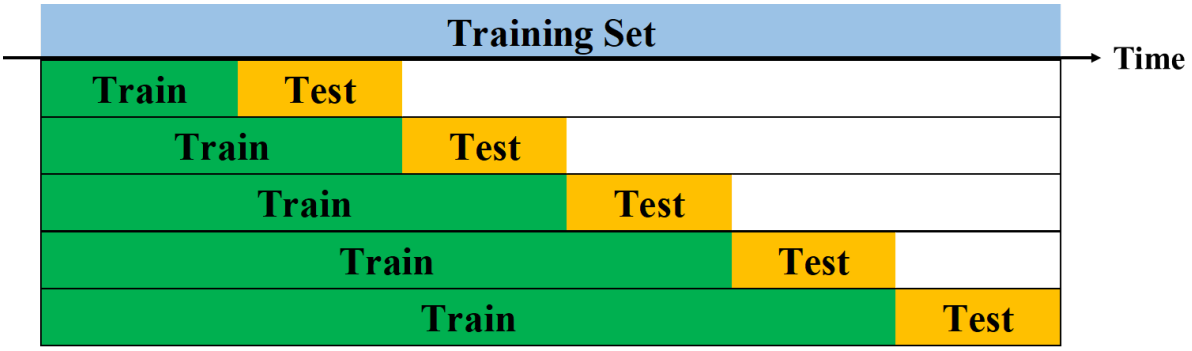
This figure depicts a simple example of a decision tree for how it works in our regression task. For a given observation, the example decision tree makes prediction based on the variables ROE and ROA of the observation: If the ROE of the observation is larger than 0.3, the example decision tree will predict its quality measure as V_1; If not, the decision tree will further check if the observation's ROA is smaller than 0.1, and gives the prediction based on this judgement result. The decision tree can grow deeper when using more independent variable to make predictions.

Figure 3. Timeline of Training and Testing Sets



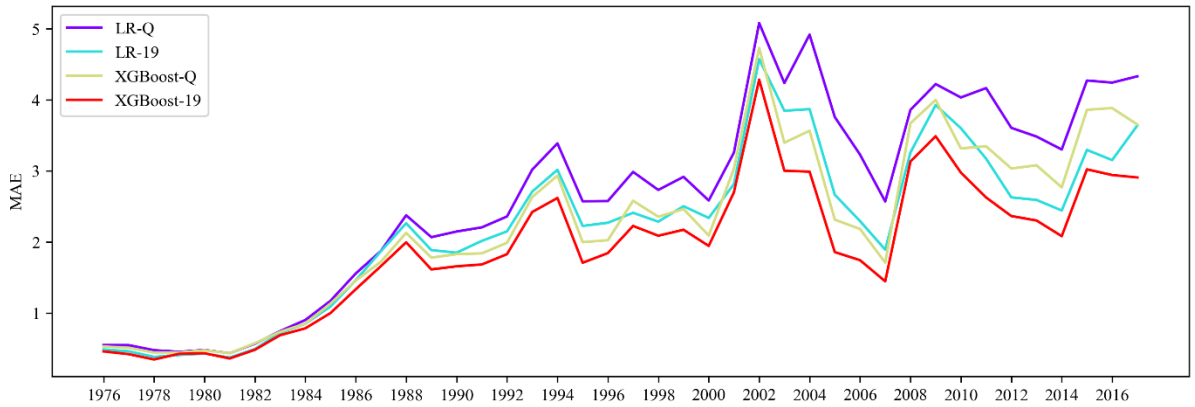
This figure shows the graphical timeline of obtaining the data for training and testing set in a four-year sliding window. For a testing set consisting of observations whose independent variables are available at calendar year *t*, the corresponding training set includes observations whose independent variables are available between calendar year *t*-3 and *t*-1. For samples in the training set, their dependent variables are obtained in June of each calendar year between calendar year *t*-2 and *t*, respectively. Therefore, in June of calendar year *t*, we can train a model using the data of the training set, and then apply the model to make predictions $\hat{Q}_t^i$ for samples of calendar year *t* based on their independent variables that are obtained at the same cross-section. In June of calendar year *t*+1, we can obtain the true quality measure $\hat{Q}_t^i$ and evaluate the performance of our predictions $\hat{Q}_t^i$. By traversing the whole sample period using the sliding window, we are able to make predictions for each cross-section in the sample period and evaluate the prediction performance.

