# Fundamental Analysis of Detailed Financial Data: A Machine Learning Approach

Xi Chen
xc13@stern.nyu.edu
Department of Technology, Operations, and Statistics

Yang Ha Cho
yhc479@stern.nyu.edu
Department of Accounting

Yiwei Dou
yd18@stern.nyu.edu
Department of Accounting

Baruch Lev
bil1@stern.nyu.edu
Department of Accounting

Stern School of Business
New York University

April 2021

The paper benefited from the comments of Christian Leuz, an anonymous reviewer, an anonymous associate editor, Aleksander Aleszczyk, Karthik Balakrishnan, Mark Bradshaw, Amy Hutton, Charles M.C. Lee, E. Jin Lee, Becky Lester, Miao Liu, Joshua Livnat, Joseph Piotroski, K. Ramesh, Joshua Ronen, Siew Hong Teoh, Chenqi Zhu, and seminar participants at Boston College, Florida International University, New York University, Rice University, UC Irvine, and Stanford University.

# Fundamental Analysis of Detailed Financial Data: A Machine Learning Approach

**ABSTRACT**

We conduct a fundamental analysis of detailed financial information to predict earnings. Since 2012, all U.S. public companies must tag quantitative amounts in financial statements and footnotes of their 10-K reports using the eXtensible Business Reporting Language (XBRL). Leveraging machine learning methods, we combine the high-dimensional XBRL-tagged financial data into a summary measure for the direction of one-year-ahead earnings changes. The measure shows significant out-of-sample predictive power: the area under the curve ranging from 67.52 to 68.66 percent is significantly higher than that of a random guess, which is 50 percent. Hedge portfolios are formed based on this measure during 2015-2018. The annual size-adjusted returns to the hedge portfolios range from 5.02 to 9.74 percent. These returns survive after accounting for transaction costs and using the five-factor Fama and French (2015) model. Our measure and strategies outperform those of Ou and Penman (1989), who extract the summary measure from 65 accounting variables using logistic regressions. Additional analyses suggest that the outperformance stems from both nonlinear predictor interactions missed by regressions and the use of more detailed financial data.

## 1. Introduction

Corporate financial reports—financial statements as well as footnote disclosures—contain numerous accounting items. Understanding the extent to which these detailed financial data help predict corporate performance is of first-order importance to accounting research and investment practitioners. Documenting predictive power of detailed financial data empirically however poses two challenges. First, the high dimensionality of these data makes a regression approach infeasible. For example, the number of predictors (i.e., financial items) often exceeds the number of observations for model estimation. Traditional regression methods break down in such a scenario. While prior work often turns to a relatively small set of selected financial ratios (e.g., Ou and Penman 1989), the extent of lost information from not using all available data is unclear. Accordingly, the predictive ability of detailed financial information in its entirety is still unknown. Second, despite listed companies' financial reports being publicly available, arranging the entire detailed financial information in a machine-readable format for a large-scale analysis is non-trivial. Commercial data vendors typically collect only part of this information. In this paper, we address these two challenges. By doing so, we conduct a fundamental analysis of comprehensive financial data in predicting future earnings and stock returns.

We address the first challenge by using two machine learning methods: random forests and stochastic gradient boosting, which have recently achieved remarkable success in real-world applications (Zhou 2012; Mullainathan and Spiess 2017; Liu 2021). Both methods, built on ensemble learning, combine a large set of estimators from decision trees. Unlike regressions, a crucial feature of these methods is their ability to estimate models where the number of predictors is greater than the number of observations. More importantly, theoretical literature offers little guidance for the selection of key financial variables and the functional forms in

financial statement analysis. The high-dimensional predictor sets may enter non-linearly with various interactions. The machine learning algorithms, on the other hand, are explicitly designed to accommodate complex associations and cast a wide net in their specification search. These algorithms are also specialized for prediction tasks as they offer high out-of-sample predictive performance by using the "regularization" (e.g., tuning a parameter such as the number of decision trees in random forests) for model selection and mitigation of overfitting.

To overcome the second challenge of arranging detailed information in a machine-readable format, we take advantage of financial reports filed in eXtensible Business Reporting Language (XBRL) format. XBRL is an extensible markup language comprised of a standard list of tags to describe business and financial information. It provides a means to convert the information from human-readable formats (e.g., paper, PDF, HTML) to a machine-readable format, comparable to the shift from paper maps to digital maps. Since 2012, all U.S. public companies must enclose quantitative amounts in financial statements and footnotes of their 10-K reports with XBRL tags (see Appendix A for two examples). As such, the XBRL-tagged data include all detailed financial information in financial reports (e.g., a line item breakdown and footnote disclosures). They are standardized and free of charge, available from companies' and the U.S. Securities and Exchange Commission (SEC)'s Websites. The XBRL-tagged financial data are point-in-time: any revision will be captured in a separate XBRL file; hence there is no backfilling or updating of the original filings. This feature avoids hindsight bias and enables the construction of implementable trading strategies. Nevertheless, XBRL documents need not be audited, and errors in these documents expose the filers to limited liability within two years after the initial adoption (SEC 2009). When the standard taxonomy does not provide a tag for a financial element, a company can create a custom tag called an "extension." Research finds

errors and unnecessary extensions in XBRL documents, as semantically equivalent tags already exist in the taxonomy, particularly in early adoption years (Debreceny et al. 2010, 2011).

Our sample is comprised of over 8,000 XBRL filings from 2012 to 2018. These filings contain more than 4,000 unique financial items in standard tags common throughout our sample period. We take all the items for the current and lagged years, divide them by total assets, and compute the annual percentage changes, which yield over 12,000 explanatory variables (i.e., 4,000 × 3 for current values, lagged values, and percentage changes). To extract a summary measure of fundamentals from the detailed financial data, we employ the machine learning models to predict the direction of one-year-ahead earnings changes.

As proposed by Ou and Penman (1989), the advantages of using the direction of one-year-ahead earnings changes are three-folds. First, conceptually, this measure can better capture fundamentals than future stock returns, which reflect both fundamentals and compensation for risk. Second, the one-year horizon disregards information about earnings more than one year ahead and thus yields conservative results toward finding no abnormal trading profits. Third, extant evidence suggests that earnings forecasts based on firm characteristics are not substantially more accurate than forecasts obtained from the random-walk model (Gerakos and Gramacy 2013; Li and Mohanram 2014). In other words, it is difficult to predict the amount of changes in earnings. As such, examining a simpler task of predicting the direction of earnings changes can be worthwhile. While losing some information, the binary specification helps mitigate the concern about low out-of-sample performance from predicting the amount of earnings changes.[1] Using this measure also permits a direct comparison between our models and those in Ou and Penman (1989).

[1] The approach of predicting the direction of earnings changes originated from Freeman et al. (1982) and was adopted by Ou and Penman (1989) and Ou (1990). Freeman et al. (1982, 643) argue that the variability in earnings

For each year in the test period, from 2015 to 2018, we use the second and third preceding years as the machine learning training period to estimate models and the preceding year as the validation period to select the model that yields the best out-of-sample performance. The chosen model is then applied to the year in the test period to produce the summary measure *Pr*, which characterizes the probability of an increase in next year's earnings. We also compare our results with Ou and Penman (1989) by constructing a summary measure from estimating logistic regressions using their 65 financial variables. Ou and Penman's (1989) method serves here as a prominent example of a regression model with expert-selected financial ratios.[2]

We find significant out-of-sample predictability of the machine learning models using the detailed financial data concerning the direction of the next year's earnings changes. The area under the Receiver Operating Characteristics (ROC) curve (AUC) in the test period ranges from 67.52 to 68.66 percent. To understand the predictive power sources, we estimate each variable's importance by computing the decrease in the AUC when that variable is randomly shuffled (Breiman 2001). The majority of the top 10 most important variables pertain to earnings components (e.g., operating income and earnings per share), suggesting that earnings are still leading indicators for valuation among financial numbers. We also classify the variables into six groups (the five financial statements and footnotes) and find that in aggregate, footnote disclosures contribute the most to our models' predictive power, followed by the balance sheet,

changes is too large to be compared to the variability in expected earnings changes conditional on explanatory variables. The substantial noise relative to the amount of data that is typically available makes it difficult to reject the random-walk hypothesis. They propose to reduce the variability in earnings changes by transforming the amount to the direction of earnings changes. While focusing on the direction of earnings changes, we examine predictability of the machine learning methods concerning the amount of earnings changes in Section 5.6.

[2] We choose Ou and Penman (1989) since it is a highly influential study and establishes a comprehensive model of predicting an earnings increase. It won the 1991 AAA Notable Contributions to Accounting Literature Award and was identified as the 11th most cited article during 1976-1993 by Brown (1996). It was cited by 1398 (283) articles on Google Scholar (Web of Science) as of April 1, 2021. Nevertheless, we examine analyst earnings forecast as another example to evaluate our models in Section 6.

income statement, and cash flow statement. Comprehensive income statement and shareholders' equity statement contribute the least. Among footnote disclosures, the top 10 most important variable list is dominated by tax-related items (e.g., valuation allowance for deferred tax assets), consistent with tax items carrying important information about future taxable income (Miller and Skinner 1998; Lev and Nissim 2004; Hanlon 2005; Thomas and Zhang 2011). We also observe meaningful nonlinear and interaction effects of preditors.

We form hedge portfolios three months after the fiscal-year end based on the machine learning summary measure *Pr* and hold them for 12 months. The size-adjusted returns to the hedge portfolios range from 5.02 to 9.74 percent. The results persist after accounting for transaction costs, using alternative earnings measures, and excluding microcaps. The returns are concentrated in long positions and thus are unlikely explained by limits to arbitrage arising from short-sell constraints. Inconsistent with risk-based explanations, the excess returns are robust to controlling for exposure to Fama and French's (2015) five risk-factors and using industry-adjusted returns.

Our summary measures and trading strategies significantly outperform Ou and Penman's (1989), which exhibit only an AUC of 61.69 percent and size-adjusted returns of 2.1 to 2.7 percent in the test period. We investigate the source of this superior performance by applying the same machine learning methods to Ou and Penman's 65 financial variables. The summary measures and trading strategies from these methods with the 65 variables significantly outperform the original Ou and Penman's and marginally underperform those we build using both machine learning and detailed financial data. The results suggest that our models' superior performance stems from primarily nonlinear predictor interactions in machine learning, which

are missed by regressions, and secondarily the use of more detailed financial information in XBRL documents.

We suspect that data quality issues offset the richness of detailed financial data in XBRL documents and thus conduct two sets of additional analyses to address these issues. First, we use Compustat as an alternative source of detailed financial information to XBRL. Compared with XBRL-tagged financial data, Compustat has the advantage of more extensive standardized adjustments to improve data quality and the disadvantage of less detailed coverage of financial information. We continue to find robust predictive power of detailed financial data from Compustat similar to XBRL-tagged data under machine learning. Thus, the influence of noise in XBRL documents (relative to Compustat) is on par with their additional financial details. Second, we find better predictive performance in more recent years than in the early period, likely due to low-quality XBRL-tagged financial data in the early years. We also partition the sample based on firm-level data quality and observe worse predictive performance in the subsample with low data quality. The results suggest that data quality issues reduce the usefulness of detailed financial data in XBRL documents.

Finally, we use analysts' forecasts to evaluate our machine learning methodology. We find that our summary measure *Pr* outperforms analysts' earnings forecasts in predicting the direction of earnings changes and is positively associated with analysts' forecast errors. It appears that these professional financial report users have not fully incorporated the detailed information in financial reports into their forecasts.

Our study makes three contributions. First and foremost, we contribute to the fundamental analysis literature (e.g., Green et al. 2013, 2017) by applying machine learning algorithms to a large set of detailed financial data. Analyzing the detailed data is increasingly

important since extant research examines only a handful of summary financial items (e.g., past earnings and book value of equity) and reports a growing dissatisfaction with the relevance of those items to investors' decisions (Ramesh and Thiagarajan 1996; Lev and Zarowin 1999; Core et al. 2003; Balachandran and Mohanram 2011; Lev 2018). Two recent papers employ machine learning methods to forecast future earnings using a small group of financial statement variables (Anand et al. 2019; Hunt et al. 2019). They do not examine and thus cannot speak to the usefulness of detailed financial data to investors, which we address in this study.

Second, we add to the XBRL literature. The SEC (2009) commented that the XBRL format of financial reports could "improve its usefulness to investors. In this format, financial statement information could be downloaded directly into spreadsheets, analyzed in a variety of ways using commercial off-the-shelf software, and used within investment models in other software formats." Despite the stated goal, the usefulness of XBRL-tagged financial data to investors remains an open question for an issue important to regulators, practitioners, and academics.[3] Existing literature finds that the adoption of XBRL influences capital market outcomes (Blankespoor et al. 2014; Dong et al. 2016; Bhattacharya et al. 2018; Kim et al. 2019a) and corporate reporting decisions (Blankespoor 2019; Kim et al. 2019b). These studies assume that XBRL-tagged financial data contain useful fundamental signals for investors. This assumption, however, is challenged by research documenting errors and unnecessary extensions in XBRL filings (Debreceny et al. 2010, 2011) and the associated adverse consequences in the

[3] Richardson et al. (2010, 446) call for more research on the usefulness of XBRL-tagged financial data: "The development and US adoption of eXtensible Business Reporting Language (XBRL)… means that users now have substantially more information in machine readable form to conduct large-scale archival analyses for the usefulness of that information for forecasting purposes. The set of information contained in financial reports is too detailed to list, but we expect to see research efforts utilizing this information to be worthwhile."

capital markets (Li and Nwaeze 2015, 2018; Kirk et al. 2016).[4] Our fundamental analysis provides direct evidence indicating that XBRL filings still inform investors' forecasts and investment decisions despite the data quality issues.

Finally, an emerging line of research uses machine learning algorithms in accounting and finance research. Several studies use these algorithms to detect accounting fraud or restatements (Cecchini et al. 2010; Perols 2011; Bao et al. 2020; Bertomeu et al. 2020). Barth et al. (2018) examine the value relevance of accounting numbers using decision trees. Ding et al. (2020) employ machine learning to improve reserve estimates in the insurance industry. Researchers also apply machine learning to refine the measurement of expected stock returns (Freyberger et al. 2020; Gu et al. 2020) and to extract information from 10-K textual disclosure (Li 2010; Frankel et al. 2016; Dyer et al. 2017; Cohen et al. 2020). We demonstrate that machine learning can help advance one of the most widely studied areas in research and practice—fundamental analysis of quantitative information in financial reports.

## 2. Background

### *2.1. Machine Learning Using Decision Trees*

We use two widely accepted machine learning methods based on decision trees. Decision trees are a popular statistical learning approach for incorporating nonlinearities and interactions. Unlike regressions, trees are built nonparametrically and designed to group observations with similar predictors. The average of the outcome variable within each group forms the forecast (i.e., the predicted value). The tree "grows" in a sequence of steps. At each step, the sample

[4] Another line of research uses tag characteristics to capture financial reporting complexity and comparability (Scherr and Ditter 2017; Hoitash and Hoitash 2018; Hoitash et al. 2018). See Perdana et al. (2015) for a survey of XBRL research.

leftover from the preceding step is split based on one predictor variable. Typically, the algorithm will try every possible cutoff for each predictor and choose the split that minimizes forecast errors ("impurity") before the next step. The split stops when a further partition cannot reduce forecast errors, or a tree attribute (e.g., tree depth $L$ or the minimum number of elements in a group $b$) reaches a prespecified threshold that can be selected adaptively using a validation sample (Hastie et al. 2009; Varian 2014).

Figure 1 presents an example with two predictors, "EPS" and "Lev" (i.e., earnings per share and leverage), to forecast the direction of earnings changes. Suppose the tree in the left panel is the final output. It describes how each observation is assigned to a group based on its predictor value. A blue box ("a node") represents a split, and a green box ("a leaf") indicates a final partition. First, the sample is sorted on EPS. Observations with EPS above the breakpoint of 0.5 are assigned to Group 1. Those with EPS below 0.5 are then further sorted by Lev: observations with Lev below 0.7 go to Group 2, while those with Lev above 0.7 are assigned to Group 3. The right panel of Figure 1 shows how the space of "EPS" and "Lev" is partitioned by this tree model. Finally, the forecast for observations in each partition is the simple average of the outcome variable among observations in that partition. We can recast the forecasts of the tree as a linear function: $\hat{y} = \beta_1 1_{\{EPS>0.5\}} + \beta_2 1_{\{EPS\leq 0.5\}} 1_{\{Lev<0.7\}} + \beta_3 1_{\{EPS\leq 0.5\}} 1_{\{Lev\geq 0.7\}}$, where $\beta_i$ denotes the mean of the outcome variable for group $i$ and $1_{\{.\}}$ is set to one when the curly bracket statement is true, and zero otherwise (Mullainathan and Spiess 2017).

The advantages of a decision tree are four-fold. First, while considering all explanatory variables, it uses only one predictor for each split and generates forecasts nonparametrically. As a result, there is no need to require a sufficient number of observations relative to the number of predictors necessary for traditional regression analysis. Second, a decision tree is invariant to

monotonic transformations of predictors. Third, it can approximate a high degree of nonlinearities. Fourth, a tree of depth $L$ allows $L-1$ way interactions. The flexibility, however, also makes decision trees prone to overfit and thus calls for regularization. We consider two tree regularizers: random forests and stochastic gradient boosting. Both combine forecasts from many different trees into a single forecast (an "ensemble learning" approach).

Random forests use two procedures to regularize decision trees. First, in the bootstrap aggregation procedure, also known as "bagging" (Breiman 2001), a tree is grown based on each of $m$ different bootstrap samples of the data, as shown in Figure 2. For a given observation, there are $m$ predictions, and the final forecast is the simple average of the $m$ predictions. Trees tend to overfit the individual bootstrap samples, which makes their individual predictions ineffective. Averaging over $m$ predictions reduces this ineffectiveness (i.e., variance in the predicted model) and enhances the predictive performance. Second, if there is a dominant predictor in the data, then most of the bagged trees will split on this predictor at a low level, leading to a significant correlation among their ultimate forecasts. The "dropout" procedure decorrelates trees by considering only a random subset of predictors ($k$ variables) for splitting in each tree. As a result, the dominant predictor may not be considered for some trees. The decreased correlation among predictors can further improve the variance reduction and mitigate the issue of overfitting.

Unlike random forests, which grow trees independently, stochastic gradient boosting builds a tree based on the previous tree's forecast errors ("boosting"), as shown in Figure 3. It starts by averaging the outcome variable as an initial prediction ($F_0(x)$), which is a weak prediction. It then fits a shallow tree (e.g., with depth $L$=1) to the residuals from the initial prediction ($r_0 = y - F_0(x)$). The fitted value is shrunken by a factor $\rho \in (0,1)$ (i.e., the learning rate) to help prevent the model from overfitting the residuals and is added to the initial prediction

$F_1(x) = F_0(x) + \rho \times \hat{r_0}$ to form an ensemble prediction. Then the next tree with the same shallow depth $L$ is used to fit the residuals from the previous prediction. This procedure is repeated $m$ times, and the output of this additive model of shallow trees is the final ensemble prediction. The "stochastic" procedure uses a random sample in each iteration to decorrelate estimates at different iterations. Friedman (2002) shows that this procedure effectively reduces the variance of the combined model.

*2.2. Detailed Financial Accounting Data in XBRL Format*

The SEC mandate of 2009 ("Interactive Data to Improve Financial Reporting") required public companies to provide their financial reports in the XBRL format by submitting them to the SEC and posting them on their corporate Websites.[5] The XBRL format disclosure is in addition to disclosure in the traditional electronic filing formats of ASCII or HTML (see Appendix A). The requirements begin for the first quarterly report for a period ending after a) June 15, 2009 for large accelerated filers with a public equity float over $5 billion, b) June 15, 2010 for other large accelerated filers (with a public equity float over $700 million), and c) June 15, 2011 for all remaining filers. In the first year of XBRL filings, companies must tag each quantitative item on the face of financial statements and each footnote as a block. In the subsequent filing years, companies must also tag the detailed quantitative disclosures within the footnotes.[6] The mandate requires filers to completely align their XBRL report to the traditional ASCII or HTML report (SEC 2009). As a result, a restated financial statement (due to errors or

[5] In 2005, the SEC established a voluntary XBRL filing program to prepare companies for the submission of XBRL filings. Through April 2008, over 75 companies have filed in the XBRL format. See Bartley et al. (2011), Efendi et al. (2016), and Hsieh and Bedard (2018) for studies on the voluntary filing program.

[6] The tagging requirement is exempt for a few types of quantitative values in footnotes, such as those in "the $1.99 pancake special," "1% fat milk," and "drilling 700 feet" (see https://www.sec.gov/corpfin/interactive-data-cdi; Question 146.16).

changes in reporting practices) does not change the original XBRL document. It will be reported in a subsequent filing (e.g., a 10-K/A), for which there is another XBRL document. This point-in-time feature avoids issues related to data backfilling in capital market research.

XBRL U.S., a non-profit organization (a spinoff from AICPA), under contract with the SEC, created the first U.S. GAAP taxonomy in 2008. Like a dictionary, the taxonomy includes a standard list of tags for financial statement items and associated contextual information for software to recognize and process without human intervention. The contextual information includes definitions, authoritative references to U.S. GAAP/SEC regulations, and calculation relationships with other tags (e.g., Accounts Receivable, Net = Accounts Receivable, Gross – Allowance for Doubtful Accounts). The FASB took over the maintenance of the taxonomy from XBRL U.S. after the SEC mandate of 2009 and updated it every year since 2011. The annual update occurs for reasons such as changes in accounting standards, technical corrections, and actual use of tags.

Preparers must tag the quantitative items in the financial reports with the appropriate elements from the standard list. Appendix A provides two examples. In the first example, the amount of cash and cash equivalents on the balance sheet is tagged by "CashAndCashEquivalentsAtCarryingValue" in the XBRL document. The opening tag also contains contextual information about the taxonomy ("us-gaap"), the unit ("usd"), the period ("AsOf29Dec2012"), and the decimal points for presentation ("-3" for in thousands). In the second example, the amount of work in process inventory, as disclosed in a footnote, is tagged by "InventoryWorkInProcess." When there is no appropriate tag in the standard list for a financial concept, a company can create a company-specific tag, called an "extension." The mandate does not require companies to obtain assurance on the XBRL filings or involve third

parties, such as auditors or consultants.[7] The XBRL documents submitted within 24 months since the initial adoption are protected from liability for failure to comply with the tagging requirements (SEC 2009).

While the mandate is intended to improve financial reports' usefulness, research documents data quality issues with the XBRL-tagged data in early adoption years. Debreceny et al. (2010) find that one quarter of the XBRL filings by the initial 400 large companies in the first round of submissions had errors such as misuse of debit/credit, missing values in calculation relationships, and wrong values. Debreceny et al. (2011) take a close look at extensions in XBRL filings of 67 large accelerated filers in the first round of submission. They find that 41 percent of them are unnecessary as appropriate tags already exist in the taxonomy, likely due to premature search in the taxonomy or inadequate understanding of the tagging structure. The errors and unnecessary extensions make it difficult to effectively use the XBRL-tagged financial data (Harris and Morsfield 2012).

Despite the complaints, the SEC, XBRL U.S., and third parties continue to invest in improving the data quality. The SEC periodically issues staff observations, updates to filer practices, and even "Dear CFO" letters on XBRL quality.[8] Michael Willis, the assistant director of the SEC Office of Structured Disclosure, states that the commission is focusing on data-driven regulation, developing data quality tools, and working with the FASB on U.S. GAAP taxonomy enhancements.[9] The Data Quality Committee of XBRL U.S. sets guidance and validation rules to

[7] Plumlee and Plumlee (2008) and Boritz and No (2009) discuss the potential challenges of XBRL documents' assurance.

[8] For example, in July 2014, the SEC Division of Corporation Finance sent letters to certain companies regarding the requirement to include calculation relationships in the XBRL filings (https://www.sec.gov/divisions/corpfin/guidance/xbrl-calculation-0714.htm).

[9] See "SEC's Increasingly Sophisticated Use of XBRL-Tagged Data" at https://www.undergrad.haslam.utk.edu/sites/default/files/files/SECs_Increasingly_Sophisticated_Use_of_XBRL_Tagged_Data.pdf.

prevent or detect inconsistencies or errors in XBRL documents. The committee also collects and publishes real-time errors in XBRL filings.[10] Third-party filing service companies such as XBRL Cloud also monitor for data quality issues in XBRL filings.[11] Using a sample of over 4000 XBRL filings from 2009 to 2010, Du et al. (2013) find a reduction in the number of errors per filing. Blankespoor (2019) computes the number of unique user-day-filing downloads of XBRL filings from the SEC EDGAR Website by year. She finds that the number rises from about 1 million in 2012 to 6 million in 2014, suggesting an increasing demand for XBRL-tagged financial data.

## 3. Data and Methodology

### *3.1. Data*

Table 1 shows our sample selection. We first obtain XBRL 10-K and 10-K/A submissions between June 15, 2012 and March 31, 2018 from the SEC Website.[12] To take advantage of detailed footnote disclosures in XBRL format, we restrict our sample to submissions with a reporting period ending on or after June 15, 2012. After merging them with pro forma earnings from I/B/E/S, we obtain 10,073 submissions.[13] We require that these companies have share price data from CRSP, yielding 8,381 submissions. Requiring non-zero

[10] The errors can be found at https://xbrl.us/data-quality/filing-results/.
[11] See https://edgardashboard.xbrlcloud.com/edgar-dashboard/.
[12] Starting from 2014, the SEC parses all the XBRL documents and puts the XBRL-tagged items in relational databases, available for bulk download at https://www.sec.gov/dera/data/financial-statement-data-sets.html. We examine annual reports for two reasons. First, many disclosures are not required for quarterly reports, making the fourth-quarter data incomparable to those in the previous three. For example, the Statement of Stockholders' Equity was not required in 10-Q filings prior to 2018. Second, this design facilitates the comparison between our study with Ou and Penman (1989).
[13] When we use US GAAP earnings per share to calculate earnings changes and do not require pro forma earnings, the final sample size increases. Our inferences are unchanged by using this sample but become weaker (see Online Appendix Table A1), consistent with US GAAP earnings being less informative about fundamentals than pro forma earnings (Bentley et al. 2018; Bradshaw et al. 2018).

total assets from the XBRL documents leads to a sample of 8,358 submissions. We leverage the point-in-time nature of XBRL submissions by retaining only the most recent financial data as of the portfolio formation date, resulting in a sample of 8,149 submissions.[14] Panel A and Panel B of Table 2 report the number of XBRL submissions by calendar period and by industry, respectively.[15] As expected, there are only 119 submissions of XBRL documents for 10-K filings in 2012 as the detailed footnote tagging for all firms is available only after June 15, 2012.

A submission contains both numerical and contextual data. Retaining only the numerical data, we obtain 167,136 unique tag names (for both custom and standard tags) from the 8,149 submissions. Figure 4a shows a histogram by the number of unique tag names. More than 30 percent of submissions use 250 to 300 unique tags, and an average submission uses 284 unique tags. For each submission, we divide the number of unique custom tags by the number of unique tags (i.e., the proportion of custom tags) and plot a histogram by this variable in Figure 4b. For about 30 percent of submissions, 15 to 20 percent of unique tags are extensions, and the average proportion of extensions is 15.5 percent. Some standard tags are deprecated, and some are added over the years due to changes in accounting standards. We identify uncommon standard tags as those that have not been used at least once in each year of our sample period 2012-2018.[16] Figure 4c presents a histogram by the proportion of uncommon standard tags (i.e., the number of unique uncommon standard tags divided by the number of unique tags). Close to 40 percent of submissions contain 2 to 4 percent uncommon standard tags, and the average proportion of these

[14] We keep only XBRL documents filed before the portfolio formation date. If a company has an XBRL 10-K submission and an XBRL 10-K/A submission to revise financial statements (but not footnotes) before the portfolio formation date, we merge the two submissions by using the revised financial statement items from the 10-K/A and the footnote items from the 10-K.
[15] We include the financial industry (banking, insurance, real estate, trading) to fully explore investment space. Nevertheless, excluding firms in this industry does not alter our inferences.
[16] For example, a standard tag "UnrecognizedTaxBenefitsResultingInNetOperatingLossCarryforward" was deprecated in the 2014 U.S. GAAP taxonomy as ASU 2013-11 about income taxes became effective in 2014.

tags is 4.5 percent, suggesting no major changes to standard tags. As our prediction analysis requires predictors to be populated across firms, we exclude all custom and uncommon standard tags, yielding 4,627 unique common standard tags.[17] Figure 4d shows a histogram by the number of these tags.

In some cases, identical tags are used to describe financial data for a co-registrant, for example, a guarantor subsidiary. We retain only the consolidated data. Also, some tags are used with dimensions (e.g., for segment reporting). Although the U.S. GAAP taxonomy provides many standard dimensions, SEC (2016) reports that 50 percent of filers use custom dimensions, which significantly compromises dimensional data comparability. As such, we discard disaggregate items tagged with dimensions. Companies use identical tags in a submission to refer to items of different reporting periods. For instance, multiple items identically tagged as "NetIncomeLoss" are found spanning different reporting periods such as current and prior years. For each of the 4,627 tags, we select current and prior fiscal year data and compute the percentage changes, which creates 13,881 predictors. Then for predictors with missing values, we fill in zeros.[18]

The FASB maps the tags in each U.S. GAAP taxonomy to financial statement categories. The map is "organized to roughly correspond to the arrangement of elements in the order in which they might be found in a financial statement" (FASB 2018). Using this map, we classify

[17] The drastic drop in the number of unique tag names (from 167,136 to 4,627) given the proportion of custom tags is due to the firm-specific nature of custom tags. For example, suppose there are 1000 documents; each document contains 200 standard tags and 50 custom tags that other firms never use. The average proportion of custom tags across the 1000 documents is 20% (=50/250), but custom tags account for 50,000 (= 50 × 1000) out of 50,200 (= 200 + 50 × 1000) unique tags.

[18] Creating an indicator variable for missing values in each predictor, which will double the number of predictors, does not alter our inferences. Dropping the percentage change predictors does not affect our inferences (See Online Appendix Table A2). Also, adding indicators for the Fama and French 30 industries to the models does not affect our inferences. This is unsurprising as many items unique to certain industries (e.g., "CapitalizedSoftwareDevelopmentCostsForSoftwareSoldToCustomers") have already captured the industry effects.

the predictors into six categories: balance sheet, income statement, cash flow statement, comprehensive income statement, shareholders' equity statement, and footnote disclosures. A tag may be associated with both a financial statement and footnote disclosures (e.g., "InventoryNet"), as a company refers to a financial statement item in a footnote when disclosing more information about that item. We classify the tag into the corresponding financial statement. This procedure allows us to classify 4,503 of 4,627 tags. The remaining 124 tags are mapped to multiple financial statements. We manually assign them to the statement with a more natural fit (see Online Appendix Table A3). Panel A of Table 3 shows that a substantial portion of the predictors belongs to footnotes. Panels B to G list the top 10 most populated (i.e., non-zero) current predictors by financial statement category and present descriptive statistics for the predictor values across the 8,149 submissions. Finally, we scale the current and lagged predictors by total assets (except for total assets itself and items on a per-share basis).

### *3.2. Methodology*

***Earnings.*** We use machine learning methods to predict the direction of one-year-ahead earnings changes. Recent studies demonstrate that earnings used by analysts are of higher quality and more value-relevant to investors, relative to GAAP earnings and non-GAAP earnings reported by managers (Bentley et al. 2018; Bradshaw et al. 2018). As such, we use the annual change in I/B/E/S-reported earnings per share as the outcome variable. We check the robustness of our results using two alternative measures (ROE and EBIT) in Section 5.1. Following Ou and Penman (1989), we adjust for the firm-specific trend by subtracting the average change in EPS over the past four years from the current EPS changes. This procedure helps mitigate the concern

that earnings increases tend to outnumber earnings decreases, and some earnings changes are anticipated due to drift. An earnings increase/decrease is coded after taking out the drift term.[19]

***Parameters.*** Table 4 shows the parameters of the two machine learning methods. The setting of these parameters follows standard practice in machine learning. In random forests, the dropout convention is to randomly select $k = \sqrt{p}$ variables for consideration in each tree, where $p$ is the number of predictors (Breiman 2001). As such, we choose the integers between 110 and 120 for this dropout procedure. We allow the machine to grow 500 to 2000 trees with an increment of 100 and bootstrap 50 percent of the sample for each tree. The minimum number of observations in a leaf (i.e., terminal node) are integers from 1 to 4. For stochastic gradient boosting, the machine can grow 500 to 2000 trees with an increment of 100 with three possible learning rates (0.005, 0.01, and 0.05). We randomly pick 50 percent of the sample to estimate each tree. The early stopping criteria for gradient boosting are typically stricter (than random forests) as the idea is to chain a series of weak learners to mitigate overfitting. As such, we set the tree depth to 1 to 4 and set the minimum number of observations in a leaf to 10.

***Sample splitting.*** In machine learning, the data are typically split into training, validation, and testing samples. Models are estimated in the training sample, selected in the validation sample, and then applied to the test sample. We use a rolling sample splitting scheme, in which the training and validation samples gradually shift forward in time, but the number of years in each sample is held constant. As shown in Figure 5, for each year in the test period from 2015 to

[19] An alternative way to account for anticipation due to drift is comparing actual earnings in fiscal year $t + 1$ with the consensus analyst forecast issued in the month following the earnings release for fiscal year $t$. In other words, one can use the sign of analysts' forecast errors to proxy for earnings changes' direction. However, if analysts incorporate financial information more than the drift into their forecasts, the predictive power of our explanatory variables will deteriorate. To make the predictability of detailed financial data independent of analysts' ability and to make our models comparable to Ou and Penman's (1989), we do not adopt this alternative way in primary analyses but report the results of using this alternative way in Online Appendix Table A4.

2018 (green), the models are trained in the second and third preceding years (blue) and validated in the preceding year (yellow) to tune the parameters as shown in Table 4. This rolling sample splitting scheme has the benefit of using more recent information for prediction.

***Excess Returns.*** Once we apply the estimated model to the year in the test period, we obtain the summary measure $\widehat{Pr}$. A hedge portfolio is then formed based on this measure. Figure 6 shows the timeline of the trading strategy. For each stock in the sample, it is assigned to a long (short) position three months after its fiscal year-end, when $\widehat{Pr}$ > 0.5 or 0.6 (<0.5 or 0.4). The positions are held for 12 months. We measure excess returns using the size-adjusted returns (*SAR*). For stock *i*, it is calculated as

$$SAR_i = \prod_{t=1}^{12}(1 + R_{it}) - \prod_{t=1}^{12}(1 + R_{st}),$$

where $R_{it}$ is the return on stock *i* in month *t*, and $R_{st}$ is the value-weighted returns on the market capitalization-matched decile portfolio in month *t*. When computing *SAR*, we use the NYSE breakpoints to assign each stock to its corresponding size decile (Hou et al., 2020). Moreover, the return data are corrected for delisting bias, as suggested by Shumway (1997) and Shumway and Warther (1999). The results are stronger when we use market-adjusted returns, for which $R_{st}$ is replaced with the value-weighted returns on the market portfolio in month *t*, $R_{mt}$ (untabulated).

## 4. Predicting the Direction of Earnings Changes

### *4.1. Primary Results*

Table 5 reports the out-of-sample prediction performance in the test period for all the observations ($\widehat{Pr} > 0.5$ and $\widehat{Pr} \leq 0.5$) and observations excluding borderline cases ($\widehat{Pr} \geq 0.6$

and $\widehat{Pr} \leq 0.4$).[20] We observe that on average, 61.9 to 67.5 percent of observations are correctly predicted. Among predicted increases, 60.05 to 65.64 percent are actually earnings increases. These statistics however depend on specific cutoffs (e.g., 0.5 or 0.6). We also report the AUC, which is equivalent to the probability that a randomly chosen earnings increase will be ranked higher by a classifier than will a randomly chosen earnings decrease observation (Fawcett 2006). The AUC ranges from 67.52 to 68.66 percent depending on methods (RF or SGB) and samples (the full sample or the sample with $\widehat{Pr} \geq 0.6$ and $\widehat{Pr} \leq 0.4$), significantly higher than 50 percent of a random guess. Following Carpenter and Bithell (2000), we construct a bootstrap *p*-value for the difference between our AUCs and 50 percent. Specifically, we use a bootstrap sample with the same size as the original test sample to compute a bootstrap AUC and repeat this 10,000 times. The *p*-value is the proportion of 10,000 bootstrap AUCs that are below 50 percent. We observe that all the *p*-values are less than 0.01, indicating that our models' predictive power is unlikely to be a random outcome.[21] The results suggest that the machine learning models extract meaningful fundamental signals from the detailed financial data.

To understand the predictive power sources, we estimate each variable's importance by computing the AUC decrease when that variable is randomly shuffled (Breiman 2001). Since a model is trained and validated for each test year, a predictor has four importance values (one for each test year of 2015-2018) under each method (random forests or stochastic gradient boosting). We compute the correlation of importance values between two consecutive years across all predictors (i.e., N = 13,881). For the three pairs of consecutive years (2015 vs. 2016, 2016 vs.

[20] The chosen parameter values for each method are reported in Online Appendix Table A5. The values are relatively stable over time and do not cluster on the lower or upper bounds, suggesting that the allowed range for each parameter is typically not binding. For example, in only one out of eight cases (two methods × four test years), the chosen number of trees is at the boundary (500; stochastic gradient boosting for the test year of 2017).

[21] To address the issue related to overlapping training/validation sets for test years (2015-2018), we construct a bootstrap *p*-value for each test year and observe that all the *p*-values are less than 0.01.

2017, and 2017 vs. 2018), the correlation coefficients are 0.98, 0.98, and 0.98, respectively, for random forests, and 0.82, 0.75, and 0.85, respectively, for stochastic gradient boosting. The results suggest that the importance of a variable in predicting the direction of one-year-ahead earnings changes is highly stable over time. As such, we average the four importance values for each predictor. Table 6 Panel A presents the top 10 most important variables. Most of them pertain to earnings, such as "NetIncomeLoss" and "EarningsPerShareBasic." The results suggest that among all the financial items in 10-K filings, earnings are still the most critical metrics for fundamental analysis. For stochastic gradient boosting, several balance-sheet items and tax-related variables also make into the top 10 list. We also observe cash flows from investing activities and SG&A in the top 70 list for stochastic gradient boosting and R&D expense in the top 70 list for random forests (untabulated). The results suggest that investment activities exhibit sizable predictive power for the direction of earnings changes in the next year, but the power is not as strong as earnings and tax-related items.[22]

Figure 7 shows the sum of variable importance by category. In aggregate, footnote disclosures contribute the most in forecasting the direction of one-year-ahead earnings changes, followed by balance sheet, income statement, and cash flow statement. Comprehensive income statement and shareholders' equity statement contribute the least to the predictive power. The results are consistent with footnote disclosures carrying important information for valuation (De Franco et al. 2011). As shown in Table 3 Panel A, footnote disclosures contain the most tags (2,443 out of 4,627), which can explain their importance in aggregate. Figure 7 also reports the mean of variable importance within each category. We observe that items from financial

[22] The results are likely due to the focus on one-year-ahead earnings changes; investments and R&D will probably be stronger contributors for longer-term earnings predictions. Given the limited number of years of data, we do not examine their importance in predicting long-term earnings and leave it to future research.

statements on average play a stronger role than footnote items, but the latter's importance is still considerable.

Table 6 Panels B and C show the top 10 important variables within each category. We observe many tax-related items (e.g., "DeferredTaxAssetsValuationAllowance") in the top 10 list for footnote disclosures, consistent with tax items carrying important information on future taxable income (Miller and Skinner 1998; Lev and Nissim 2004; Hanlon 2005; Thomas and Zhang 2011). For example, Miller and Skinner (1998) manually collect valuation allowance for deferred tax assets for 200 companies and find that it is negatively associated with future taxable income.

To visualize the marginal effect of tax items on $\widehat{Pr}$, we construct partial dependence plots (Hastie et al. 2009).[23] Figure 8a shows a nonlinear negative effect of the valuation allowance for deferred tax assets (the top 1 predictor from footnotes) under random forests, consistent with Miller and Skinner (1998). We also observe an interaction effect in Figure 8b: $\widehat{Pr}$ becomes higher when the valuation allowance is lower and lagged operating income (the top 1 predictor under random forests) is higher. The results suggest that the valuation allowance provides additional details on the growth of operating income by revealing management's assessment of future taxable income. Figure 8c presents a nonlinear negative effect of tax benefits related to the exercise of employee stock options (Hanlon and Shevlin 2002; the top 1 predictor from footnotes) under stochastic gradient boosting. Figure 8d shows an interaction effect: $\widehat{Pr}$ becomes

[23] In a one-way partial dependence plot, for each value of a predictor (in the x-axis), we force all observations in the training sample to assume that value for the predictor without changing any data points for other predictors, compute the forecasts using the chosen model, and average forecasts across all observations. The value of the average forecast is for the y-axis. In a two-way partial dependence plot, for each value combination of two predictors (in both the x-axis and y-axis), we force all observations in the training sample to assume the value combination for the two predictors without changing any data points for other predictors, compute the forecasts using the chosen model, and average forecasts across all observations. The value of the average forecast is coded by color.

lower when both the tax benefits and lagged retained earnings (the top 1 predictor under stochastic gradient boosting) are higher. The results are consistent with Bartov and Mohanram's (2004) finding that the exercise of executive stock options predicts disappointing earnings and reveals management's private information on the reversal of previously inflated earnings.

Finally, we compare the out-of-sample performance between our models and Ou and Penman's (1989), who estimate the summary measure by applying logistic regressions to 65 accounting variables from Compustat.[24] For each year in the test period (2015-2018), we use the past three years to estimate a logistic model and then apply the model to the test year. Follow their variable selection approach, we first run a univariate logistic regression for each of the 65 variables and retain only variables that load significantly at the 10 percent level. Second, a multivariate logistic regression is estimated using all remaining variables. We then drop all variables with coefficients that are not significant at the 10 percent level. In a final stage, for the remaining variables, we delete the variables that do not load significantly at the 10 percent level stepwise until all explanatory variables have statistically significant coefficients at the 10 percent level. We refer to this logistic model with Ou and Penman's 65 variables as OP/Logit. To better understand the difference between our models and OP/Logit, we also apply the machine learning methods to the 65 variables and refer to the two models as OP/RF and OP/SGB.