Figure 4. Five-fold Time-Series Cross-validation



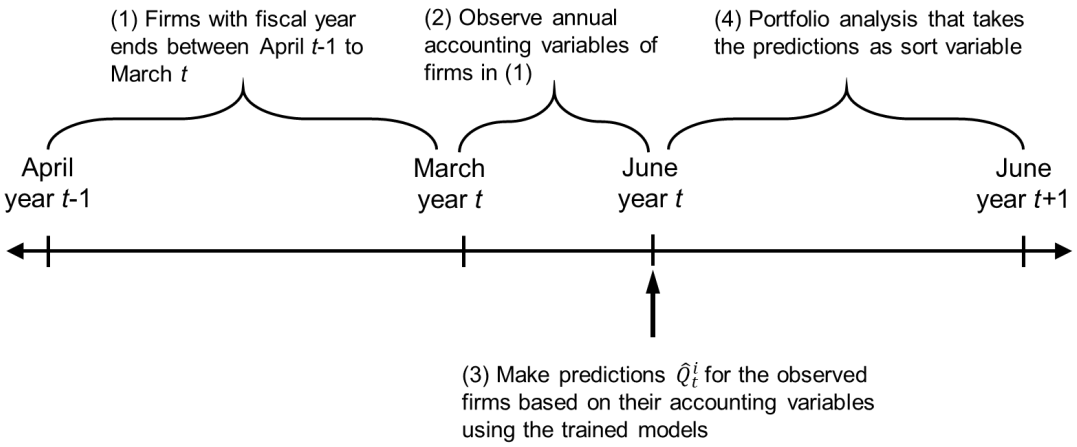
This figure illustrates how to perform time-series cross-validation (Hyndman and Athanasopoulos 2018). Specifically, the observations in a training set are first sorted in chronological order based on their fiscal year end (FYE), and then divided into six equal-size groups (labelled 1 to 6) in the chronological order. Group 1 is the earliest and group 6 the latest in calendar time. The five-fold cross validation is conducted as follows: the first fold takes group 1 as the training set and group 2 as the testing set; the second fold takes the groups 1 and 2 as the training set and group 3 as the testing set; we repeat this procedure for a total of five times. Given multiple alternative values of each hyperparameter, we can obtain a set of hyperparameter combinations in which the elements are the Cartesian product of the value set of each hyperparameter. By performing training and testing on each fold, the hyperparameter combination that achieves the best average performance on the five testing sets is selected. In our study, we use the average of the minimum absolute prediction error as our performance evaluation metric to avoid potential influence of outliers.

Figure 5. Time-varying prediction performance of each prediction model



This figure tabulates the values of MAE for all the models in each test year. It is found that the performance of all models suffers for the years around financial crises (e.g., 1988, 1994, 2002, 2008). More importantly, the performance of our best model XGBoost always dominates the performance of the other models. In addition, the performance of the benchmark model LR-Q is almost always the worst for all years.

Figure 6. Timeline of the Portfolio Analysis



This figure shows the timeline of the portfolio analysis. Our firm quality proxies are developed using publicly available information prior to July 1 of calendar year *t*. Hence, there is no look-ahead biases in the abnormal returns generated from our trading strategy. Following Asness et al. (2019), we hold each hedging portfolio for one year, rebalanced every month using equal weighting.

Table 1. Hyperparameters for XGBoost

Hyperparameters	Definitions	Candidate values
objective	The learning task and the learning objective	reg:pseudohubererror
seed	The seed of random number	42
num_round	Number of trees	500
eta	Learning rate	0.01
max_depth	Maximum depth of a tree	2, 3, 4, 5
min_child_weight	Minimum sum of instance weight (hessian) needed in a child	1, 2
subsample	Subsample ratio of the training instances	0.6, 0.7, 0.8, 0.9
colsample_bytree	Subsample ratio of features when constructing each tree	0.6, 0.7, 0.8, 0.9

This table shows the hyperparameters and the candidate values for the hyperparameter tuning process. We list the standard name of the hyperparameters in XGBoost tutorials and explain their definitions. The first four rows show the parameters that have fixed values and the rest have multiple candidate values. As our dependent variable is continuous, we set the hyperparameter 'objective' as 'reg:pseudohubererror', which means that the learning task is regression and the objective of the model training is to minimize the Pseudo-Huber loss (i.e., a twice differentiable alternative to absolute loss). We use a fixed seed of random number to make sure the training set can be replicated. The number of trees is set as 500 and the learning rate is 0.01. To control overfitting of the model training, we mainly tune the hyperparameters 'max_depth', 'min_child_weight', 'subsample' and 'colsample_bytree'. The former two control the model complexity, and the latter two are used to add randomness to make training robust to noise.