Figure 9 reports the ROC curves for these three models and our two models (XBRL/RF and XBRL/SGB). We find that our models significantly outperform OP/Logit by a large margin. The XBRL/RF (XBRL/SGB) model exhibits an AUC of 67.52 (67.54) percent, compared with 61.79 percent for OP/Logit. We also observe an AUC of 66.63 (66.87) percent for the OP/RF

[24] Ou and Penman (1989) use 68 financial variables. We exclude three variables (%Δ in total uses of funds, %Δ in total sources of funds, and %Δ in funds) as they are no longer reported. None of the three variables is statistically significantly associated with the direction of one-year-ahead earnings changes in Ou and Penman (1989).

(OP/SGB) model, which is significantly higher than that of OP/Logit and marginally lower than XBRL/RF (XBRL/SGB). Following Carpenter and Bithell (2000), we construct a bootstrap *p*-value for the AUC difference. Specifically, for each comparison between two data/method combinations, we use a bootstrap sample with the same size as the original test sample to compute a bootstrap AUC for each combination and repeat this 10,000 times. The *p*-value is the proportion of 10,000 bootstrap AUC differences that are below zero. We observe that all the *p*-values are less than 0.1 except for XBRL/SGB vs. OP/SGB, for which the *p*-value is 0.108. Thus, the improvements in predictive power are unlikely to be random outcomes.[25] The results suggest that our models' superior performance comes from primarily flexible functional forms in machine learning and secondarily more detailed financial information in XBRL documents, which are not used under human experts' guidance and potentially suffer from data quality issues. We address the data quality issues in Sections 4.2. and 4.3.

### *4.2. Using Compustat as an Alternative Source of Detailed Financial Information*

We use Compustat as an alternative source of detailed financial information to XBRL. Compared with XBRL-tagged data, Compustat has its own advantages, such as more extensive standardized adjustments to improve data quality and disadvantages, such as less detailed coverage of financial information.[26] There are 883 financial items from Compustat, for which we take current values, lagged values, and percentage changes, resulting in 2,649 predictors. We scale the current and lagged predictors by total assets (except for total assets itself and items on a

[25] To address the issue related to overlapping training/validation sets for test years (2015-2018), we construct the pairwise bootstrap *p*-values for each test year and observe that all the *p*-values are less than 0.1 except for XBRL/RF vs. OP/RF and XBRL/SGB vs. OP/SGB in 2015, which is unsurprising, given the data quality issues in the early years in the training sample (2012-2013) for 2015.

[26] The adjustments create discrepancies between the accounting numbers in Compustat and 10-K filings (Chychyla and Kogan 2015).

per-share basis). Table 7 Panel A reports the predictive power of detailed financial information from Compustat similar to XBRL-tagged data. The results suggest that the influence of noise in XBRL documents is on par with their additional financial details relative to Compustat.

*4.3. Variation in Data Quality*

We conduct two additional tests based on variation in XBRL data quality. First, since errors in XBRL documents expose the filers to limited liability within two years after the initial adoption (SEC 2009), we classify the test year 2015 as the early period, the training period (2012-2013) for which is fully covered by the liability protection, and 2016-2018 as the late period. As shown in Table 7 Panel B, our model exhibits a higher AUC in the late period than the early period. The bootstrap *p*-value for the AUC difference between the two periods is 0.049 for random forests and 0.307 for stochastic gradient boosting. The results suggest that data quality issues in early adoption years decrease the usefulness of detailed financial data in XBRL documents.[27]

Second, we use the proportion of unique custom and uncommon standard tags in an XBRL submission as an inverse measure of data quality at the firm level. This measure captures the amount of information lost due to the use of extensions and uncommon tags that cannot be used for modeling and thus were removed before the analysis. We split the test sample by the year median and report the AUC of each subsample in Table 7 Panel C. Our model exhibits a higher AUC for firms with high data quality than other firms. The bootstrap *p*-value for the AUC

[27] We also repeat this analysis using all Compustat items, which do not experience the same data quality changes as XBRL documents. We observe, as expected, an insignificant AUC difference between the early and late periods (see Online Appendix Table A6).

difference between the two subsamples is less than 0.01 for both methods. The results suggest that low data quality reduces the predictive power of detailed financial data in XBRL documents.

## 5. Portfolio Returns

### *5.1. Primary Results*

Table 8 reports the size-adjusted returns over the 12 months on the portfolios constructed according to the estimated summary measure $\widehat{Pr}$ using machine learning and detailed financial data. In the top panel for random forests, we observe a monotonic increase in the 12-month SAR as $\widehat{Pr}$ moves from below 0.1 to above 0.8. A similar pattern is observed in the bottom panel for stochastic gradient boosting except for the two extreme sort portfolios ($\widehat{Pr} \leq 0.2$ and $\widehat{Pr} > 0.8$ ), which consist of only a handful of stocks and thus are subject to substantial noise. A hedge portfolio with a long (short) position for stocks with $\widehat{Pr} > 0.5$ ($\widehat{Pr} \leq 0.5$) yields size-adjusted returns of 5.02 percent for random forests and 6.57 percent for stochastic gradient boosting. These returns account for 38.7 and 50.7 percent of returns from a strategy with perfect foresight of the direction of one-year-ahead earnings changes.

To evaluate the extent to which the returns are generated by chance, we place them in the distribution of hedge returns under the null hypothesis that $\widehat{Pr}$ is unrelated to subsequent stock returns. Specifically, for each model, we randomly draw with replacement the same number of stocks as those in the long and short positions, compute the 12-month size-adjusted returns for this pseudo hedge portfolio, and repeat this process 10,000 times. The *p*-values less than 0.0001 for both returns (5.02 and 6.57 percent) suggest that they are unlikely to be random outcomes. When we exclude the borderline cases and take a long (short) position for stocks with $\widehat{Pr} > 0.6$

($\widehat{Pr} \leq 0.4$), the size-adjusted returns are more impressive, 9.43 percent for random forests and 9.74 percent for stochastic gradient boosting.

Figure 10 reports the returns on the short and long positions of our strategies by year. Most of the returns come from the long rather than the short positions. The results suggest that the hedge returns are unlikely explained by limits to arbitrage arising from short-sell constraints. We also observe that the returns are larger in more recent years (2016-2018) than the first year (2015) of the test period, consistent with the findings in Table 7 Panel B.

Finally, we compare our strategies with Ou and Penman's (1989) and those based on the application of our machine learning methods to their 65 accounting variables. Figure 11 reports the size-adjusted returns for the five models, as discussed in the previous section. As the results are similar for the 0.5 cutoff in Figure 11a and 0.4/0.6 cutoffs in Figure 11b, we focus on the latter. The strategy of applying machine learning methods to the 65 variables generates size-adjusted returns of 8.14 percent for OP/RF and 5.88 percent for OP/SGB. These returns are significantly higher than 2.69 percent from Ou and Penman's original strategy (OP/Logit), and marginally lower than 9.43 percent for XBRL/RF and 9.74 percent for XBRL/SGB.[28] The results mirror the findings of the predictability of the five models in Section 4.1.

### *5.2. Transaction Costs*

To assess the impacts of transaction costs on our trading profits, we follow Novy-Marx and Velikov (2016) and estimate the effective bid-ask spread using a Bayesian Gibbs sampler on

[28] We conduct a bootstrap test for the difference in returns between each pair of portfolios (i.e., OP/Logit vs. OP/RF, OP/RF vs. XBRL/RF, OP/Logit vs. OP/SGB, and OP/SGB vs. XBRL/SGB). Specifically, we randomly draw with replacement the same number of stocks as those in the long and short positions for each portfolio in a pair and compute the 12-month size-adjusted returns for the pseudo hedge portfolio. We then take a difference in returns between the two pseudo hedge portfolios and repeat this process 10,000 times. The *p*-value is based on the actual difference with respect to the distribution of simulated differences. The *p*-values are less than 0.001 for all pairs.

a generalized Roll (1984) model, as proposed by Hasbrouck (2009). The mean (median) transaction cost from 1993 to 2005 is 0.0129 (0.0138), consistent with Hasbrouck (2009). The summary statistics for the estimated transaction costs from 2009 to 2019 (see Online Appendix Table A7) are consistent with Novy-Marx and Velikov's (2016) finding that the round-trip transaction costs are less than 1 percent over the period of 2000-2009. We also observe a decrease in transaction costs over time, from 0.805 percent in 2009 to 0.335 percent in 2019. Since 12 percent of stock-year observations during 2011-2019 have insufficient daily returns to estimate the spread, following Novy-Marx and Velikov (2016), we impute the values of stock-years with similar market capitalization (*MKVLT*) and idiosyncratic volatility (*IVOL*).[29] Specifically, in each calendar year, we rank all stocks on *MKVLT* and *IVOL*. For stock *i*, we select stock *j* with the shortest Euclidean distance in rank space of the two characteristics: $distance_{ij} = \sqrt{\left(rankMKVLT_i - \text{rank}MKVLT_j\right)^2 + \left(rankIVOL_i - \text{rank}IVOL_j\right)^2}$. Table 9 reports the size-adjusted returns net of the transaction costs. We continue to observe statistically and economically important returns, ranging from 4.77 to 9.30 percent.

*5.3. Alternative Earnings Measures*

We use pro forma earnings per share to measure the direction of earnings changes, as recent studies demonstrate that the earnings used by analysts are of higher quality and more value-relevant to investors, relative to GAAP earnings and non-GAAP earnings reported by managers (Bentley et al. 2018; Bradshaw et al. 2018). To assess our results' sensitivity to this measurement choice, we use two alternative measures: ROE and EBIT per share. As shown in

[29] The idiosyncratic volatility is calculated as the standard deviation of residuals from regressing daily returns of a given year on Fama-French three factors by firm, with the requirement of at least two months of daily return observations.

Table 10 Panel A, we continue to find robust size-adjusted returns net of transaction costs on a hedge portfolio with a long (short) position for stocks with $\widehat{Pr} > 0.6$ ($\widehat{Pr} \leq 0.4$).

*5.4. Excluding Microcaps*

Microcaps refer to stocks with market capitalization below the 20th percentile of the NYSE stocks (Fama and French 2008; Novy-Marx and Velikov 2016). These tiny stocks are on average only about 3 percent of the total market capitalization but account for about 60 percent of the total number of stocks. Microcaps typically exhibit the largest dispersion in signals of trading strategies and thus often account for more than 60 percent of the stocks in extreme sort portfolios. As they are relatively illiquid, strategies that take disproportionately large positions in these stocks are more expensive to trade. We exclude microcaps from the portfolios and report the size-adjusted returns net of transaction costs on a hedge portfolio with a long (short) position for stocks with $\widehat{Pr} > 0.6$ ($\widehat{Pr} \leq 0.4$) in Table 10 Panel B. We continue to observe robust excess returns on the hedge portfolios.

*5.5. Risk-based Explanations*

The size-adjusted returns have accounted for compensations for risk to some extent. We further control for the conventional risk factors. Specifically, we estimate the following five-factor model of Fama and French (2015):

$$R_{pt} - R_{Ft} = a_i + b_i(R_{Mt} - R_{Ft}) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{it},$$

where $R_{pt}$ is the monthly returns on a hedge portfolio with a long (short) position for stocks with $\widehat{Pr} > 0.6$ ($\widehat{Pr} \leq 0.4$) and $R_{Ft}$ is the one-month T-bill rate. The explanatory variables are the market returns in excess of the one-month T-bill rate ($R_{Mt} - R_{Ft}$), and returns on size (*SMB*),

book-to-market (*HML*), profitability (*RMW*), and investment (*CMA*) portfolios. As shown in Table 10 Panel C columns (1) and (3), our portfolios generate monthly excess returns of 0.79 to 0.94 percent, which translate into annualized returns of 9.48 to 11.28 percent. We compute the Newey-West standard errors and find that the excess returns are also significantly higher than zero with *p*-values < 0.05 and 0.01. We also adjust the dependent variable by subtracting Fama-French 30 industry monthly returns for each stock in the portfolio (Richardson et al. 2010). Columns (2) and (4) report monthly industry-adjusted excess returns of 0.66 to 0.69 percent, significantly higher than zero with *p*-values < 0.1 and 0.05.

*5.6. Predicting the Amount of Earnings Changes*

We predict the direction of earnings changes since prior research demonstrates that it is difficult to predict the amount of earnings changes (Gerakos and Gramacy 2013; Li and Mohanram 2014), and transforming the amount to the direction reduces the variability in earnings changes (Freeman et al. 1982). Nevertheless, we use the two machine learning methods to predict the amount of earnings changes following the same rolling windows in Figure 5 and timeline in Figure 6. Consistent with prior research, we observe a low out-of-sample $R^2$ of 8 percent (5.8 percent) for random forests (stochastic gradient boosting). A hedge portfolio with a long (short) position for stocks with the predicted earnings changes $\Delta\widehat{\text{Earnings}} > 0$ ($\Delta\widehat{\text{Earnings}} \leq 0$) yields size-adjusted returns of 0.098 percent (*p*-value = 0.4668) for random forests and 0.109 percent (*p*-value = 0.4601) for stochastic gradient boosting. The results suggest that focusing on the direction of earnings changes is preferable to make our machine learning models successful.

## 6. Analysts' Earnings Forecasts

In Sections 4 and 5, we use predictions from Ou and Penman's (1989) model as a prominent example to evaluate our summary measure $\widehat{Pr}$. In this section, we use analysts' earnings forecasts as another example. We take the consensus (i.e., median) analyst forecast in the month following the portfolio formation and compare it with the realized earnings in fiscal year *t* to determine whether analysts forecast an earnings increase or decrease. Table 11 Panel A shows an AUC of 63.62 percent for analysts' prediction of an earnings increase, which is significantly lower than those of our models, which are reproduced in columns (2)-(3) (bootstrap *p*-value < 0.01). A hedge portfolio with a long (short) position for stocks with a predicted earnings increase (decrease) yields size-adjusted returns of 2.49 percent, lower than those from our models.

Finally, we examine whether analysts fully understand the implications of detailed financial data on future earnings changes. We compute analysts' forecast errors as the actual earnings in fiscal year *t* + 1 minus the consensus (i.e., median) analyst forecast in the month following the portfolio formation, scaled by the close price on the portfolio formation date, multiplied by 100. We regress analysts' forecast errors on $\widehat{Pr}$, the log of market cap, and book-to-market ratios. Table 11 Panel B shows a significant positive coefficient on $\widehat{Pr}$ for random forests (0.48 with a *p*-value < 0.1) and stochastic gradient boosting (0.59 with a *p*-value < 0.05). The results suggest that analysts fail to fully incorporate detailed financial information into their forecasts.[30]

[30] We also examine whether analysts' earnings forecasts help improve our hedge returns to the extent that some earnings increases/decreases are anticipated and thus do not help earn future excess returns. A hedge portfolio with a long (short) position for stocks with $\widehat{Pr} > 0.5$ and an earnings decrease predicted by analysts ($\widehat{Pr} \leq 0.5$ and an earnings increase predicted by analysts) yields size-adjusted returns of 6.40 percent for random forests and 7.01 percent for stochastic gradient boosting, higher than the original returns (5.02 and 6.57 percent, respectively).

## 7. Conclusion

We conduct a fundamental analysis of a large set of detailed financial information aimed at predicting earnings. Since 2012, all U.S. public companies must tag quantitative amounts in financial statements and footnotes of their 10-K reports using XBRL. Applying machine learning methods (random forests and stochastic gradient boosting), we combine the detailed financial data into a summary measure for the direction of one-year-ahead earnings changes. The measure shows significant out-of-sample predictive power concerning the direction of earnings changes. The AUC ranging from 67.52 to 68.66 percent is significantly higher than that of a random guess, which is 50 percent. Hedge portfolios are formed based on this measure during the period 2015-2018. The annual size-adjusted returns to the hedge portfolios range from 5.02 to 9.74 percent. These returns persist after accounting for transaction costs and risk.

Our measure and strategies outperform those of Ou and Penman (1989), who extract the summary measure from 65 accounting variables using logistic regressions. Additional analyses suggest that the outperformance stems from primarily nonlinear predictor interactions, missed by regressions, and secondarily the use of more detailed financial data in XBRL documents. The former indicates that machine learning can unleash more predictive power of financial statements concerning future earnings and returns. The latter suggests that, despite data quality issues, XBRL-tagged detailed financial data (with little access cost) still serve as useful inputs for fundamental analysis. Overall, our evidence suggests that applying machine learning to the detailed financial data in XBRL documents can reveal valuable fundamental signals that have not been fully impounded into stock prices.

**References**


Anand, V., R. Brunner, K. Ikegwu, and T. Sougiannis. 2019. Predicting profitability using machine learning. Working Paper. University of Illinois at Urbana-Champaign.

Balachandran, S., and P. Mohanram. 2011. Is the decline in value relevance of accounting driven by increased conservatism? *Review of Accounting Studies* 16, 272-301.

Bao, Y., B. Ke, B. Li, J. Yu, and J. Zhang. 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research* 58, 199-235.

Barth, M., K. Li, and C. McClure. 2018. Evolution in value relevance of accounting information. Working Paper. Stanford University.

Bartley, J., A. Chen, and E. Taylor. 2011. A comparison of XBRL filings to corporate 10-Ks—Evidence from the voluntary filing program. *Accounting Horizon* 25, 227-245.

Bartov, E., and P. Mohanram. 2004. Private information, earnings manipulations, and executive stock-option exercises. *The Accounting Review* 79, 889-920.

Bentley, J., T. Christensen, K. Gee, and B. Whipple. 2018. Disentangling managers' and analysts' non-GAAP reporting. *Journal of Accounting Research* 56, 1039-1081.

Bertomeu, J., E. Cheynel, E. Floyd, and W. Pan. 2020. Using machine learning to detect misstatements. *Review of Accounting Studies*, Forthcoming.

Bhattacharya, N., Y. Cho, and J. Kim. 2018. Leveling the playing field between large and small institutions: Evidence from the SEC's XBRL mandate. *The Accounting Review* 93, 51-71.

Blankespoor, E. 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. *Journal of Accounting Research* 57, 919-967.

Blankespoor, E., B. Miller, and H. White. 2014. Initial evidence on the market impact of the XBRL mandate. *Review of Accounting Studies* 19, 1468-1503.

Boritz, E., and W. No. 2009. Assurance on XBRL-related documents: The case of United Technologies Corporation. *Journal of Information Systems* 23, 49-78.

Bradshaw, M., T. Christensen, K. Gee, and B. Whipple. 2018. Analysts' GAAP earnings forecasts and their implications for accounting research. *Journal of Accounting and Economics* 66, 46-66.

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

Brown, L. 1996. Influential accounting articles, individuals, Ph.D. granting institutions and faculties: A citational analysis. *Accounting, Organization and Society* 21, 723-754.

Carpenter, J., and J. Bithell. 2000. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19, 1141-1164.

Cecchini, M., H. Aytug, G. Koehler, and P. Pathak. 2010. Detecting management fraud in public companies. *Management Science* 56, 1146-1160.

Chychyla, R., and A. Kogan. 2015. Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings. *Journal of Information Systems* 29, 37-72.

Cohen, L., C. Malloy, and Q. Nguyen. 2020. Lazy prices. *Journal of Finance* 75, 1371-1415.

Core, J., W. Guay, and A. Van Buskirk. 2003. Market valuations in the New Economy: An investigation of what has changed. *Journal of Accounting and Economics* 34, 43-67.

De Franco, G., M.H. Wong, and Y. Zhou. 2011. Accounting adjustments and the valuation of financial statement note information in 10-K filings. *The Accounting Review* 86, 1577-1604.

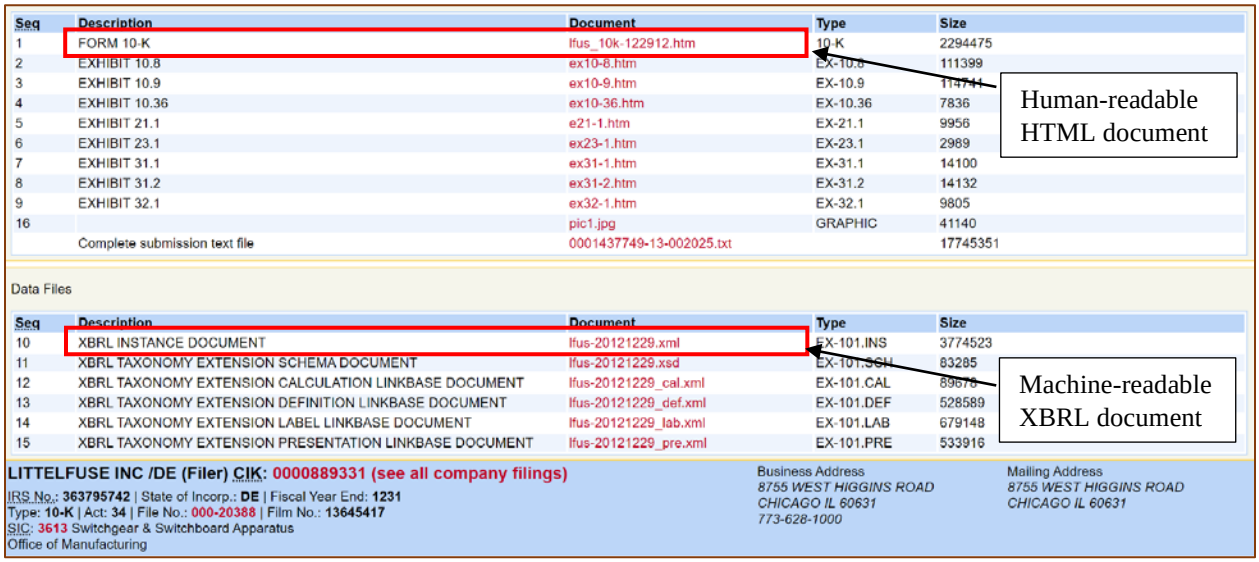
Debreceny, R., S. Farewell, M. Piechocki, C. Felden, and A. Graning. 2010. Does it add up? Early evidence on the data quality of XBRL filings to the SEC. *Journal of Accounting and Public Policy* 29, 296-306.
Debreceny, R., S. Farewell, M. Piechocki, C. Felden, A. Graning, and A. d'Eri. 2011. Flex or Break? Extensions in XBRL disclosures to the SEC. *Accounting Horizon* 25, 631-657.
Ding, K., B. Lev, X. Peng, T. Sun, and M. Vasarhelyi. 2020. Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies* 25, 1098-1134.
Dong, Y., O. Li, Y. Lin, and C. Ni. 2016. Does information processing cost affect firm-specific information acquisition? Evidence from XBRL adoption. *Journal of Financial and Quantitative Analysis* 51, 435-462.
Du, H., M. Vasarhelyi, and X. Zheng. 2013. XBRL mandate: thousands of filing errors and so what? *Journal of Information Systems* 27, 61-78.
Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64, 221-245.
Efendi, J., J. Park, C. Subramaniam. 2016. Does the XBRL reporting format provide incremental information value? A study using XBRL disclosures during the voluntary filing program. *Abacus* 52, 259-285.
Fama, E., and K. French. 2008. Dissecting anomalies. *Journal of Finance* 63, 1653-1678.
Fama, E., and K. French. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116, 1-22.
FASB. 2018. SEC reporting taxonomy technical guide. Available at https://www.fasb.org/cs/ContentServer?d=Touch&c=Document_C&pagename=FASB%2FDocument_C%2FDocumentPage&cid=1176169716122
Fawcett, T. 2006. An introduction to ROC Analysis. *Pattern Recognition Letters* 27, 861-874.
Frankel, R., J. Jennings, and J. Lee. 2016. Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics* 62, 209-227.
Freeman, R., J. Ohlson, and S. Penman. 1982. Book rate-of-return and prediction of earnings changes: An empirical investigation. *Journal of Accounting Research* 20, 639-653.
Freyberger, J., N. Andreas, and M. Weber. 2020. Dissecting characteristics nonparametrically. *Review of Financial Studies* 33, 2326-2377.
Friedman, J. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 367-378.
Gerakos, J., and R. Gramacy. 2013. Regression-based earnings forecasts. Chicago Booth Research Paper No. 12-26.
Green, J., J. Hand, and F. Zhang. 2013. The superview of return predictive signals. *Review of Accounting Studies* 18, 692-730.
Green, J., J. Hand, and F. Zhang. 2017. The characteristics that provide independent information about average US monthly stock returns. *Review of Financial Studies* 30 4389-4436.
Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33, 2223-2273.
Hanlon, M. 2005. The persistence and pricing of earnings, accruals, and cash flows when firms have large book-tax differences. *The Accounting Review* 80, 137-166.
Hanlon, M., and T. Shevlin. 2002. Accounting for tax benefits of employee stock options and implications for research. *Accounting Horizon* 16, 1-16.

Harris, T., and S. Morsfield. 2012. An evaluation of the current state and future of XBRL and interactive data for investors and analysts. White Paper. Columbia University.
Hasbrouck, J. 2009. Trading costs and returns for US equities: Estimating effective costs from daily data. *Journal of Finance* 64, 1446-1477.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. Springer, New York.
Hsieh, T., and J. Bedard. 2018. Impact of XBRL on voluntary adopters' financial reporting quality and cost of equity capital. *Journal of Emerging Technologies in Accounting* 15, 45-65.
Hoitash, R., and U. Hoitash. 2018. Measuring accounting reporting complexity with XBRL. *The Accounting Review* 93, 259-287.
Hoitash, R., U. Hoitash, A. Kurt, and R. Verdi. 2018. An input-based measure of financial statement comparability. Working Paper.
Hou, K., C. Xue, and L. Zhang. 2020. Replicating anomalies. *Review of Financial Studies* 33, 2019-2133.
Hunt, J., J. Myers, and L. Myers. 2019. Improving earnings predictions with machine learning. Working Paper. Mississippi State University.
Kim, J., B. Li, and Z. Liu. 2019a. Information-processing costs and breadth of ownership. *Contemporary Accounting Research* 36, 2408-2436.
Kim, J., J. Kim, and J. Lim. 2019b. Does XBRL adoption constrain earnings management? Early evidence from Mandated US Filers. *Contemporary Accounting Research* 36, 2610-2634.
Kirk, M., J. Vincent, and D. Williams. 2016. From print to practice: XBRL extension use and analyst forecast properties. Working Paper.
Lev, B. 2018. The deteriorating usefulness of financial report information and how to reverse it. *Accounting and Business Research* 48, 465-493.
Lev, B., and D. Nissim. 2004. Taxable income, future earnings, and equity values. *The Accounting Review* 79, 1039-1074.
Lev, B., and P. Zarowin. 1999. The boundaries of financial reporting and how to extend them. *Journal of Accounting Research* 37, 353-385.
Li, F. 2010. The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research* 48, 1049-1102.
Li, K., and P. Mohanram. 2014. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies* 19, 1152-1185.
Li, S., and E. Nwaeze. 2015. The Association between extensions in XBRL disclosures and financial information environment. *Journal of Information Systems* 29, 73-99.
Li, S., and E. Nwaeze. 2018. Impact of extensions in XBRL disclosure on analysts' forecast behavior. *Accounting Horizon* 32, 57-79.
Liu, M. 2021. Assessing human information processing in lending decisions: A machine learning approach. Working Paper. Boston University.
Miller, G., and D. Skinner. 1998. Determinants of the valuation allowance for deferred tax assets under SFAS No. 109. *The Accounting Review* 213-233.
Mullainathan, S., and J. Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31, 87-106.
Novy-Marx, R., and M. Velikov. 2016. A taxonomy of anomalies and their trading costs. *Review of Financial Studies* 29, 104-147.

Ou, J. 1990. The information content of nonearnings accounting numbers as earnings predictors. *Journal of Accounting Research* 28, 144-163.
Ou, J. and S. Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics* 11, 295-329.
Perdana, A., A. Robb, and F. Rohde. 2015. An integrative review of synthesis of XBRL research in academic journals. *Journal of Information Systems* 29, 115-153.
Perols, J. 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory* 30, 19-50.
Plumlee, R.D., and M. Plumlee. 2008. Assurance on XBRL for financial reporting. *Accounting Horizon* 22, 353-368.
Ramesh, K., and R. Thiagarajan. 1996. Inter-tempioral decline in earnings response coefficients. Working Paper.
Richardson, S., I. Tuna, and P. Wysocki. 2010. Accounting anomalies and fundamental analysis: A review of recent research advances. *Journal of Accounting and Economics* 50, 410-454.
Roll, R. 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39, 1127-1139.
Scherr, E., and D. Ditter. 2017. Customization versus standardization in electronic financial reporting: Early evidence from the SEC XBRL mandate. *Journal of Information Systems* 31, 125-148.
SEC. 2009. Interactive data to improve financial reporting. Final rule. Available at https://www.sec.gov/rules/final/2009/33-9002.pdf
SEC. 2016. Staff Observations of Custom Axis Tags. Available at https://www.sec.gov/structureddata/reportspubs/osd_assessment_custom-axis-tags.html
Shumway, T. 1997. The delisting bias in CRSP data. Journal of Finance 52, 327-340.
Shumway, T., and V. Warther. 1999. The delisting bias in CRAP's Nasdaq data and its implications for the size effect. Journal of Finance 54, 2361-2379.
Thomas, J., and F. Zhang. 2011. Tax expense momentum. *Journal of Accounting Research* 49, 791-821.
Varian, H. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28, 3-28.
Zhou, Z. 2012. *Ensemble Learning Methods: Foundations and Algorithms*. CRC Press, Florida.

## Appendix A: Examples of XBRL-tagged Financial Items

This appendix shows where an XBRL document is filed and how financial items are tagged in the XBRL document. The following screenshot shows where a human-readable HTML document and the corresponding machine-readable XBRL document are located on the SEC EDGAR Website for Littelfuse, an electronic manufacturer.




| Seq | Description | Document | Type | Size |
|---|---|---|---|---|
| 1 | FORM 10-K | lfus_10k-122912.htm | 10-K | 2294475 |
| 2 | EXHIBIT 10.8 | ex10-8.htm | EX-10.8 | 111399 |
| 3 | EXHIBIT 10.9 | ex10-9.htm | EX-10.9 | 114744 |
| 4 | EXHIBIT 10.36 | ex10-36.htm | EX-10.36 | 7836 |
| 5 | EXHIBIT 21.1 | e21-1.htm | EX-21.1 | 9956 |
| 6 | EXHIBIT 23.1 | ex23-1.htm | EX-23.1 | 2989 |
| 7 | EXHIBIT 31.1 | ex31-1.htm | EX-31.1 | 14100 |
| 8 | EXHIBIT 31.2 | ex31-2.htm | EX-31.2 | 14132 |
| 9 | EXHIBIT 32.1 | ex32-1.htm | EX-32.1 | 9805 |
| 16 | | pic1.jpg | GRAPHIC | 41140 |
| | Complete submission text file | 0001437749-13-002025.txt | | 17745351 |

Data Files

| Seq | Description | Document | Type | Size |
|---|---|---|---|---|
| 10 | XBRL INSTANCE DOCUMENT | lfus-20121229.xml | EX-101.INS | 3774523 |
| 11 | XBRL TAXONOMY EXTENSION SCHEMA DOCUMENT | lfus-20121229.xsd | EX-101.SCH | 83285 |
| 12 | XBRL TAXONOMY EXTENSION CALCULATION LINKBASE DOCUMENT | lfus-20121229_cal.xml | EX-101.CAL | 89678 |
| 13 | XBRL TAXONOMY EXTENSION DEFINITION LINKBASE DOCUMENT | lfus-20121229_def.xml | EX-101.DEF | 528589 |
| 14 | XBRL TAXONOMY EXTENSION LABEL LINKBASE DOCUMENT | lfus-20121229_lab.xml | EX-101.LAB | 679148 |
| 15 | XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE DOCUMENT | lfus-20121229_pre.xml | EX-101.PRE | 533916 |

**LITTELFUSE INC /DE (Filer) CIK: 0000889331 (see all company filings)**

IRS No.: **363795742** | State of Incorp.: **DE** | Fiscal Year End: **1231**
Type: **10-K** | Act: **34** | File No.: **000-20388** | Film No.: **13645417**
SIC: **3613** Switchgear & Switchboard Apparatus
Office of Manufacturing

Business Address
*8755 WEST HIGGINS ROAD*
*CHICAGO IL 60631*
*773-628-1000*

Mailing Address
*8755 WEST HIGGINS ROAD*
*CHICAGO IL 60631*

**Example 1: Items on the face of financial statements**

Cash and cash equivalents from the human-readable HTML document:

**CONSOLIDATED BALANCE SHEETS**

| (In thousands of USD) | December 29, 2012 | December 31, 2011 |
|---|---|---|
| **ASSETS** | | |
| Current assets: | | |
| Cash and cash equivalents | $ 235,404 | $ 164,016 |
| Short-term investments | — | 13,997 |
| Accounts receivable, less allowances (2012 - $13,508; 2011 - $12,306) | 100,559 | 92,088 |
| Inventories | 75,580 | 75,575 |
| Deferred income taxes | 11,890 | 11,895 |
| Prepaid expenses and other current assets | 16,532 | 14,219 |
| Assets held for sale | 5,500 | 6,592 |

Cash and cash equivalents from the machine-readable XBRL document:

```
<us-gaap:CashAndCashEquivalentsAtCarryingValue unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">235404000</us-gaap:CashAndCashEquivalentsAtCarryingValue>
<us-gaap:CashAndCashEquivalentsAtCarryingValue unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">164016000</us-gaap:CashAndCashEquivalentsAtCarryingValue>
<us-gaap:ShortTermInvestments unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">13997000</us-gaap:ShortTermInvestments>
<us-gaap:AccountsReceivableNetCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">100559000</us-gaap:AccountsReceivableNetCurrent>
<us-gaap:AccountsReceivableNetCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">92088000</us-gaap:AccountsReceivableNetCurrent>
<us-gaap:InventoryNet unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">75580000</us-gaap:InventoryNet>
<us-gaap:InventoryNet unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">75575000</us-gaap:InventoryNet>
<us-gaap:DeferredTaxAssetsNetCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">11890000</us-gaap:DeferredTaxAssetsNetCurrent>
<us-gaap:DeferredTaxAssetsNetCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">11895000</us-gaap:DeferredTaxAssetsNetCurrent>
<us-gaap:PrepaidExpenseAndOtherAssetsCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">16532000</us-gaap:PrepaidExpenseAndOtherAssetsCurrent>
<us-gaap:PrepaidExpenseAndOtherAssetsCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">14219000</us-gaap:PrepaidExpenseAndOtherAssetsCurrent>
<us-gaap:AssetsHeldForSaleCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">5500000</us-gaap:AssetsHeldForSaleCurrent>
<us-gaap:AssetsHeldForSaleCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">6592000</us-gaap:AssetsHeldForSaleCurrent>
```

**Example 2: Items in the footnotes**

Work in process inventory from the human-readable HTML document:

**3. Inventories**

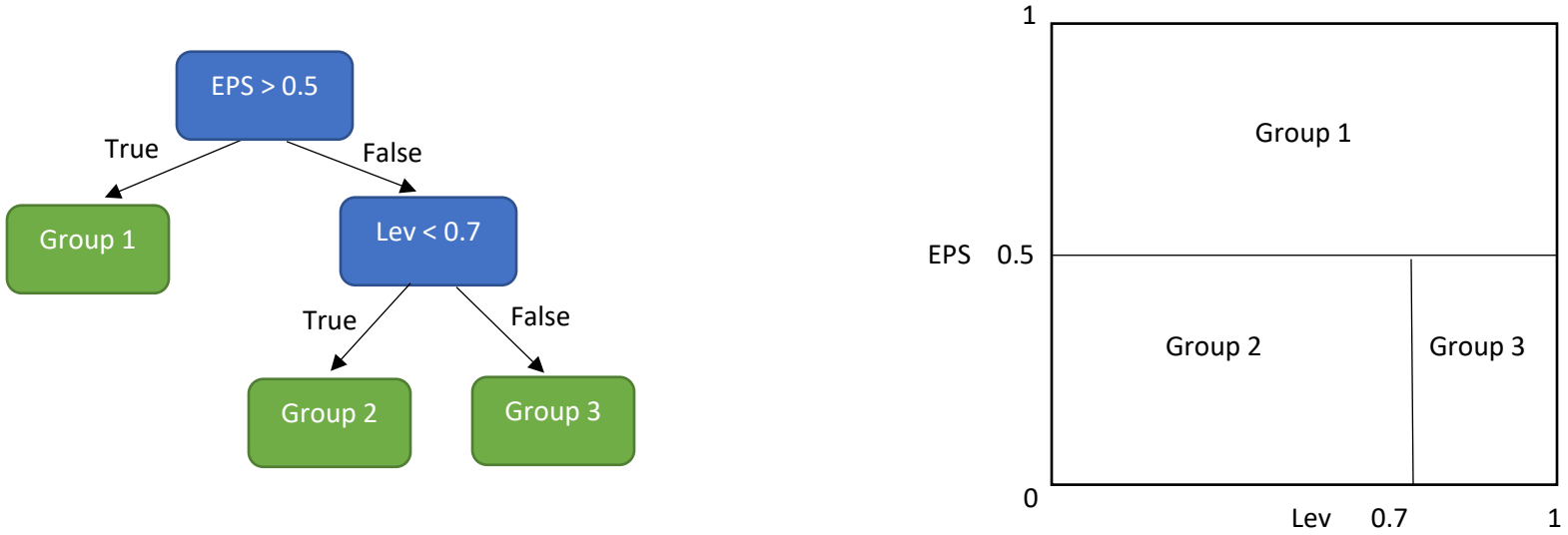
The components of inventories at December 29, 2012 and December 31, 2011 are as follows (in thousands):

| | 2012 | 2011 |
|---|---|---|
| Raw materials | $ 21,689 | $ 26,919 |
| Work in process | 11,868 | 10,704 |
| Finished goods | 42,023 | 37,952 |
| Total | $ 75,580 | $ 75,575 |

Work in process inventory from the machine-readable XBRL document:

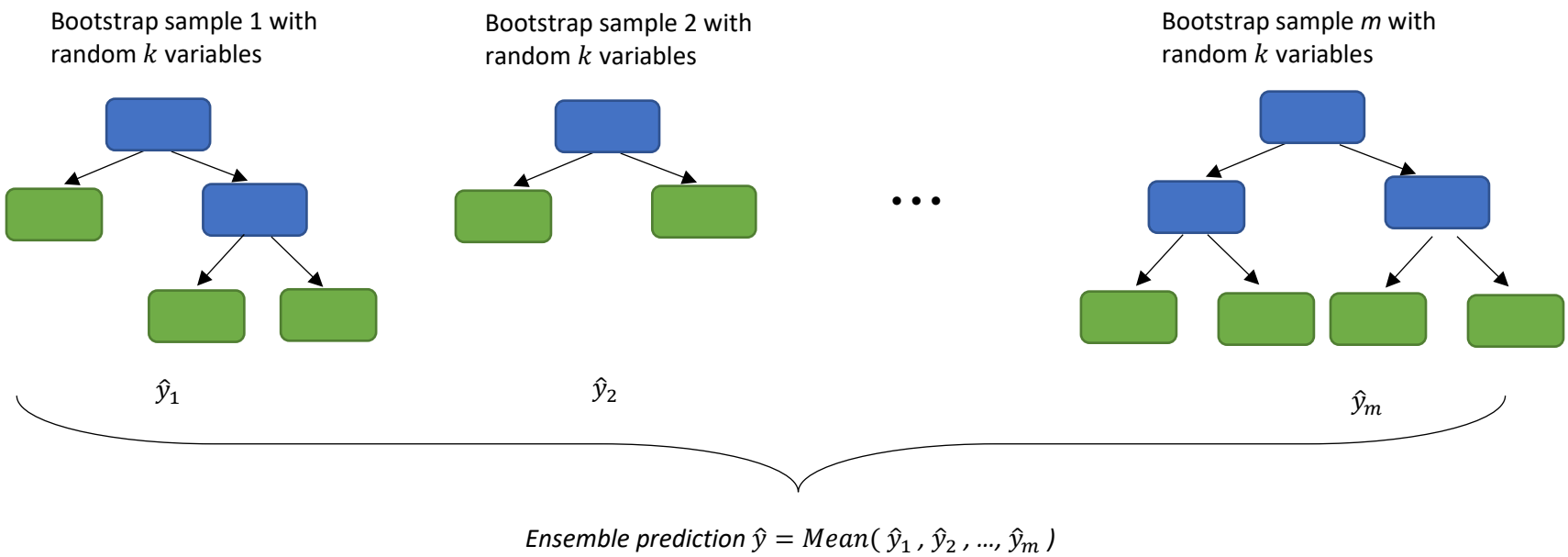
```
<us-gaap:InventoryRawMaterials unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">21689000</us-gaap:InventoryRawMaterials>
<us-gaap:InventoryRawMaterials unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">26919000</us-gaap:InventoryRawMaterials>
<us-gaap:InventoryWorkInProcess unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">11868000</us-gaap:InventoryWorkInProcess>
<us-gaap:InventoryWorkInProcess unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">10704000</us-gaap:InventoryWorkInProcess>
<us-gaap:InventoryFinishedGoods unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">42023000</us-gaap:InventoryFinishedGoods>
<us-gaap:InventoryFinishedGoods unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">37952000</us-gaap:InventoryFinishedGoods>
```

**Figure 1 A Decision Tree Example**



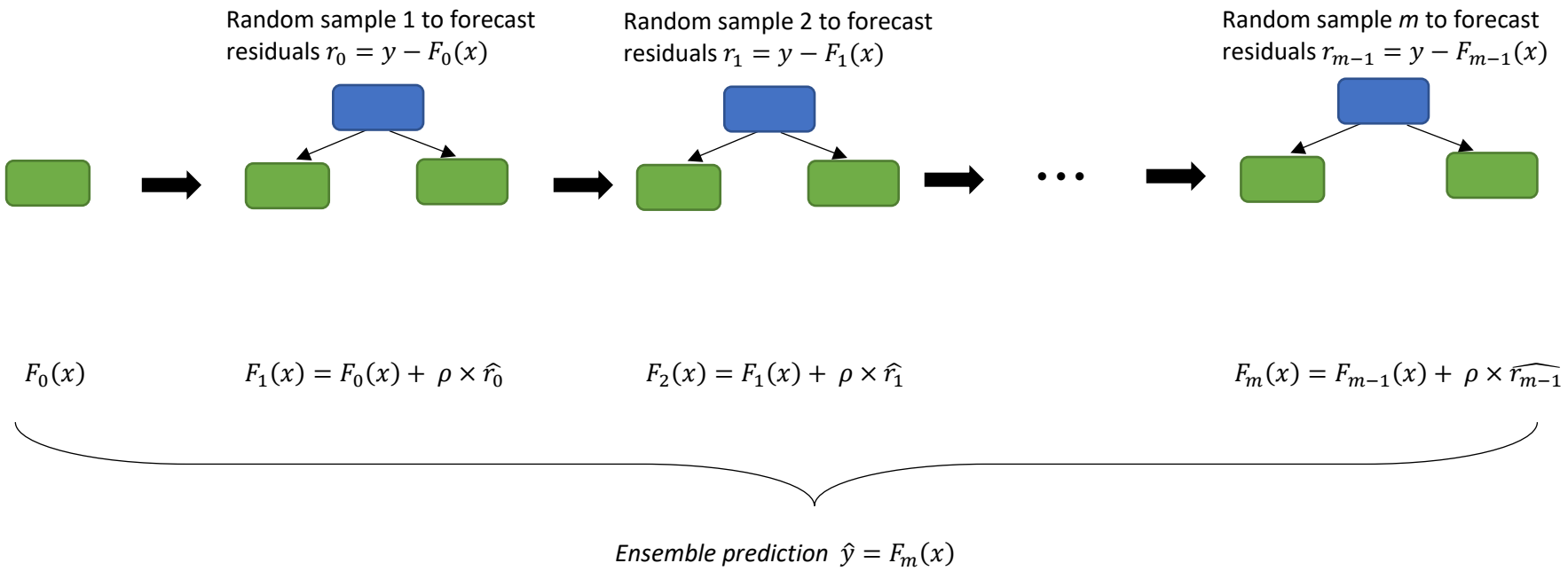

In this figure, the left panel presents an example with two predictors, "EPS" and "Lev" (i.e., earnings per share and leverage), to forecast the direction of one-year-ahead earnings changes. Suppose the tree in the left panel is the final output. It describes how each observation is assigned to a group based on its predictor value. A blue box ("a node") represents a split and a green box ("a leaf") indicates a final partition. First, the sample is sorted on EPS. Observations with EPS above the breakpoint of 0.5 are assigned to Group 1. Those with EPS below 0.5 are then further sorted by Lev: observations with below 0.7 go to Group 2, while those with Lev above 0.7 are assigned to Group 3. The right panel shows how the space of "EPS" and "Lev" is partitioned by this tree model.