Table 2. Sample Selection

Steps of Sample Selection	The Number of Kept Observations	Sample Period
Merged Compustat/CRSP data	238607	1951-2018
Delete firms that are not in the 49 industries defined in Fama and French (1997)	238199	1951-2018
Delete financial industry firms	199225	1951-2018
Delete observations before 1970	185449	1970-2018
Compute variables	176876	1973-2017
Delete observations with missing independent variables	127830	1973-2017
Delete observations with missing dependent variables	118267	1973-2017
Delete observations with abnormal dependent variable values	113336	1973-2017
Delete duplicated observations of the same firm in one cross-section	113296	1973-2017

This table describes sample selection procedures. Our sample starts from all available common stocks in the merged Compustat/CRSP database at the end of 2018, which consists of 238,607 firm-year observations for the calendar years 1951-2018. We drop firms that are not in the 49 industries defined in Fama and French (1997) because the construction of firm quality requires non-missing industry cost of capital. Considering that many accounting variables are not meaningful for financial firms, we also exclude firms in the financial service industry. We obtain a sample of 199,225 firm years for the period 1951-2018. Many Compustat accounting variables contain missing values. As the missing value problem is very severe in the early years, we restrict the sample period to 1970-2018. As the calculation of certain independent variables requires data for at least four years, our sample is further limited to 176,876 firm years for the period 1973-2018. After filling the missing values of each raw accounting data item from Compustat to the extent possible based on accounting knowledge, we drop all the firm years that still have missing values for any of the accounting variables. This restriction reduces the sample to 127,830 firm years for the period 1973-2018. Similarly, requiring a non-missing dependent variable reduces the sample to 118267 firm years. Following Frankel and Lee (1998), we further drop firm-years with negative or extremely small book value of common equity (referred to as abnormal dependent variable values in the table) for the calculation of dependent variable. Extremely small book value of common equity is defined as firm-years whose book values are smaller than the 1% percentile of all firms in that year. This step reduces the sample to 113,336 firm years. Finally, we drop duplicate firm IDs (PERMNO) for each year due to fiscal year change, and the final sample consists of 113296 firm-years.

Table 3. Out-of-Sample Performance of Quality Prediction

Panel A: Out-of-Sample Performance of Quality Prediction for Different Models		
	MSE	**MAE**
LR-Q	34.654	2.626
LR-19	27.896	2.206
XGBoost-Q	36.535	2.227
XGBoost-19	27.592	1.907
XGBoost-19m	27.654	1.908
Panel B: Comparison of Prediction Performance Using Newey-West Adjusted *t*-test		
	MSE	**MAE**
LR-Q vs. LR-19	3.28**	3.73**
LR-Q vs. XGBoost-Q	-3.34**	3.87**
LR-Q vs. XGBoost-19	3.09**	4.19**
LR-19 vs. XGBoost-Q	-3.41**	-0.34
LR-19 vs. XGBoost-19	0.58	4.49**
XGBoost-Q vs. XGBoost-19	3.31**	3.83**
XGBoost-19 vs. XGBoost-19m	-1.94	-1.07

This table shows the summary statistics (mean and standard deviation) of the out-of-sample performance for the three machine learning models versus the benchmark model over the test period 1976-2018. For each model, Panel A reports the summary statistics for the two performance evaluation metrics, MSE and MAE. Panel B conducts formal statistical tests on the performance difference for each pair of machine learning models. The values in Panel B are *t*-stat. A negative value means that the left model performs better than the right, and vice versa. Double asterisk indicates the difference is significant at 1% level (single asterisk indicates the 5% level) or better for individual tests. The *t*-stat is Newey-West adjusted following Newey and West (1987), and the lag is set to 5.