**Figure 2 A Random Forest Example**




This figure shows how an ensemble prediction is generated by random forests. A tree is grown based on each of $m$ different bootstrap samples of the data considering only a random subset of predictors ($k$ variables) for splitting. For a given observation, there are $m$ predictions, and the final forecast is the simple average of the $m$ predictions.

**Figure 3 A Stochastic Gradient Boosting Example**




This figure shows how an ensemble prediction is generated by stochastic gradient boosting. It starts by averaging the outcome variable as an initial prediction ($F_0(x)$). It then fits a shallow tree (e.g., with depth $L$=1) to the residuals from the initial prediction ($r_0 = y - F_0(x)$). The fitted value is shrunken by a factor $\rho \in (0,1)$ (i.e., the learning rate) to help prevent the model from overfitting the residuals and added to the initial prediction $F_1(x) = F_0(x) + \rho \times \widehat{r_0}$ to form an ensemble prediction. Then the next tree with the same shallow depth $L$ is used to fit the residuals from the previous prediction. This is repeated $m$ times and the output of this additive model of shallow trees is the final ensemble prediction. To reduce the correlation among estimates at different iterations, the "stochastic" procedure introduces randomness by using a random sample in each iteration.

## Figure 4 Tag Distribution Across XBRL Submissions

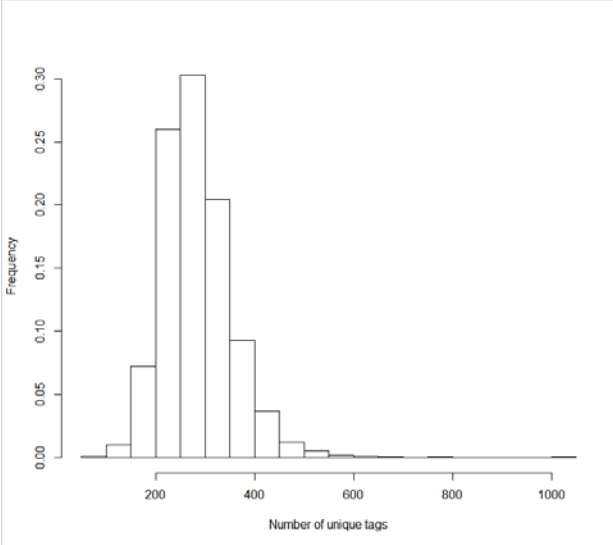



Figure 4a: Histogram by the number of unique tags

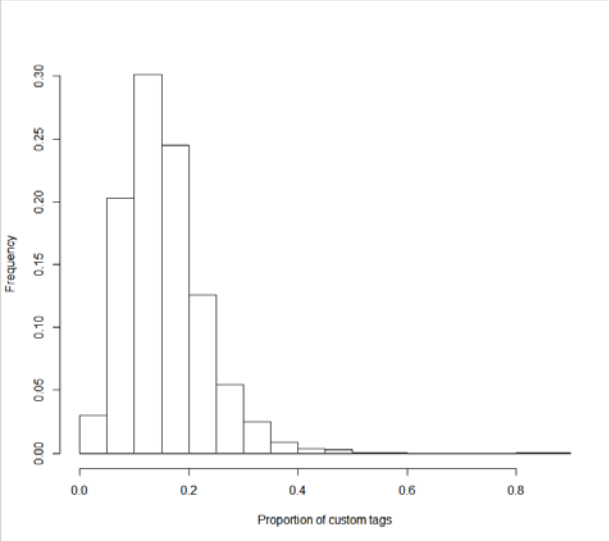



Figure 4b: Histogram by the proportion of custom tags

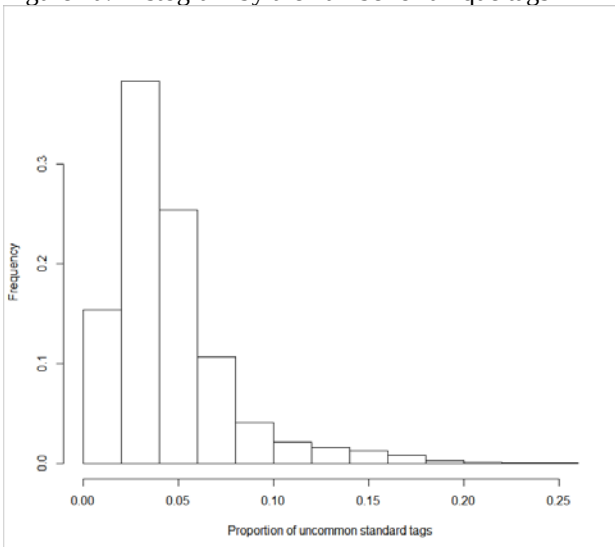



Figure 4c: Histogram by the proportion of uncommon standard tags

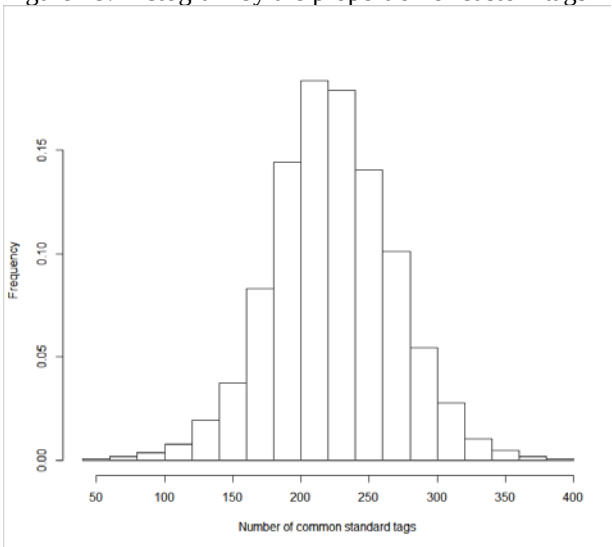


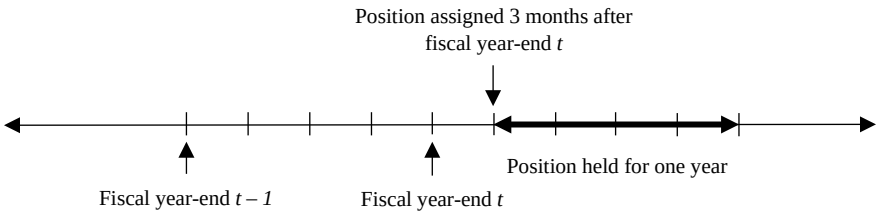

Figure 4d: Histogram by the number of common standard tags

Frequency refers to the proportion of XBRL documents in the 8,149 submissions. Figure 4a shows the histogram by the number of total unique tags (including both custom and standard tags). Figure 4b shows the histogram by the proportion of custom tags, calculated as the number of unique custom tags divided by the number of unique tags. Figure 4c shows the histogram by the proportion of uncommon standard tags, calculated as the number of unique uncommon standard tags divided by the number of unique tags. Uncommon standard tags are standard tags that have not be used at least once in each year. Figure 4d shows the histogram by the number of common standard tags, which are standard tags that have been used at least once in each year. We use 4,627 unique common standard tags in subsequent analyses.

## Figure 5 Rolling Windows for Machine Learning


This figure illustrates the rolling window procedure through which the machine learning models are trained, validated, and used to predict the direction of the one-year-ahead change in earnings. For each year in the test period from 2015 to 2018 (green), the models are trained in the second and third preceding years (blue) and validated in the preceding year (yellow) to tune the parameters in our machine learning models (as shown in Table 4).

## Figure 6 Timeline of the Trading Strategy




This figure shows the timeline of the trading strategy. For each stock in the sample, it is assigned to a long (short) position three months after its fiscal year-end, when $\widehat{Pr}$ > 0.5 or 0.6 (<0.5 or 0.4). The positions are held for twelve months.

## Figure 7 Group Importance

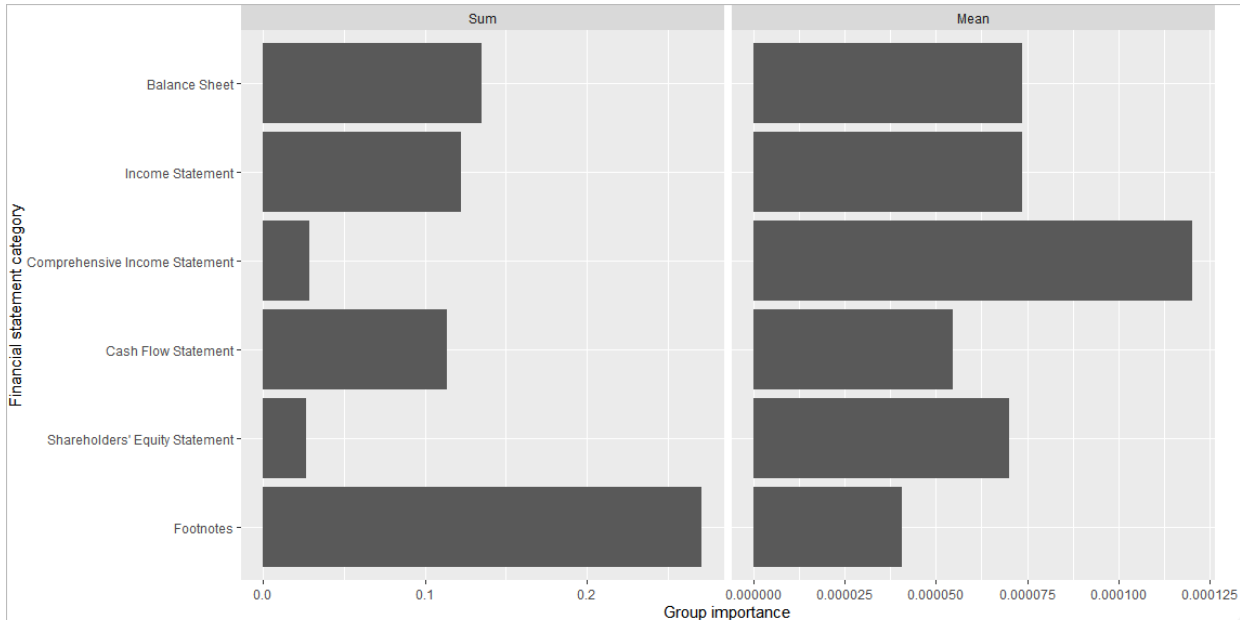



Figure 7a Group importance for random forests

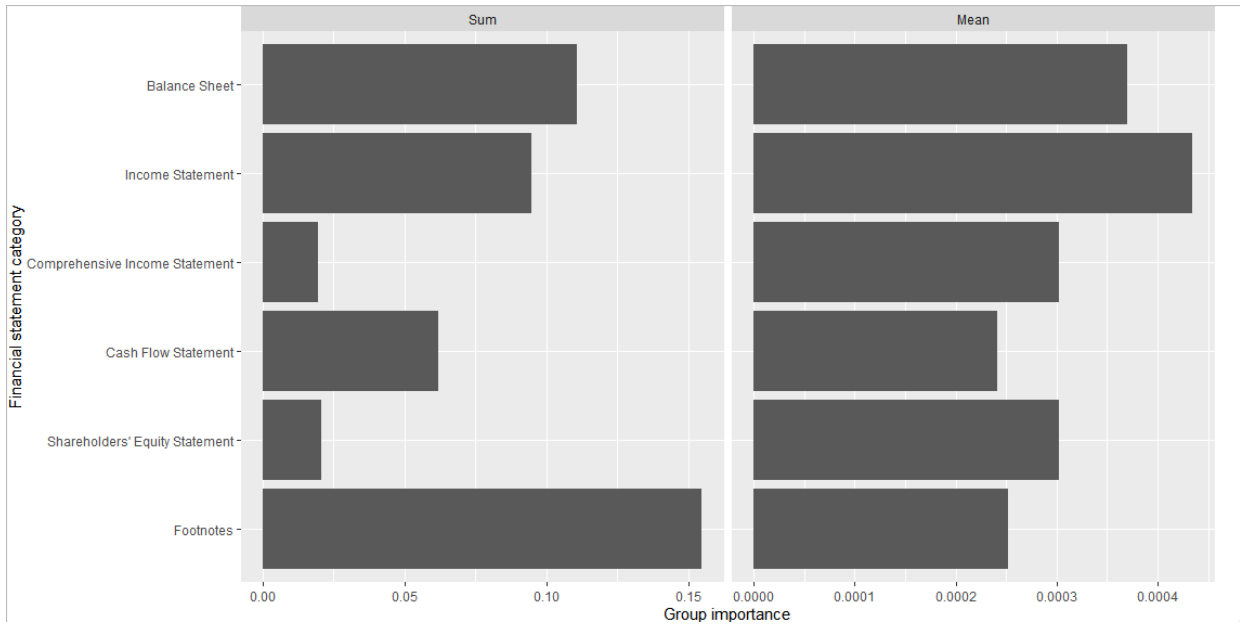



Figure 7b Group importance for stochastic gradient boosting

---

Figures 7a and 7b show the importance of predictors grouped by financial statement category for random forests and stochastic gradient boosting, respectively. Each predictor is classified into balance sheet, income statement, comprehensive income statement, cash flow statement, shareholders' equity statement, or footnotes. The importance of a predictor is computed as the decrease in the AUC when that variable is randomly shuffled. The sum and mean of the predictor importance grouped by financial statement category are reported.

## Figure 8 Partial Dependence Plots

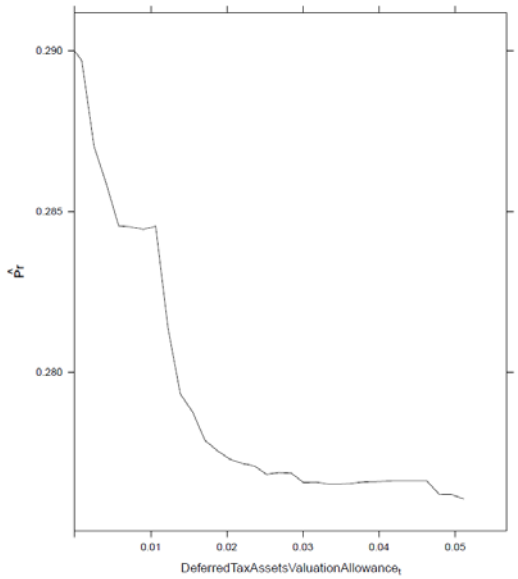



Figure 8a: Valuation allowance for deferred tax assets (random forests)

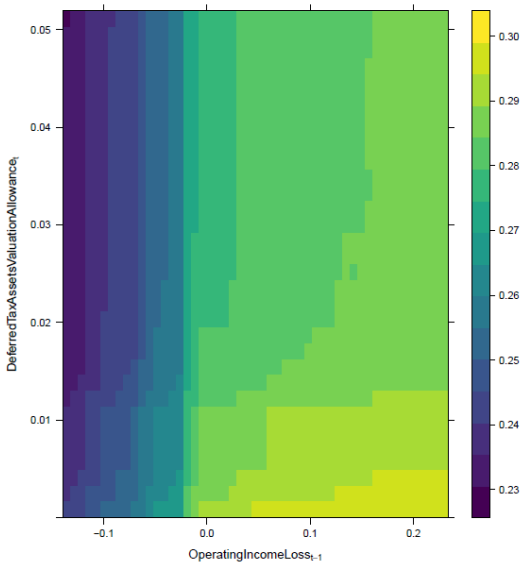



Figure 8b: Valuation allowance for deferred tax assets and operating income (random forests)

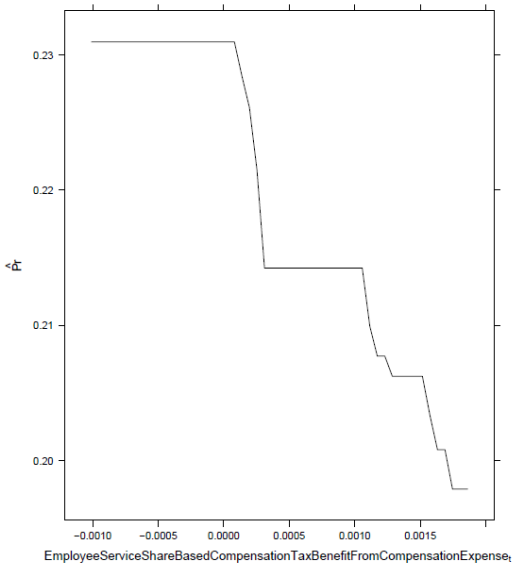



Figure 8c: Tax benefits of share-based compensation (stochastic gradient boosting)

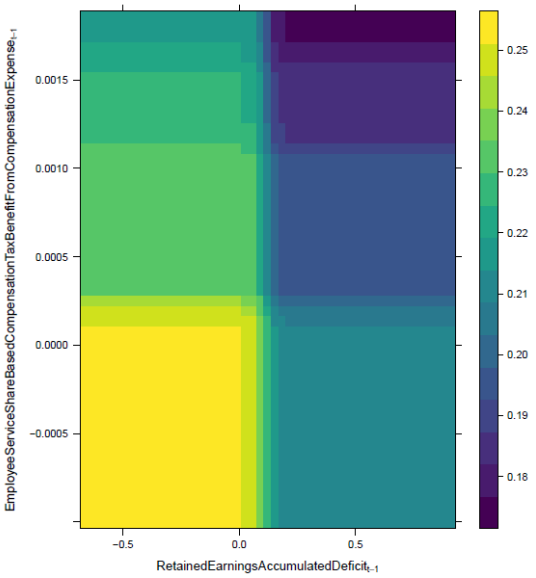
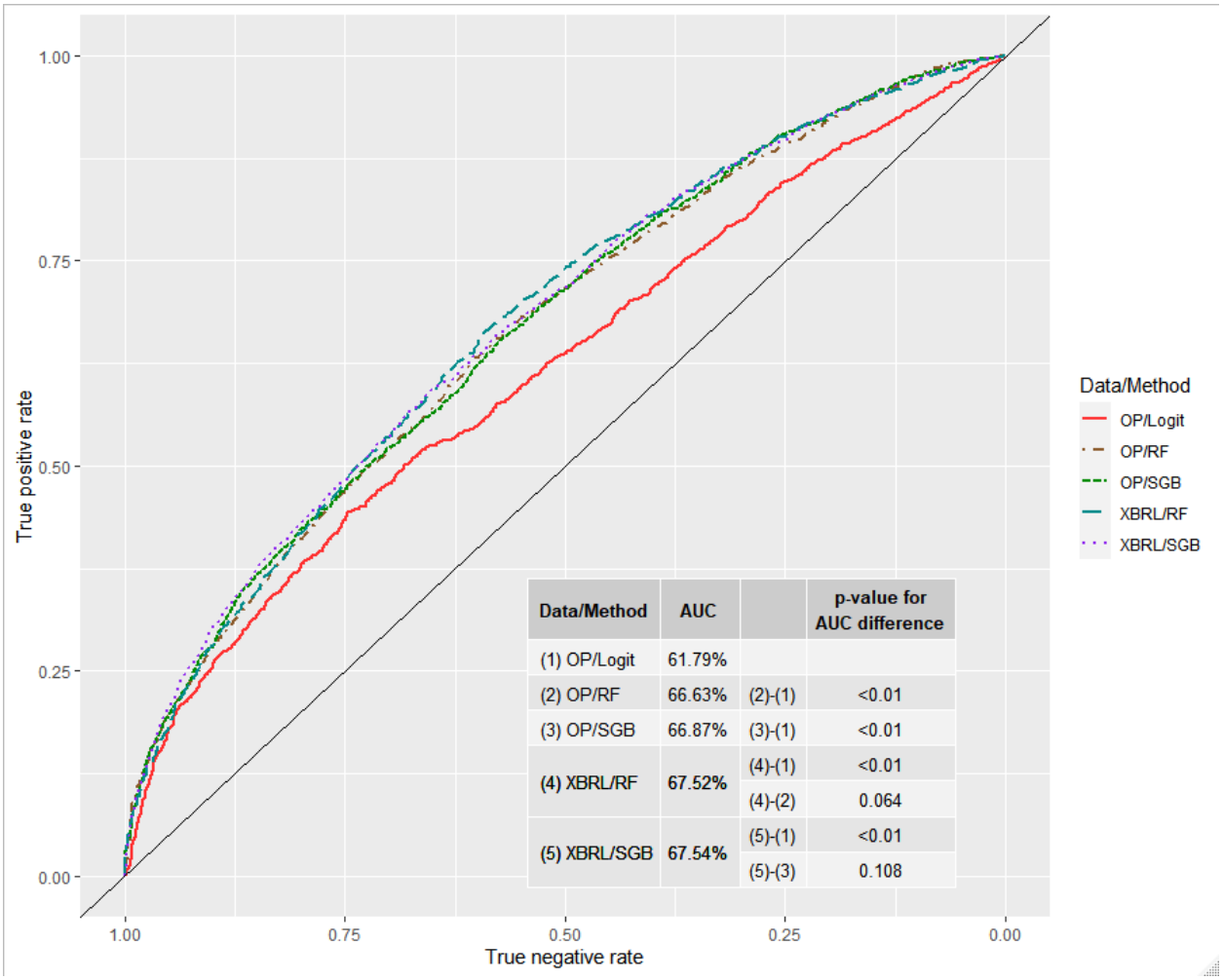



Figure 8d: Tax benefits of share-based compensation and retained earnings (stochastic gradient boosting)

Figures 8a and 8c show one-way partial dependence plots and Figures 8b and 8d show two-way partial dependence plots. In a one-way partial dependence plot, for each value of a predictor (in the x-axis), we force all observations in the training sample to assume that value for that predictor without changing any data points for other predictors, compute the forecasts using the chosen model, and average forecasts across all observations. The value of the average forecast is for the y-axis. In a two-way partial dependence plot, for each value combination of two predictors (in both the x-axis and y-axis), we force all observations in the training sample to assume the value combination for those two predictors without changing any data points for other predictors, compute the forecasts using the chosen model, and average forecasts across all observations. The value of the average forecast is coded by color.

**Figure 9 Comparison of Out-of-sample AUC for Different Data/Method Combinations**




| Data/Method | AUC | | p-value for AUC difference |
|---|---|---|---|
| (1) OP/Logit | 61.79% | | |
| (2) OP/RF | 66.63% | (2)-(1) | <0.01 |
| (3) OP/SGB | 66.87% | (3)-(1) | <0.01 |
| **(4) XBRL/RF** | **67.52%** | (4)-(1) | <0.01 |
| | | (4)-(2) | 0.064 |
| **(5) XBRL/SGB** | **67.54%** | (5)-(1) | <0.01 |
| | | (5)-(3) | 0.108 |

This figure compares out-of-sample AUC for different data/method combinations. The data consist of Ou and Penman's (1989) variables (OP) and our XBRL items (XBRL). The employed methods are logistic regression (Logit), random forests (RF), and stochastic gradient boosting (SGB). For each comparison between two data/method combinations (e.g., OP/Logit vs. XBRL/RF), the bootstrap *p*-value is the proportion of 10,000 bootstrap AUC differences that are below zero. We use a bootstrap sample with the same size as the original sample to compute the bootstrap AUC for each combination and the AUC difference between the two combinations.

## Figure 10 Abnormal Returns by Year

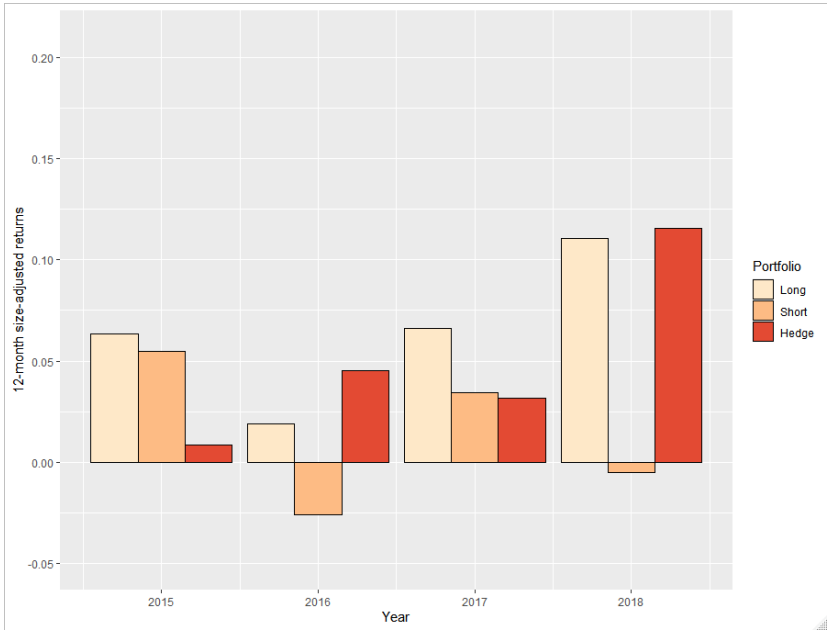



Figure 10a Random forests with $\widehat{Pr}$ > 0.5 and <0.5


Figure 10b Random forests with $\widehat{Pr}$ > 0.6 and <0.4


Figure 10c Stochastic gradient boosting with $\widehat{Pr}$ > 0.5 and <0.5

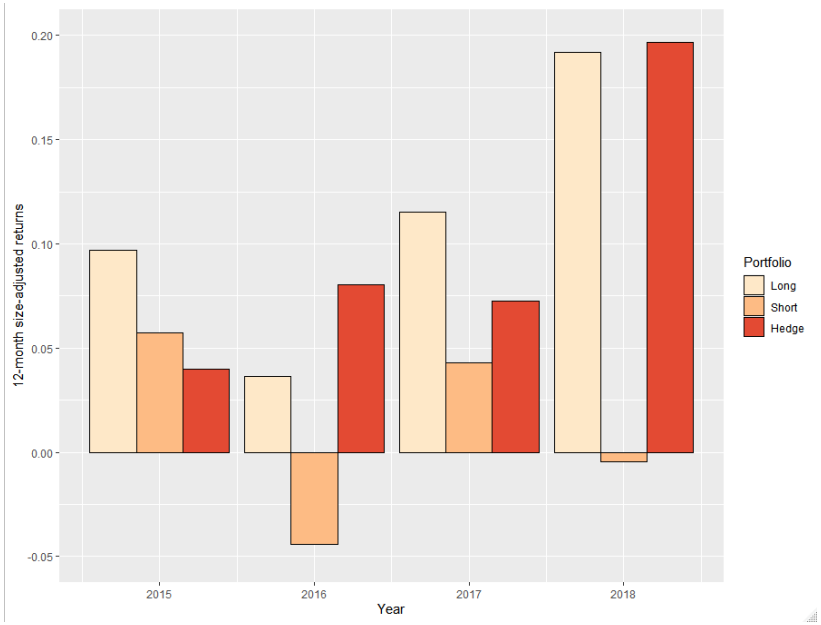



Figure 10d Stochastic gradient boosting with $\widehat{Pr}$ > 0.6 and <0.5

---

Figure 10a presents 12-month size-adjusted returns from taking long positions in stocks with $\widehat{Pr} > 0.5$ and short positions in stocks with $\widehat{Pr} \leq 0.5$ using random forests. Figure 10b presents 12-month size-adjusted returns from taking long positions in stocks with $\widehat{Pr} \geq 0.6$ and short positions in stocks with $\widehat{Pr} \leq 0.4$ using random forests. Figure 10c shows 12-month size-adjusted returns from taking long positions in stocks with $\widehat{Pr} > 0.5$ and short positions in stocks with $\widehat{Pr} \leq 0.5$ using stochastic gradient boosting. Figure 10d shows 12-month size-adjusted returns from taking long positions in stocks with $\widehat{Pr} \geq 0.6$ and short positions in stocks with $\widehat{Pr} \leq 0.4$ using stochastic gradient boosting. The size-adjusted return for a hedge portfolio is the difference between the size-adjusted abnormal returns of the long and short positions.

**Figure 11 Comparison of Abnormal Returns for Different Data/Method Combinations**

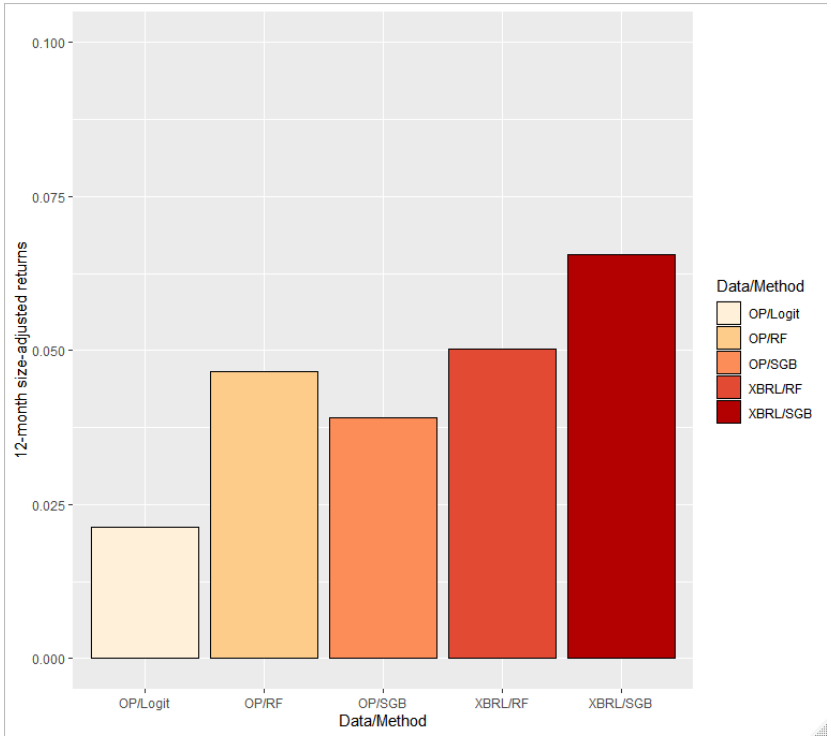



Figure 11a Abnormal returns with $\widehat{Pr}$ > 0.5 and <0.5

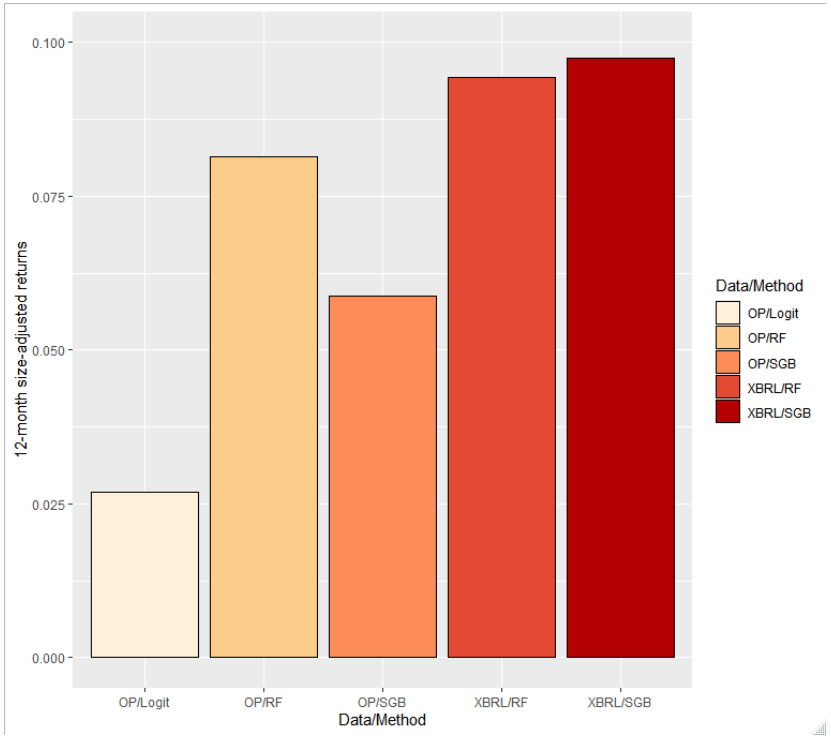



Figure 11b Abnormal returns with $\widehat{Pr}$ > 0.6 and <0.4

Figure 11a compares mean 12-month size-adjusted returns for different data/method combinations from taking long positions in stocks with $\widehat{Pr} > 0.5$ and short positions in stocks with $\widehat{Pr} \leq 0.5$. Figure 11b compares mean 12-month size-adjusted returns for different data/method combinations from taking long positions in stocks with $\widehat{Pr} \geq 0.6$ and short positions in stocks with $\widehat{Pr} \leq 0.4$. The data consist of Ou and Penman's (1989) variables (OP) and our XBRL items (XBRL). The employed methods are logistic regression (Logit), random forests (RF), and stochastic gradient boosting (SGB).

## Table 1 Sample Selection

| | Number of submissions |
|---|---|
| (1) XBRL filings for 10-K and 10-K/A between June 15, 2012 and March 31, 2018 that can be matched to pro forma earnings from I/B/E/S | 10,073 |
| (2) Requiring stock price data available from CRSP | 8,381 |
| (3) Requiring non-zero total assets | 8,358 |
| (4) Retaining the most recent XBRL filings as of the portfolio formation date | 8,149 |

This table shows the sample selection procedure. We start our sample with XBRL filings for 10-K and 10-K/A between June 15, 2012 and March 31, 2018 that can be matched to pro forma earnings from I/B/E/S. To ensure compliance with mandatory footnote disclosure in the XBRL format, we require that an XBRL filing has a reporting period ending on or after June 15, 2012. We require a filing to have stock price data available from CRSP and non-zero total assets. Exploiting the point-in-time nature of XBRL-tagged financial data, we only retain the most recent filings as of the portfolio formation date.

## Table 2 Sample Distribution

**Panel A: XBRL filings by calendar period**

| Calendar Period | Number of submissions |
|---|---|
| 2012Q3-2012Q4 | 119 |
| 2013Q1-2013Q4 | 1,206 |
| 2014Q1-2014Q4 | 1,304 |
| 2015Q1-2015Q4 | 1,375 |
| 2016Q1-2016Q4 | 1,371 |
| 2017Q1-2017Q4 | 1,460 |
| 2018Q1 | 1,314 |
| Total | 8,149 |

**Panel B: XBRL filings by industry**

| Industry | Number of submisssions |
|---|---|
| Food Products | 149 |
| Beer & Liquor | 23 |
| Tobacco Products | 21 |
| Recreation | 79 |
| Printing and Publishing | 66 |
| Consumer Goods | 121 |
| Apparel | 94 |
| Healthcare, Medical Equipment, Pharmaceutical Products | 588 |
| Chemicals | 205 |
| Textiles | 26 |
| Construction and Construction Materials | 210 |
| Steel Works | 102 |
| Fabricated Products and Machinery | 303 |
| Electrical Equipment | 87 |
| Automobiles and Trucks | 170 |
| Aircraft, ships, and railroad equipment | 56 |
| Precious Metals, Non-Metallic, and Industrial Metal Mining | 92 |
| Coal | 10 |
| Petroleum and Natural Gas | 390 |
| Utilities | 248 |
| Communication | 165 |
| Personal and Business Services | 1,353 |
| Business Equipment | 995 |
| Business Supplies and Shipping Containers | 126 |
| Transportation | 194 |
| Wholesale | 202 |
| Retail | 374 |
| Restaurants, Hotels, Motels | 222 |
| Banking, Insurance, Real Estate, Trading | 1,344 |
| Other | 134 |
| Total | 8,149 |

Panel A shows the number of XBRL filings by year for the final sample of 8,149 filings. Panel B provides the number of XBRL filings by Fama-French 30-industry classification.

## Table 3 Summary Statistics

**Panel A: Number of predictors by financial statement category**

| Financial Statement Category | Number of Current Predictors | Number of Lagged Predictors | Number of %Δ Predictors | Total |
|---|---|---|---|---|
| Balance Sheet | 639 | 639 | 639 | 1,917 |
| Income Statement | 740 | 740 | 740 | 2,220 |
| Cash Flow Statement | 87 | 87 | 87 | 261 |
| Comprehensive Income Statement | 131 | 131 | 131 | 393 |
| Shareholders' Equity Statement | 587 | 587 | 587 | 1,761 |
| Footnotes | 2,443 | 2,443 | 2,443 | 7,329 |
| Total | 4,627 | 4,627 | 4,627 | 13,881 |

**Panel B: Top 10 most populated current predictors (i.e., non-zero values) from Balance Sheet**

| Predictor | Frequency | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| Assets $_t$ | 8,149 | 17,281.54 | 648.90 | 2,247.50 | 7,739.48 |
| LiabilitiesAndStockholdersEquity $_t$ | 8,138 | 17,290.44 | 648.93 | 2,245.22 | 7,730.65 |
| RetainedEarningsAccumulatedDeficit $_t$ | 7,882 | 2,672.38 | -71.66 | 195.64 | 1,402.41 |
| CashAndCashEquivalentsAtCarryingValue $_t$ | 7,836 | 762.66 | 43.25 | 133.65 | 479.34 |
| PropertyPlantAndEquipmentNet $_t$ | 7,641 | 2,638.33 | 47.47 | 219.90 | 1,035.82 |
| StockholdersEquity $_t$ | 7,635 | 3,699.93 | 219.88 | 702.92 | 2,282.85 |
| AccumulatedDepreciationDepletionAndAmortizationPropertyPlantAndEquipment $_t$ | 7,349 | 2,052.62 | 50.78 | 238.90 | 956.60 |
| CommonStockSharesAuthorized $_t$ | 7,066 | 142,369.72 | 100.00 | 210.00 | 500.00 |
| AccumulatedOtherComprehensiveIncomeLossNetOfTax $_t$ | 7,022 | -297.63 | -107.91 | -8.58 | -0.01 |
| PropertyPlantAndEquipmentGross $_t$ | 6,935 | 4,659.43 | 108.59 | 485.09 | 1,982.24 |

**Panel C: Top 10 most populated current predictors (i.e., non-zero values) from Income Statement**

| Predictor | Frequency | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| IncomeTaxExpenseBenefit $_t$ | 7,903 | 153.03 | 0.75 | 18.73 | 96.02 |
| WeightedAverageNumberOfSharesOutstandingBasic $_t$ | 7,312 | 316.50 | 32.75 | 66.16 | 165.82 |
| WeightedAverageNumberOfDilutedSharesOutstanding $_t$ | 7,292 | 306.07 | 33.23 | 68.00 | 169.64 |
| NetIncomeLoss $_t$ | 7,280 | 314.61 | -0.77 | 30.80 | 164.65 |
| EarningsPerShareBasic $_t$ | 7,227 | 1.28 | 0.10 | 0.75 | 2.02 |
| EarningsPerShareDiluted $_t$ | 7,210 | 1.20 | 0.08 | 0.70 | 1.92 |
| OperatingIncomeLoss $_t$ | 6,669 | 475.35 | 3.61 | 66.28 | 310.50 |
| AmortizationOfIntangibleAssets $_t$ | 5,890 | 76.53 | 2.80 | 11.84 | 37.86 |
| InterestExpense $_t$ | 5,672 | 162.74 | 5.99 | 29.78 | 107.64 |
| IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethodInvestments $_t$ | 4,794 | 493.18 | 4.64 | 71.61 | 329.66 |

**Panel D: Top 10 most populated current predictors (i.e., non-zero values) from Cash Flow Statement**

| Predictor | Frequency | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| DeferredIncomeTaxExpenseBenefit $_t$ | 7,252 | 105.66 | -12.28 | -0.04 | 11.87 |
| CashAndCashEquivalentsPeriodIncreaseDecrease $_t$ | 7,166 | 20.01 | -26.00 | 2.86 | 51.53 |
| ShareBasedCompensation $_t$ | 6,939 | 45.94 | 4.88 | 13.44 | 36.00 |
| PaymentsToAcquirePropertyPlantAndEquipment $_t$ | 6,041 | 307.97 | 9.89 | 39.36 | 148.60 |
| NetCashProvidedByUsedInInvestingActivities $_t$ | 5,672 | -705.14 | -507.93 | -115.04 | -20.38 |
| NetCashProvidedByUsedInOperatingActivities $_t$ | 5,647 | 997.09 | 44.45 | 175.49 | 660.75 |
| NetCashProvidedByUsedInFinancingActivities $_t$ | 5,626 | -267.87 | -194.89 | -15.86 | 59.51 |
| IncreaseDecreaseInAccountsReceivable $_t$ | 5,063 | 36.00 | -2.68 | 5.30 | 28.23 |
| Depreciation $_t$ | 4,963 | 201.04 | 10.33 | 34.20 | 113.19 |
| DepreciationDepletionAndAmortization $_t$ | 4,943 | 351.03 | 16.66 | 59.63 | 197.85 |