Table 4. Out-of-Sample Performance of Quality Prediction for Two- and Three- year Expansion of PVRI

Panel A: Out-of-Sample Performance of Quality Prediction for Different Model

	Two-year expansion of PVRI		Three-year expansion of PVRI	
	MSE	**MAE**	**MSE**	**MAE**
LR-Q	47.866	2.952	70.499	3.385
LR-19	41.097	2.660	61.879	3.130
XGBoost-Q	50.019	2.578	73.430	2.973
XGBoost-19	41.092	2.349	63.122	2.782
XGBoost-19m	41.165	2.352	63.258	2.784

Panel B: Comparison of Prediction Performance Using Newey-West Adjusted *t*-test

	Two-year expansion of PVRI		Three-year expansion of PVRI	
	MSE	**MAE**	**MSE**	**MAE**
LR-Q vs. LR-19	3.30**	3.59**	3.34**	3.29**
LR-Q vs. XGBoost-Q	-3.42**	3.49**	-3.41**	3.33**
LR-Q vs. XGBoost-19	3.19**	3.99**	3.21**	3.80**
LR-19 vs. XGBoost-19	0.01	3.77**	-1.26	3.29**
LR-19 vs. XGBoost-Q	-3.38**	1.14	3.44**	-1.61
XGBoost-Q vs. XGBoost-19	3.37**	3.71**	3.42**	3.48**
XGBoost-19 vs. XGBoost-19m	-0.73	-1.25	-2.12*	-1.51

This table shows the performance results for the two-year expansion and three-year expansion of the valuation model. For each model, Panel A reports the summary statistics for the two performance evaluation metrics, MSE and MAE. Panel B conducts formal statistical tests on the performance difference for each pair of machine learning models. The values in Panel B are *t*-stat. A negative value means that the left model performs better than the right, and vice versa. Double asterisk indicates the difference is significant at 1% level (single asterisk indicates the 5% level) or better for individual tests. The *t*-stat is Newey-West adjusted following Newey and West (1987), and the lag is set to 5.

Table 5. Out-of-Sample Performance of Quality Prediction for XGBoost Models Using More Input Variables

Panel A: Out-of-Sample Performance of Quality Prediction for Different Models		
	MSE	**MAE**
XGBoost-19m	27.654	1.908
XGBoost-63m	27.316	1.872
XGBoost-87m	26.887	1.853
XGBoost-318m	26.712	1.858
Panel B: Comparison of Prediction Performance Using Newey-West Adjusted t-test		
	MSE	**MAE**
XGBoost-19m vs. XGBoost-63m	1.90	4.59**
XGBoost-19m vs. XGBoost-87m	3.12**	4.44**
XGBoost-19m vs. XGBoost-318m	3.60**	4.49**
XGBoost-63m vs. XGBoost-87m	3.41**	3.00**
XGBoost-63m vs. XGBoost-318m	2.11*	1.75
XGBoost-87m vs. XGBoost-318m	0.76	-0.81

This table shows the performance results for the XGBoost models using more input variables. For each model, Panel A reports the summary statistics for the two performance evaluation metrics, MSE and MAE. Panel B conducts formal statistical tests on the performance difference for each pair of machine learning models. The values in Panel B are *t*-stat. A negative value means that the left model performs better than the right, and vice versa. Double asterisk indicates the difference is significant at 1% level (single asterisk indicates the 5% level) or better for individual tests. The *t*-stat is Newey-West adjusted following Newey and West (1987), and the lag is set to 5.

Table 6. The Contemporaneous Stock Pricing of Firm Quality

	(1)	(2)	(3)
The true firm quality measure	0.20***		
	[16.28]		
Predicted firm quality from XGBoost-19m		0.19***	
		[17.54]	
Asness' Q score			0.12***
			[9.91]
Firm_size	0.33***	0.32***	0.37***
	[12.49]	[12.10]	[15.71]
One_year_return	0.19***	0.23***	0.24***
	[12.17]	[15.56]	[17.23]
Firm_age	-0.02	-0.02	-0.01
	[-1.39]	[-1.14]	[-0.66]
Profit_Uncertainty	0.24***	0.24***	0.26***
	[16.69]	[16.03]	[21.83]
Dividend_Payer	-0.08***	-0.09***	-0.08***
	[-3.94]	[-4.66]	[-3.94]
Profit_Uncertainty_by_Dividend_Payer	-0.02	-0.02	-0.01
	[-1.11]	[-1.12]	[-0.38]
Adjusted R2	0.43	0.42	0.40
Nobs	42	42	42
Industry FE	Y	Y	Y

This table reports the time series averages of the regression coefficients of model (8). We adjust the standard errors for heteroskedasticity and autocorrelation of five lags (Newey and West 1987). Adjusted R^2 is the time series average of the adjusted R-squared of the cross-sectional regression. T-statistics are shown below the coefficient estimate. 10% statistical significance is indicated by *, 5% statistical significance is indicated by **, and 1% statistical significance is indicated by ***.

Table 7. Risk-Adjusted Return to the Combination of Changes in Firm Quality and Book-to-Market Ratio

Panel A: Change in Asness' Q score relative to the Q-score in year *t*-1 (1976/07 - 2018/06)

Quality Change Group	1 (Low quality change)			2~9			10 (High quality change)			H-L (10 and 10 – 1 and 1)
BM Group	**1 (Expensive)**	**2~9**	**10 (Cheap)**	**1 (Expensive)**	**2~9**	**10 (Cheap)**	**1 (Expensive)**	**2~9**	**10 (Cheap)**	
q^5-factor alpha	0.08	0.38***	1.64***	0.03	0.26***	1.03***	0.19	0.39***	0.92***	0.84***
	[0.27]	[2.65]	[4.69]	[0.20]	[3.01]	[5.03]	[0.81]	[3.23]	[4.13]	[2.67]
Beta	1.13	1.02	0.85	1.09	0.94	0.92	1.17	0.99	0.84	-0.29
Sharpe Ratio	0.21	0.57	0.85	0.31	0.68	0.91	0.31	0.67	0.88	0.88
Information Ratio	0.07	0.66	1.01	0.05	0.74	1.11	0.17	0.70	0.75	0.53
Adjusted R2	0.72	0.89	0.47	0.89	0.94	0.75	0.77	0.89	0.59	0.25
Nobs	24	187	23	188	1494	186	24	186	23	47

Panel B: Change in predicted firm quality based on XGBoost-19m relative to the realized firm quality in year *t*-1 (1976/07 - 2018/06)

q^5-factor alpha	0.22	0.36**	1.23***	0.01	0.22***	0.86***	0.01	0.77***	2.41***	2.18***
	[1.12]	[2.42]	[3.76]	[0.05]	[2.88]	[4.48]	[0.04]	[2.84]	[5.08]	[4.7]
Beta	1.14	0.98	0.9	1.09	0.93	0.9	1.13	1.06	1	-0.15
Sharpe Ratio	0.39	0.51	0.64	0.32	0.7	0.88	0.19	0.57	0.93	0.93
Information Ratio	0.23	0.58	0.74	0.01	0.61	1.02	0.01	0.85	1.09	0.95
Adjusted R2	0.77	0.87	0.54	0.91	0.94	0.76	0.69	0.83	0.45	0.22
Nobs	24	187	23	188	1494	186	24	186	23	47

This table shows the monthly alpha for our hedging portfolios. Deciles are created along the dimension of changes in firm quality (Asness' Q score in year t minus Q score in year *t*-1, or the change in predicted firm quality based on XGBoost-19m relative to the realized firm quality in year *t*-1), and within each decile, deciles of book-to-market ratio are created. For brevity, the results of the middle deciles are condensed by putting them into one portfolio. Following Asness et al (2019), we hold each hedging portfolio for one year, rebalanced every month using equal weighting. Alpha is the intercept in a time-series regression on the whole sample period of monthly excess return. The explanatory variables are the five factors from q^5 factor model including the market factor MKT, the size factor ME, the return on equity factor ROE (i.e., the profitability factor), and the investment factor IA and the expected investment growth factor EG. Returns and alphas are in monthly percentage, t-statistics are shown below the coefficient estimates. 10% statistical significance is indicated by *, 5% statistical significance is indicated by **, and 1% statistical significance is indicated by ***. Beta is the realized loading on the market portfolio in q^5 factor model. Information ratio is equal to the q^5 factor alpha divided by the

standard deviation of the estimated residuals in the time-series regression. Sharpe ratios and information ratios are annualized. Nobs is the number of stocks in the portfolio. The t-stat is Newey-West adjusted (following Newey and West (1987)). The lag is set as 5 following Asness et al (2019).