**Panel E: Top 10 most populated current predictors (i.e., non-zero values) from Comprehensive Income Statement**

| Predictor | Frequency | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| ComprehensiveIncomeNetOfTax $_t$ | 6,899 | 450.00 | -1.73 | 56.98 | 276.68 |
| OtherComprehensiveIncomeLossNetOfTax $_t$ | 4,616 | -27.54 | -30.69 | -0.60 | 10.00 |
| OtherComprehensiveIncomeLossForeignCurrencyTransactionAndTranslationAdjustmentNetOfTax $_t$ | 3,554 | -47.83 | -25.10 | -1.23 | 1.90 |
| ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest $_t$ | 3,206 | 769.20 | 14.39 | 139.37 | 584.73 |
| OtherComprehensiveIncomeLossNetOfTaxPortionAttributableToParent $_t$ | 2,384 | -33.09 | -18.90 | -0.42 | 6.62 |
| OtherComprehensiveIncomeLossPensionAndOtherPostretirementBenefitPlansAdjustmentNetOfTax $_t$ | 2,235 | -13.37 | -9.37 | -0.02 | 9.00 |
| ComprehensiveIncomeNetOfTaxAttributableToNoncontrollingInterest $_t$ | 2,209 | 38.64 | -0.12 | 2.20 | 20.00 |
| OtherComprehensiveIncomeUnrealizedHoldingGainLossOnSecuritiesArisingDuringPeriodNetOfTax $_t$ | 2,038 | 3.94 | -0.60 | 0.00 | 1.00 |
| OtherComprehensiveIncomeUnrealizedGainLossOnDerivativesArisingDuringPeriodNetOfTax $_t$ | 1,666 | 0.96 | -2.20 | 0.06 | 2.84 |
| OtherComprehensiveIncomeLossDerivativesQualifyingAsHedgesNetOfTax $_t$ | 1,491 | -0.01 | -2.00 | 0.20 | 3.67 |

**Panel F: Top 10 most populated current predictors (i.e., non-zero values) from Shareholders' Equity Statement**

| Predictor | Frequency | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| CommonStockSharesOutstanding $_t$ | 5,529 | 97,794.45 | 31.99 | 60.39 | 146.21 |
| AdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriodRecognitionValue $_t$ | 4,786 | 39.81 | 4.77 | 13.00 | 31.90 |
| TreasuryStockValue $_t$ | 4,087 | 2,296.51 | 22.90 | 190.00 | 1,107.70 |
| StockIssuedDuringPeriodSharesStockOptionsExercised $_t$ | 3,854 | 413.75 | 0.10 | 0.44 | 1.30 |
| StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest $_t$ | 3,782 | 6,379.97 | 490.33 | 1,378.19 | 4,630.00 |
| StockIssuedDuringPeriodValueStockOptionsExercised $_t$ | 2,774 | 17.33 | 0.63 | 3.18 | 11.70 |
| CommonStockDividendsPerShareDeclared $_t$ | 2,609 | 0.98 | 0.30 | 0.66 | 1.28 |
| TreasuryStockValueAcquiredCostMethod $_t$ | 2,319 | 1,213.40 | 7.47 | 58.95 | 319.33 |
| StockIssuedDuringPeriodValueShareBasedCompensation $_t$ | 2,120 | 47.01 | 1.00 | 6.97 | 29.26 |
| DividendsCommonStockCash $_t$ | 2,035 | 316.86 | 17.02 | 62.93 | 198.00 |

**Panel G: Top 10 most populated current predictors (i.e., non-zero values) from Footnotes**

| Predictor | Frequency | Mean | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| OperatingLeasesFutureMinimumPaymentsDueInTwoYears $_t$ | 7,091 | 64.62 | 4.10 | 12.91 | 42.96 |
| OperatingLeasesFutureMinimumPaymentsDueCurrent $_t$ | 7,062 | 73.22 | 4.87 | 15.46 | 50.29 |
| OperatingLeasesFutureMinimumPaymentsDueInThreeYears $_t$ | 7,060 | 143.75 | 3.31 | 10.60 | 35.33 |
| OperatingLeasesFutureMinimumPaymentsDueInFourYears $_t$ | 6,935 | 46.65 | 2.68 | 8.80 | 29.00 |
| OperatingLeasesFutureMinimumPaymentsDueInFiveYears $_t$ | 6,570 | 40.10 | 2.30 | 7.48 | 25.00 |
| CurrentStateAndLocalTaxExpenseBenefit $_t$ | 6,371 | 14.23 | 0.17 | 1.59 | 7.20 |
| CurrentIncomeTaxExpenseBenefit $_t$ | 6,347 | 285.83 | 2.55 | 20.40 | 90.52 |
| DeferredFederalIncomeTaxExpenseBenefit $_t$ | 6,310 | 11.73 | -8.69 | 0.28 | 13.00 |
| OperatingLeasesFutureMinimumPaymentsDue $_t$ | 6,309 | 428.10 | 20.96 | 70.00 | 254.30 |
| OperatingLeasesFutureMinimumPaymentsDueThereafter $_t$ | 6,288 | 202.01 | 5.87 | 25.00 | 97.00 |

Panel A shows the number of current predictors, lagged predictors, and percentage changes by financial statement category. Panels B, C, D, E, F, and G provide lists of top 10 most populated (i.e., non-zero) current predictors from balance sheet, income statements, cash flow statements, comprehensive income statements, shareholders' equity statements, and footnotes, respectively, and descriptive statistics for the predictor values. Frequency counts the number of XBRL filings with a non-zero predictor value. Except for per share items, all predictor values are in millions.

## Table 4 Parameters for Machine Learning

| Parameters | Random forests | Stochastic gradient boosting |
|---|---|---|
| # of variables ($k$) | From 110 to 120 | |
| # of trees ($m$) | 500, 600, 700,... 2000 | 500, 600, 700,…, 2000 |
| Learning rate ($\rho$) | | 0.005, 0.01, 0.05 |
| Tree depth ($L$) | | 1, 2, 3, 4 |
| Min. # of obs. in a leaf ($b$) | 1, 2, 3, 4 | 10 |
| Bagging | 0.5 | 0.5 |

This table presents the parameter values considered in training the respective machine learning model. # of variables ($k$) is the number of variables (i.e., predictors) to be randomly selected when forming a split in a tree. # of trees ($m$) is the number of trees to be grown. Learning rate ($\rho$) is the extent to which each tree iteration contributes to the base tree. Tree depth ($L$) is the maximum depth of each tree. Min. # of obs. in a leaf ($b$) is the minimum number of observations in the terminal nodes of each tree. Bagging is the fraction of observations to be randomly selected to grow a tree.

## Table 5 Summary of Out-of-sample Prediction Performance

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ |
| Number of observations | 5,520 | 3,338 | 5,520 | 3,649 |
| Number of earnings increases | 2,552 | 1,362 | 2,552 | 1,547 |
| Number of earnings decreases | 2,968 | 1,976 | 2,968 | 2,102 |
| % correctly predicted | 61.90 | 67.50 | 62.26 | 66.90 |
| % of predicted increases that are actual earnings increases | 60.10 | 65.64 | 60.05 | 64.50 |
| AUC (%) | 67.52 | 68.62 | 67.54 | 68.66 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |

This table presents a summary of prediction performance using different probability cutoffs. The full prediction sample consists of 5,520 firm-year observations from 2015 to 2018. For each method of random forest and stochastic gradient boosting, two sets of probability thresholds are considered. In the first set of probability thresholds, $\widehat{Pr} > 0.5$ and $\widehat{Pr} \leq 0.5$, we assign stocks with predicted probability of an increase in next year's earnings greater than (less than or equal to) 0.5 to the long (short) position. In the second set of probability thresholds, $\widehat{Pr} \geq 0.6$ and $\widehat{Pr} \leq 0.4$, we assign stocks with predicted probability of an increase in next year's earnings greater than or equal to 0.6 (less than or equal to 0.4) to the long (short) position. The AUC does not depend on the thresholds, but varies with the sample and the model. The bootstrap *p*-value is the proportion of 10,000 bootstrap AUCs that are below 50%. We use a bootstrap sample with the same size as the original sample to compute each bootstrap AUC.

# Table 6 Importance of Predictors

## Panel A: Top 10 most important predictors

| Random forests | Stochastic gradient boosting |
|---|---|
| OperatingIncomeLoss $_{t-1}$ | RetainedEarningsAccumulatedDeficit $_{t-1}$ |
| NetIncomeLoss $_{t}$ | %ΔLiabilitiesCurrent |
| ComprehensiveIncomeNetOfTax $_{t-1}$ | EarningsPerShareBasicAndDiluted $_{t}$ |
| ComprehensiveIncomeNetOfTax $_{t}$ | OperatingIncomeLoss $_{t-1}$ |
| OperatingIncomeLoss $_{t}$ | EmployeeServiceShareBasedCompensationTaxBenefitFromCompensationExpense $_{t}$ |
| NetIncomeLoss $_{t-1}$ | IncomeTaxReconciliationChangeInDeferredTaxAssetsValuationAllowance $_{t}$ |
| EarningsPerShareBasic $_{t-1}$ | NetIncomeLoss $_{t}$ |
| EarningsPerShareDiluted $_{t-1}$ | %ΔStockholdersEquity |
| EarningsPerShareDiluted $_{t}$ | TreasuryStockValue $_{t-1}$ |
| IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethod Investments $_{t-1}$ | ShortTermInvestments $_{t-1}$ |

## Panel B: Top 10 most important predictors by financial statement category (random forests)

| Balance Sheet | Cash Flow Statement |
|---|---|
| RetainedEarningsAccumulatedDeficit $_{t}$ | NetCashProvidedByUsedInOperatingActivities $_{t-1}$ |
| RetainedEarningsAccumulatedDeficit $_{t-1}$ | NetCashProvidedByUsedInOperatingActivities $_{t}$ |
| %ΔLiabilitiesCurrent | ShareBasedCompensation $_{t}$ |
| %ΔStockholdersEquity | IncomeTaxesPaid $_{t}$ |
| OtherAssetsNoncurrent $_{t-1}$ | %ΔDeferredIncomeTaxExpenseBenefit |
| LiabilitiesCurrent $_{t}$ | NetCashProvidedByUsedInOperatingActivitiesContinuingOperations $_{t}$ |
| AssetsCurrent $_{t}$ | IncreaseDecreaseInAccountsReceivable $_{t-1}$ |
| OtherAssetsNoncurrent $_{t}$ | NetCashProvidedByUsedInInvestingActivities $_{t-1}$ |
| %ΔEmployeeRelatedLiabilitiesCurrent | CashAndCashEquivalentsPeriodIncreaseDecrease $_{t}$ |
| AccumulatedOtherComprehensiveIncomeLossNetOfTax $_{t}$ | %ΔCashAndCashEquivalentsPeriodIncreaseDecrease |

| Income Statement | Comprehensive Income Statement |
|---|---|
| OperatingIncomeLoss $_{t-1}$ | ComprehensiveIncomeNetOfTax $_{t-1}$ |
| NetIncomeLoss $_{t}$ | ComprehensiveIncomeNetOfTax $_{t}$ |
| OperatingIncomeLoss $_{t}$ | %ΔComprehensiveIncomeNetOfTax |
| NetIncomeLoss $_{t-1}$ | OtherComprehensiveIncomeLossNetOfTax $_{t-1}$ |
| EarningsPerShareBasic $_{t-1}$ | ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest $_{t-1}$ |
| EarningsPerShareDiluted $_{t-1}$ | OtherComprehensiveIncomeLossNetOfTax $_{t}$ |
| EarningsPerShareDiluted $_{t}$ | %ΔOtherComprehensiveIncomeLossNetOfTax |
| IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethod Investments $_{t-1}$ | %ΔOtherComprehensiveIncomeLossForeignCurrencyTransactionAndTranslationAdjustmentNetOfTax |
| EarningsPerShareBasic $_{t}$ | OtherComprehensiveIncomeLossForeignCurrencyTransactionAndTranslationAdjustmentNetOfTax $_{t}$ |
| EarningsPerShareBasicAndDiluted $_{t}$ | ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest $_{t}$ |

| Shareholders' Equity Statement | Footnotes |
|---|---|
| AdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriodRecognitionValue $_{t}$ | DeferredTaxAssetsValuationAllowance $_{t}$ |
| CommonStockSharesOutstanding $_{t}$ | IncomeLossFromContinuingOperationsBeforeIncomeTaxesDomestic $_{t}$ |
| TreasuryStockValue $_{t-1}$ | IncomeTaxReconciliationIncomeTaxExpenseBenefitAtFederalStatutoryIncomeTaxRate $_{t}$ |
| %ΔCommonStockSharesOutstanding | IncomeLossFromContinuingOperationsBeforeIncomeTaxesDomestic $_{t-1}$ |
| TreasuryStockValue $_{t}$ | DeferredTaxAssetsValuationAllowance $_{t-1}$ |
| CommonStockSharesOutstanding $_{t-1}$ | IncomeTaxReconciliationIncomeTaxExpenseBenefitAtFederalStatutoryIncomeTaxRate $_{t-1}$ |
| AdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriodRecognitionValue $_{t-1}$ | CurrentIncomeTaxExpenseBenefit $_{t}$ |
| %ΔAdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriodRecognitionValue | CurrentFederalTaxExpenseBenefit $_{t}$ |
| %ΔTreasuryStockValue | CurrentFederalTaxExpenseBenefit $_{t-1}$ |
| %ΔStockholdersEquityIncludingPortionAttributableToNoncontrollingInterest | CurrentIncomeTaxExpenseBenefit $_{t-1}$ |

**Panel C: Top 10 most important predictors by financial statement category (stochastic gradient boosting)**

| Balance Sheet | Cash Flow Statement |
|---|---|
| RetainedEarningsAccumulatedDeficit $_{t-1}$ | PaymentsForRepurchaseOfCommonStock $_{t-1}$ |
| %ΔLiabilitiesCurrent | %ΔShareBasedCompensation |
| %ΔStockholdersEquity | PaymentsForRepurchaseOfCommonStock $_{t}$ |
| ShortTermInvestments $_{t-1}$ | EffectOfExchangeRateOnCashAndCashEquivalents $_{t}$ |
| %ΔEmployeeRelatedLiabilitiesCurrent | IncomeTaxesPaid $_{t}$ |
| AccountsPayableCurrent $_{t-1}$ | ProceedsFromSaleOfAvailableForSaleSecurities $_{t-1}$ |
| AccruedIncomeTaxesCurrent $_{t-1}$ | NetCashProvidedByUsedInInvestingActivities $_{t-1}$ |
| %ΔAccruedIncomeTaxesCurrent | %ΔIncomeTaxesPaid |
| LongTermDebtAndCapitalLeaseObligations $_{t-1}$ | NetCashProvidedByUsedInInvestingActivities $_{t}$ |
| TradingSecurities $_{t-1}$ | ProceedsFromInsuranceSettlementInvestingActivities $_{t-1}$ |

| Income Statement | Comprehensive Income Statement |
|---|---|
| EarningsPerShareBasicAndDiluted $_{t}$ | ComprehensiveIncomeNetOfTax $_{t-1}$ |
| OperatingIncomeLoss $_{t-1}$ | OtherComprehensiveIncomeLossForeignCurrencyTransactionAndTranslationAdjustmentNetOfTax $_{t}$ |
| NetIncomeLoss $_{t}$ | %ΔOtherComprehensiveIncomeLossNetOfTaxPortionAttributableToParent |
| %ΔIncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethodInvestments | OtherComprehensiveIncomeLossNetOfTax $_{t}$ |
| %ΔNetIncomeLoss | OtherComprehensiveIncomeLossNetOfTaxPortionAttributableToParent $_{t}$ |
| BusinessCombinationAcquisitionRelatedCosts $_{t-1}$ | %ΔOtherComprehensiveIncomeUnrealizedHoldingGainLossOnSecuritiesArisingDuringPeriodTax |
| EarningsPerShareBasic $_{t-1}$ | OtherComprehensiveIncomeUnrealizedHoldingGainLossOnSecuritiesArisingDuringPeriodTax $_{t-1}$ |
| ProfitLoss $_{t}$ | OtherComprehensiveIncomeLossDerivativesQualifyingAsHedgesNetOfTax $_{t-1}$ |
| AntidilutiveSecuritiesExcludedFromComputationOfEarningsPerShareAmount $_{t}$ | OtherComprehensiveIncomeUnrealizedHoldingGainLossOnSecuritiesArisingDuringPeriodNetOfTax $_{t-1}$ |
| NetIncomeLossAvailableToCommonStockholdersBasic $_{t}$ | OtherComprehensiveIncomeLossPensionAndOtherPostretirementBenefitPlansNetUnamortizedGainLossArisingDuringPeriodBeforeTax $_{t-1}$ |

| Shareholders' Equity Statement | Footnotes |
|---|---|
| TreasuryStockValue $_{t-1}$ | EmployeeServiceShareBasedCompensationTaxBenefitFromCompensationExpense $_{t}$ |
| StockIssuedDuringPeriodValueNewIssues $_{t}$ | IncomeTaxReconciliationChangeInDeferredTaxAssetsValuationAllowance $_{t}$ |
| StockIssuedDuringPeriodValueNewIssues $_{t-1}$ | %ΔAllocatedShareBasedCompensationExpense |
| StockholdersEquityOther $_{t-1}$ | UndistributedEarningsOfForeignSubsidiaries $_{t}$ |
| StockIssuedDuringPeriodValueStockOptionsExercised $_{t}$ | CurrentForeignTaxExpenseBenefit $_{t}$ |
| %ΔStockIssuedDuringPeriodValueNewIssues | %ΔShareBasedCompensationArrangementByShareBasedPaymentAwardOptionsOutstandingNumber |
| %ΔDividendsCommonStockCash | %ΔDeferredTaxAssetsNetNoncurrent |
| TreasuryStockValue $_{t}$ | %ΔCapitalLeasesLesseeBalanceSheetAssetsByMajorClassAccumulatedDeprecation |
| %ΔCommonStockDividendsPerShareDeclared | %ΔShareBasedCompensationArrangementByShareBasedPaymentAwardOptionsOutstandingWeightedAverageExercisePrice |
| %ΔTreasuryStockValue | UnrecognizedTaxBenefitsDecreasesResultingFromSettlementsWithTaxingAuthorities $_{t}$ |

Panel A provides a list of top 10 most important predictors for random forests and stochastic gradient boosting. Panel B presents a list of top 10 most important predictors by financial statement category for random forests. Panel C presents a list of top 10 most important predictors by financial statement category for stochastic gradient boosting. Each predictor is classified into balance sheet, cash flow statement, income statement, comprehensive income statement, shareholders' equity statement, or footnotes. Importance of a predictor is computed as the decrease in the AUC when that variable is randomly shuffled.

## Table 7 Additional Analyses for Out-of-sample Prediction Performance

**Panel A: Using detailed financial data from Compustat**

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ |
| AUC (%) | 67.50 | 69.40 | 67.39 | 68.66 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |

**Panel B: Temporal changes in data quality**

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.6$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.6$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ |
| Period | Early | Late | Early | Late |
| AUC (%) | 66.96 | 70.88 | 68.86 | 69.98 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |
| Bootstrap *p*-value for AUC difference | 0.049 | | 0.307 | |

**Panel C: Partition on firm-level data quality**

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.6$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.6$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ |
| Data quality | Low | High | Low | High |
| AUC (%) | 65.99 | 70.82 | 64.70 | 71.83 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |
| Bootstrap *p*-value for AUC difference | <0.01 | | <0.01 | |

Panel A presents AUCs of using Compustat as an alternative source of detailed financial information to XBRL. The bootstrap *p*-value for AUC vs. 50% is the proportion of 10,000 bootstrap AUCs that are below 50%. We use a bootstrap sample with the same size as the original sample to compute each bootstrap AUC. Panel B shows AUCs of two subsamples by period. The early period is 2015 and the late period is 2016-2018. The bootstrap *p*-value for AUC difference is the proportion of 10,000 bootstrap AUC differences that are below zero. We use a bootstrap sample with the same size as the original subsample to compute the bootstrap AUC for each subsample and the AUC difference between the two subsamples. Panel C shows AUCs of two subsamples based on a firm-level data quality measure. High (low) data quality means the proportion of custom and uncommon standard tags in a submission is below (above) the year median value.