Table 8. Risk-Adjusted Return to the Combination of Changes in Firm Quality and Book-to-Market Ratio Based on XGBoost-19m for Different Subperiods

Panel A: Subperiod 1 (1976/07 - 1990/06)										
Quality Change Group	**1 (Low quality change)**			**2~9**			**10 (High quality change)**			**H-L (10 and 10 – 1 and 1)**
BM Group	**1 (Expensive)**	**2~9**	**10 (Cheap)**	**1 (Expensive)**	**2~9**	**10 (Cheap)**	**1 (Expensive)**	**2~9**	**10 (Cheap)**	
***q*^{5}-factor alpha**	-0.28	0.11	0.56	-0.14	0.23***	0.42**	-0.71	0.14	1.50**	1.78***
	[-0.88]	[0.67]	[1.1]	[-0.92]	[3.70]	[2.19]	[-1.55]	[0.56]	[2.21]	[2.99]
Beta	1.21	1.06	0.98	1.09	0.95	0.89	1.17	1.06	1.00	-0.21
Sharpe Ratio	0.32	0.42	0.49	0.26	0.63	0.79	0.03	0.50	0.76	0.76
Information Ratio	-0.32	0.24	0.51	-0.35	1.30	0.79	-0.56	0.22	0.91	0.99
Adjusted R2	0.84	0.92	0.63	0.95	0.99	0.87	0.71	0.88	0.49	0.25
Nobs	24	183	23	184	1465	183	24	182	22	46
Panel B: Subperiod 2 (1990/07 - 2004/06)										
***q*^{5}-factor alpha**	0.88**	1.05***	2.37***	0.39	0.39	1.42***	1.20**	2.06***	4.31***	3.43***
	[2.42]	[3.29]	[4.30]	[1.18]	[1.63]	[2.89]	[2.07]	[3.31]	[5.06]	[3.77]
Beta	0.96	0.86	0.99	0.99	0.83	0.66	0.89	0.87	0.66	-0.29
Sharpe Ratio	0.36	0.61	0.92	0.26	0.84	1.22	0.29	0.78	1.29	1.29
Information Ratio	0.91	1.43	1.30	0.57	0.79	1.45	0.71	1.84	1.70	1.34
Adjusted R2	0.77	0.85	0.61	0.86	0.86	0.62	0.68	0.81	0.40	0.21
Nobs	27	207	26	208	1660	207	27	206	26	53
Panel C: Subperiod 3 (2004/07 - 2018/06)										
***q*^{5}-factor alpha**	0.20	0.06	0.68	-0.14	0.23***	1.06***	-0.38	0.16	2.01**	1.81**
	[0.74]	[0.48]	[1.32]	[-1.18]	[2.95]	[4.15]	[-1.20]	[0.65]	[2.43]	[2.2]
Beta	0.97	0.92	0.78	1.05	0.95	0.90	1.11	1.09	1.03	0.05
Sharpe Ratio	0.54	0.52	0.44	0.49	0.65	0.72	0.25	0.39	0.71	0.71
Information Ratio	0.26	0.13	0.38	-0.37	0.91	1.33	-0.32	0.26	0.91	0.78
Adjusted R2	0.75	0.89	0.42	0.93	0.97	0.85	0.73	0.89	0.50	0.20

Nobs	22	170	21	171	1358	169	22	169	21	43

This table shows the monthly alpha for our hedging portfolios based on XGBoost-19m for different subperiods. Deciles are created along the dimension of changes in firm quality (Asness' Q score in year t minus Q score in year *t*-1, or the change in predicted firm quality based on XGBoost-19m relative to the realized firm quality in year *t*-1), and within each decile, deciles of book-to-market ratio are created. For brevity, the results of the middle deciles are condensed by putting them into one portfolio. Following Asness et al (2019), we hold each hedging portfolio for one year, rebalanced every month using equal weighting. Alpha is the intercept in a time-series regression on the whole sample period of monthly excess return. The explanatory variables are the five factors from q^5 factor model including the market factor MKT, the size factor ME, the return on equity factor ROE (i.e., the profitability factor), and the investment factor IA and the expected investment growth factor EG. Returns and alphas are in monthly percentage, t-statistics are shown below the coefficient estimates. 10% statistical significance is indicated by *, 5% statistical significance is indicated by **, and 1% statistical significance is indicated by ***. Beta is the realized loading on the market portfolio in q^5 factor model. Information ratio is equal to the q^5 factor alpha divided by the standard deviation of the estimated residuals in the time-series regression. Sharpe ratios and information ratios are annualized. Nobs is the number of stocks in the portfolio. The t-stat is Newey-West adjusted (following Newey and West (1987)). The lag is set as 5 following Asness et al (2019).

Table 9. Risk-Adjusted Return to the Combination of Changes in Firm Quality and Book-to-Market Ratio Based on XGBoost Models using More Input Variables

Panel A: XGBoost-63m (1976/07 - 2018/06)										
Quality Change Group	**1 (Low quality change)**			**2~9**			**10 (High quality change)**			**H-L (10 and 10 – 1 and 1)**
BM Group	**1 (Expensive)**	**2~9**	**10 (Cheap)**	**1 (Expensive)**	**2~9**	**10 (Cheap)**	**1 (Expensive)**	**2~9**	**10 (Cheap)**	
***q*^{5}-factor alpha**	0.27	0.51***	1.00***	0.01	0.21***	0.92***	0.18	0.67***	2.11***	1.84***
	[1.14]	[3.09]	[3.75]	[0.10]	[2.69]	[4.69]	[0.62]	[2.72]	[4.6]	[4.14]
Beta	1.09	0.98	0.84	1.10	0.94	0.90	1.12	1.06	1.02	-0.07
Sharpe Ratio	0.30	0.52	0.64	0.35	0.70	0.90	0.21	0.58	0.84	0.84
Information Ratio	0.26	0.79	0.66	0.03	0.58	1.06	0.13	0.79	0.99	0.83
Adjusted R2	0.76	0.87	0.53	0.90	0.94	0.75	0.69	0.84	0.47	0.21
Nobs	24	187	23	188	1494	186	24	186	23	47
Panel B: XGBoost-87m (1976/07 - 2018/06)										
***q*^{5}-factor alpha**	0.15	0.40**	1.14***	0.04	0.22***	0.93***	0.15	0.69***	1.91***	1.77***
	[0.63]	[2.57]	[3.68]	[0.30]	[2.85]	[4.64]	[0.49]	[2.70]	[4.33]	[3.95]
Beta	1.14	0.99	0.92	1.08	0.94	0.89	1.14	1.06	1.07	-0.08
Sharpe Ratio	0.27	0.48	0.68	0.35	0.70	0.91	0.19	0.59	0.79	0.79
Information Ratio	0.14	0.63	0.70	0.08	0.61	1.08	0.10	0.81	0.90	0.78
Adjusted R2	0.77	0.88	0.53	0.90	0.94	0.76	0.69	0.84	0.48	0.20
Nobs	24	187	23	188	1494	186	24	186	23	47
Panel C: XGBoost-318m (1976/07 - 2018/06)										
***q*^{5}-factor alpha**	0.13	0.38**	1.24***	0.01	0.23***	0.91***	0.16	0.68***	2.1***	1.97***
	[0.60]	[2.12]	[3.75]	[0.09]	[2.95]	[4.7]	[0.49]	[2.94]	[4.57]	[4.13]
Beta	1.12	1.01	0.87	1.09	0.93	0.89	1.12	1.04	1.07	-0.05
Sharpe Ratio	0.26	0.47	0.58	0.34	0.70	0.91	0.20	0.60	0.85	0.85
Information Ratio	0.13	0.57	0.71	0.02	0.63	1.06	0.11	0.84	0.97	0.86
Adjusted R2	0.79	0.87	0.54	0.90	0.94	0.76	0.68	0.85	0.47	0.18

Nobs	24	187	23	188	1494	186	24	186	23	47

This table shows the monthly alpha for our hedging portfolios based on XGBoost models using more input variables. Deciles are created along the dimension of changes in firm quality (the change in predicted firm quality based on XGBoost-63m, XGBoost-87m or XGBoost-318m relative to the realized firm quality in year *t*-1), and within each decile, deciles of book-to-market ratio are created. For brevity, the results of the middle deciles are condensed by putting them into one portfolio. Following Asness et al (2019), we hold each hedging portfolio for one year, rebalanced every month using equal weighting. Alpha is the intercept in a time-series regression on the whole sample period of monthly excess return. The explanatory variables are the five factors from q^5 factor model including the market factor MKT, the size factor ME, the return on equity factor ROE (i.e., the profitability factor), and the investment factor IA and the expected investment growth factor EG. Returns and alphas are in monthly percentage, t-statistics are shown below the coefficient estimates. 10% statistical significance is indicated by *, 5% statistical significance is indicated by **, and 1% statistical significance is indicated by ***. Beta is the realized loading on the market portfolio in q^5 factor model. Information ratio is equal to the q^5 factor alpha divided by the standard deviation of the estimated residuals in the time-series regression. Sharpe ratios and information ratios are annualized. Nobs is the number of stocks in the portfolio. The t-stat is Newey-West adjusted (following Newey and West (1987)). The lag is set as 5 following Asness et al (2019).