## Table 8 Size-Adjusted Returns

| Random forests | | | | | |
|---|---|---|---|---|---|
| $\widehat{Pr}$ portfolio | $\widehat{Pr}$ values | *N* | % correctly predicted | 12-month SAR (%) | |
| 1 | $\widehat{Pr} \leq 0.1$ | 31 | 83.87 | -7.71 | |
| 2 | $0.1 < \widehat{Pr} \leq 0.2$ | 336 | 79.17 | -0.53 | |
| 3 | $0.2 < \widehat{Pr} \leq 0.3$ | 1,009 | 69.87 | 1.88 | |
| 4 | $0.3 < \widehat{Pr} \leq 0.4$ | 1,473 | 59.27 | 1.71 | |
| 5 | $0.4 < \widehat{Pr} \leq 0.5$ | 1,432 | 49.09 | 1.70 | |
| 6 | $0.5 < \widehat{Pr} \leq 0.6$ | 754 | 61.67 | 3.72 | |
| 7 | $0.6 < \widehat{Pr} \leq 0.7$ | 266 | 71.05 | 5.68 | |
| 8 | $0.7 < \widehat{Pr} \leq 0.8$ | 150 | 84.00 | 15.36 | |
| 9 | $\widehat{Pr} > 0.8$ | 69 | 92.75 | 23.90 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} > 0.5$ | | | 5.02*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.5$ | | | | |
| Perfect foresight | | | | 12.97 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} \geq 0.6$ | | | 9.43*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.4$ | | | | |
| Perfect foresight | | | | 12.85 | |
| Stochastic gradient boosting | | | | | |
| $\widehat{Pr}$ portfolio | $\widehat{Pr}$ values | *N* | % correctly predicted | 12-month SAR (%) | |
| 1 | $\widehat{Pr} \leq 0.2$ | 33 | 87.88 | 6.26 | |
| 2 | $0.2 < \widehat{Pr} \leq 0.3$ | 1,141 | 73.09 | 0.59 | |
| 3 | $0.3 < \widehat{Pr} \leq 0.4$ | 1,908 | 58.96 | 1.61 | |
| 4 | $0.4 < \widehat{Pr} \leq 0.5$ | 1,323 | 49.66 | 1.45 | |
| 5 | $0.5 < \widehat{Pr} \leq 0.6$ | 548 | 61.86 | 4.34 | |
| 6 | $0.6 < \widehat{Pr} \leq 0.7$ | 284 | 73.94 | 9.45 | |
| 7 | $0.7 < \widehat{Pr} \leq 0.8$ | 209 | 84.21 | 14.27 | |
| 8 | $\widehat{Pr} > 0.8$ | 74 | 90.54 | 10.36 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} > 0.5$ | | | 6.57*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.5$ | | | | |
| Perfect foresight | | | | 12.97 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} \geq 0.6$ | | | 9.74*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.4$ | | | | |
| Perfect foresight | | | | 13.03 | |

This table presents 12-month size-adjusted returns on portfolios based on $\widehat{Pr}$, the estimated probability of an increase in next year's earnings in the test period (2015-2018). For each portfolio, the number of observations (N), the proportion of accurate predictions (% correctly predicted), and 12-month size-adjusted return (SAR) are reported. % correctly predicted is the proportion of observations with a correct prediction using $\widehat{Pr} = 0.5$ as a cutoff. We consider two hedge portfolios. The first hedge portfolio takes a long (short) position in stocks with $\widehat{Pr} > 0.5$ ($\leq 0.5$). The second hedge portfolio takes a long (short) position in stocks with $\widehat{Pr} \geq 0.6$ ($\leq 0.4$). The resulting 12-month size-adjusted return is reported for each hedge portfolio. The *p*-values in parentheses pertain to 12-month size-adjusted abnormal returns and are calculated from a bootstrap distribution of 10,000 pseudo abnormal returns under the null hypothesis that our predictors do not have any predictive power. For each of 10,000 iterations, we randomly assign stocks to the long and short positions and calculate a pseudo 12-month size-adjusted return. The 12-month size-adjusted returns for the perfect foresight strategy are calculated from taking a long (short) position in stocks with an increase (a decrease) in next year's earnings. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

## Table 9 Size-Adjusted Returns Net of Transaction Costs

| Random forest | | | | | |
|---|---|---|---|---|---|
| $\widehat{Pr}$ portfolio | $\widehat{Pr}$ values | *N* | % correctly predicted | 12-month SAR (%) | |
| 1 | $\widehat{Pr} \leq 0.1$ | 31 | 83.87 | -8.08 | |
| 2 | $0.1 < \widehat{Pr} \leq 0.2$ | 336 | 79.17 | -1.01 | |
| 3 | $0.2 < \widehat{Pr} \leq 0.3$ | 1,009 | 69.87 | 1.41 | |
| 4 | $0.3 < \widehat{Pr} \leq 0.4$ | 1,473 | 59.27 | 1.23 | |
| 5 | $0.4 < \widehat{Pr} \leq 0.5$ | 1,432 | 49.09 | 1.15 | |
| 6 | $0.5 < \widehat{Pr} \leq 0.6$ | 754 | 61.67 | 3.08 | |
| 7 | $0.6 < \widehat{Pr} \leq 0.7$ | 266 | 71.05 | 4.82 | |
| 8 | $0.7 < \widehat{Pr} \leq 0.8$ | 150 | 84.00 | 14.32 | |
| 9 | $\widehat{Pr} > 0.8$ | 69 | 92.75 | 23.02 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} > 0.5$ | | | 4.77*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.5$ | | | | |
| Perfect foresight | | | | 12.95 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} \geq 0.6$ | | | 8.98*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.4$ | | | | |
| Perfect foresight | | | | 12.80 | |
| Stochastic gradient boosting | | | | | |
| $\widehat{Pr}$ portfolio | $\widehat{Pr}$ values | *N* | % correctly predicted | 12-month SAR (%) | |
| 1 | $\widehat{Pr} \leq 0.2$ | 33 | 87.88 | 5.68 | |
| 2 | $0.2 < \widehat{Pr} \leq 0.3$ | 1,141 | 73.09 | 0.10 | |
| 3 | $0.3 < \widehat{Pr} \leq 0.4$ | 1,908 | 58.96 | 1.13 | |
| 4 | $0.4 < \widehat{Pr} \leq 0.5$ | 1,323 | 49.66 | 0.94 | |
| 5 | $0.5 < \widehat{Pr} \leq 0.6$ | 548 | 61.86 | 3.65 | |
| 6 | $0.6 < \widehat{Pr} \leq 0.7$ | 284 | 73.94 | 8.59 | |
| 7 | $0.7 < \widehat{Pr} \leq 0.8$ | 209 | 84.21 | 13.28 | |
| 8 | $\widehat{Pr} > 0.8$ | 74 | 90.54 | 9.38 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} > 0.5$ | | | 6.25*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.5$ | | | | |
| Perfect foresight | | | | 12.95 | |
| Hedge portfolio | | | | | |
| Long | $\widehat{Pr} \geq 0.6$ | | | 9.30*** | (<0.0001) |
| Short | $\widehat{Pr} \leq 0.4$ | | | | |
| Perfect foresight | | | | 12.98 | |

This table presents 12-month size-adjusted returns net of transaction costs on portfolios based on $\widehat{Pr}$, the estimated probability of an increase in next year's earnings in the test period (2015-2018). The transaction costs are estimated as the effective bid-ask spread following Novy-Marx and Velikov (2016). For each portfolio, the number of observations (N), the proportion of accurate predictions (% correctly predicted), and 12-month size-adjusted return (SAR) are reported. % correctly predicted is the proportion of observations with a correct prediction using $\widehat{Pr} = 0.5$ as a cutoff. We consider two hedge portfolios. The first hedge portfolio takes a long (short) position in stocks with $\widehat{Pr} > 0.5$ ($\leq 0.5$). The second hedge portfolio takes a long (short) position in stocks with $\widehat{Pr} \geq 0.6$ ($\leq 0.4$). The resulting 12-month size-adjusted return is reported for each hedge portfolio. The *p*-values in parentheses pertain to 12-month size-adjusted abnormal returns and are calculated from a bootstrap distribution of 10,000 pseudo abnormal returns under the null hypothesis that our predictors do not have any predictive power. For each of 10,000 iterations, we randomly assign stocks to the long and short positions and calculate a pseudo 12-month size-adjusted return. The 12-month size-adjusted returns for the perfect foresight strategy are calculated from taking a long (short) position in stocks with an increase (a decrease) earnings in the next year. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

## Table 10 Additional Analyses For Abnormal Returns

**Panel A: Alternative earnings measures**

| | ROE $_{t+1}$ | | EBIT $_{t+1}$ | |
|---|---|---|---|---|
| | RF | SGB | RF | SGB |
| Probability thresholds | | | | |
| Long | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ |
| SAR net of transaction costs (%) | 6.46*** | 5.43*** | 8.92*** | 8.52*** |
| *p*-value | (<0.0001) | (<0.0001) | (<0.0001) | (<0.0001) |
| Perfect foresight SAR net of transaction costs (%) | 9.98 | 10.68 | 14.32 | 13.35 |

**Panel B: Excluding microcaps**

| | Random forests | Stochastic gradient boosting |
|---|---|---|
| Probability thresholds | | |
| Long | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ |
| SAR net of transaction costs (%) | 7.65*** | 7.34*** |
| *p*-value | (<0.0001) | (<0.0001) |
| Perfect foresight SAR net of transaction costs (%) | 10.69 | 10.74 |

**Panel C: Controlling for five-factors**

| Dep. Var. = | $R_{pt} - R_{Ft}$ | | | |
|---|---|---|---|---|
| | Random forests | | Stochastic gradient boosting | |
| | (1) | (2) | (3) | (4) |
| Intercept | 0.79** | 0.66* | 0.94*** | 0.69** |
| | (0.34) | (0.33) | (0.27) | (0.31) |
| $R_{Mt} - R_{Ft}$ | 0.01 | -0.06 | 0.05 | -0.04 |
| | (0.11) | (0.14) | (0.13) | (0.18) |
| $SMB_t$ | 0.33** | 0.52*** | 0.53*** | 0.67*** |
| | (0.14) | (0.16) | (0.13) | (0.17) |
| $HML_t$ | -0.54** | -0.30 | -0.56** | -0.31 |
| | (0.25) | (0.24) | (0.22) | (0.21) |
| $RMW_t$ | -0.26 | -0.10 | -0.34 | -0.22 |
| | (0.23) | (0.28) | (0.21) | (0.29) |
| $CMA_t$ | -0.32 | -0.03 | 0.09 | 0.23 |
| | (0.28) | (0.26) | (0.25) | (0.24) |
| Industry-adjusted | No | Yes | No | Yes |
| *N* | 51 | 51 | 51 | 51 |
| $R^2$ | 0.32 | 0.19 | 0.37 | 0.26 |

Panel A presents 12-month size-adjusted returns (SAR) net of transaction costs on portfolios based on $\widehat{Pr}$, the estimated probability of an increase in next year's earnings using two alternative earnings measures (ROE$_{t+1}$ and EBIT$_{t+1}$) in the test period (2015-2018). ROE is defined as net income divided by book value of equity at the fiscal year-end. EBIT is defined as earnings before interest and tax deflated by the number of common shares outstanding at the fiscal year-end. The transaction costs are estimated as the effective bid-ask spread following Novy-Marx and Velikov (2016). Panel B presents 12-month size-adjusted returns (SAR) of net of transaction costs on portfolios based on $\widehat{Pr}$, the estimated probability of an increase in next year's earnings, excluding microcaps. Microcaps are defined as those with market capitalization of less than the 20$^{th}$ percentile of NYSE market capitalization. For each method of random forest and stochastic gradient boosting, long (short) positions are taken in stocks with predicted probability of an increase in next year's earnings greater than or equal to 0.6 (less than or equal to 0.4). The *p*-values in parentheses are calculated from a bootstrap distribution of 10,000 pseudo abnormal returns under the null

hypothesis that our predictors do not have any predictive power. For each of 10,000 iterations, we randomly assign stocks to the long and short positions and calculate a pseudo 12-month size-adjusted return net of transaction costs. The 12-month size-adjusted returns net of transaction costs for the perfect foresight strategy are calculated from taking long (short) positions in stocks with an increase (a decrease) in next year's earnings. Panel C presents results from regressing excess returns of our trading strategies on Fama-French (2015) five-factors. For each method of random forest and stochastic gradient boosting, long (short) positions are taken in stocks with predicted probability of an increase in next year's earnings greater than or equal to 0.6 (less than or equal to 0.4). In columns (1) and (3), the dependent variable is the monthly hedge portfolio returns in excess of 1-month T-bill rate. In columns (2) and (4), we adjust the dependent variable by subtracting Fama-French 30 industry monthly returns for each stock in the portfolio. The explanatory variables are the market returns in excess of 1-month T-bill rate on market ($R_{Mt} - R_{Ft}$), and returns on the size (*SMB*), book-to-market (*HML*), profitability (*RMW*), and investment (*CMA*) portfolios. Returns are in percentage. Newey-West standard errors are presented in parentheses. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

## Table 11 Analysts' Earnings Forecasts

**Panel A: Predicting the direction of one-year-ahead earnings changes and abnormal returns**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Analyst forecasts | Random forests | Stochastic gradient boosting |
| Long | Forecast an increase | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} \geq 0.6$ |
| Short | Forecast a decrease | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ |
| Hedge portfolio SAR (%) | 2.49 | 9.43 | 9.74 |
| AUC (%) | 63.62 | 68.62 | 68.66 |
| | | (2) – (1) | (3) – (1) |
| Bootstrap *p*-value for AUC difference | | <0.01 | <0.01 |

**Panel B: Analysts' forecast errors**

| Dep. Var. = | Analysts' forecast errors | |
|---|---|---|
| | Random forests | Stochastic gradient boosting |
| | (1) | (2) |
| Intercept | -1.10*** | -1.17*** |
| | (0.22) | (0.23) |
| $\widehat{Pr}$ | 0.48* | 0.59** |
| | (0.26) | (0.29) |
| *log(MKVLT)* | 0.09*** | 0.10*** |
| | (0.02) | (0.02) |
| *BTM* | -0.08 | -0.08 |
| | (0.07) | (0.07) |
| *N* | 4,256 | 4,256 |
| $R^2$ | 0.006 | 0.006 |

Panel A column (1) shows the AUC for analysts' prediction of an earnings increase and 12-month size-adjusted returns (SAR) on a hedge portfolio with a long (short) position for stocks with a predicted earnings increase (decrease). We take the consensus (i.e., median) analyst forecast in the month following the portfolio formation and compare it with the realized earnings in fiscal year *t* to find out whether analysts forecast an earnings increase or decrease. The results in columns (2)-(3) are reproduced from Tables 5 and 8. Panel B presents results from regressing analysts' forecast errors on the predicted probability of increasing earnings ($\widehat{Pr}$), log of market value of equity (*log(MKVLT)*) and book-to-market value of equity (*BTM*). Analysts' forecast errors are calculated as the actual earnings in fiscal year $t + 1$ minus the consensus (i.e., median) analyst forecast in the month following the portfolio formation, scaled by the close price on the portfolio formation date, multiplied by 100. Standard errors are presented in parentheses. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

# Online Appendix
# Fundamental Analysis of Detailed Financial Data: A Machine Learning Approach

This appendix provides supplemental materials that support the manuscript "Fundamental Analysis of Detailed Financial Data: A Machine Learning Approach."

The materials include the following tables.

## Table A1: Predicting direction of earnings changes without requiring pro forma earnings

Panel A presents descriptive statistics between our sample and the sample without requiring pro forma earnings. *ROA* is the return on assets, *MKVLT* is market capitalization, *BTM* is book-to-market, and *LEV* is book leverage. Panel B presents out-of-sample prediction performance when we use the new sample and US GAAP earnings to compute the direction of earnings changes. For each method of random forest and stochastic gradient boosting, two sets of probability thresholds are considered. In the first set of probability thresholds, $\widehat{Pr} > 0.5$ and $\widehat{Pr} \leq 0.5$, we assign stocks with predicted probability of an increase in next year's earnings greater than (less than or equal to) 0.5 to the long (short) position. In the second set of probability thresholds, $\widehat{Pr} \geq 0.6$ and $\widehat{Pr} \leq 0.4$, we assign stocks with predicted probability of an increase in next year's earnings greater than or equal to 0.6 (less than or equal to 0.4) to the long (short) position. The resulting 12-month size-adjusted return (SAR) is reported for each hedge portfolio. The AUC does not depend on the thresholds, but varies with the sample and the model. The bootstrap *p*-value is the proportion of 10,000 bootstrap AUCs that are below 50%. We use a bootstrap sample with the same size as the original sample to compute each bootstrap AUC.

**Panel A: Descriptive statistics**

| | Our sample (N=8,149) | | | | Without requiring pro forma earnings (N=20,512) | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Q1 | Median | Q3 | Mean | Q1 | Median | Q3 |
| *ROA* | 0.013 | 0.002 | 0.030 | 0.065 | -0.061 | -0.031 | 0.014 | 0.056 |
| *MKVLT* | 7.692 | 6.510 | 7.645 | 8.870 | 6.707 | 5.255 | 6.745 | 8.108 |
| *BTM* | 0.499 | 0.221 | 0.398 | 0.677 | 0.576 | 0.225 | 0.454 | 0.781 |
| *LEV* | 0.230 | 0.043 | 0.203 | 0.349 | 0.211 | 0.008 | 0.139 | 0.340 |

**Panel B: Out-of-sample prediction performance**

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ |
| AUC (%) | 60.85 | 62.47 | 60.69 | 61.66 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |
| Hedge portfolio SAR (%) | 3.32 | 4.50 | 3.08 | 6.39 |

## Table A2: Dropping the %Δ variables

This table presents out-of-sample prediction performance when we drop the 4,627 percentage change predictors (as shown in Table 3 Panel A). For each method of random forest and stochastic gradient boosting, two sets of probability thresholds are considered. In the first set of probability thresholds, $\widehat{Pr} > 0.5$ and $\widehat{Pr} \leq 0.5$, we assign stocks with predicted probability of an increase in next year's earnings greater than (less than or equal to) 0.5 to the long (short) position. In the second set of probability thresholds, $\widehat{Pr} \geq 0.6$ and $\widehat{Pr} \leq 0.4$, we assign stocks with predicted probability of an increase in next year's earnings greater than or equal to 0.6 (less than or equal to 0.4) to the long (short) position. The resulting 12-month size-adjusted return (SAR) is reported for each hedge portfolio. The AUC does not depend on the thresholds, but varies with the sample and the model. The bootstrap *p*-value is the proportion of 10,000 bootstrap AUCs that are below 50%. We use a bootstrap sample with the same size as the original sample to compute each bootstrap AUC.

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ |
| AUC (%) | 68.00 | 69.50 | 67.31 | 68.72 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |
| Hedge portfolio SAR (%) | 5.13 | 9.29 | 6.07 | 10.17 |

## Table A3: Classification of tags associated with multiple financial statements

We classify 4,503 of 4,627 tags into five financial statements and footnote disclosures based on the U.S. GAAP taxonomy. The remaining 124 tags are associated with multiple financial statements. This table shows how we classify them into the financial statement categories.

| Balance Sheet |
|---|
| Cash |
| CashAndCashEquivalentsAtCarryingValue |
| RestrictedCashAndCashEquivalents |
| RestrictedCashAndCashEquivalentsAtCarryingValue |
| DisposalGroupIncludingDiscontinuedOperationCashAndCashEquivalents |
| DividendsPayableCurrentAndNoncurrent |
| TemporaryEquityCarryingAmountIncludingPortionAttributableToNoncontrollingInterests |
| RestrictedCashAndCashEquivalentsNoncurrent |

| Income Statement |
|---|
| CostOfGoodsSoldDepreciation |
| CostOfGoodsSoldAmortization |
| CostOfServicesDepreciation |
| CostOfServicesAmortization |
| InventoryWriteDown |
| ProvisionForLoanAndLeaseLosses |
| ResearchAndDevelopmentInProcess |
| DepreciationNonproduction |
| AmortizationOfAcquisitionCosts |
| AmortizationOfIntangibleAssets |
| AmortizationOfDeferredSalesCommissions |
| AmortizationOfRegulatoryAsset |
| AmortizationOfLeasedAsset |
| AmortizationOfDeferredLeasingFees |
| AmortizationOfNuclearFuelLease |
| AmortizationOfAdvanceRoyalty |
| AmortizationOfDeferredPropertyTaxes |
| AmortizationOfDeferredHedgeGains |
| OtherAmortizationOfDeferredCharges |
| OtherDepreciationAndAmortization |
| DepletionOfOilAndGasProperties |
| RecapitalizationCosts |
| CarryingCostsPropertyAndExplorationRights |
| OtherRestructuringCosts |
| RestructuringCharges |
| EnvironmentalRemediationExpense |
| ImpairmentOfLongLivedAssetsToBeDisposedOf |
| ImpairmentOfLongLivedAssetsHeldForUse |
| ImpairmentOfIntangibleAssetsIndefinitelivedExcludingGoodwill |
| GoodwillImpairmentLoss |
| ImpairmentOfRealEstate |
| ImpairmentOfOngoingProject |
| ImpairmentOfLeasehold |
| ImpairmentOfIntangibleAssetsFinitelived |
| ExplorationAbandonmentAndImpairmentExpense |
| ImpairmentOfOilAndGasProperties |
| ImpairmentLossesRelatedToRealEstatePartnerships |
| DisposalGroupNotDiscontinuedOperationLossGainOnWriteDown |

OtherAssetImpairmentCharges
AssetImpairmentCharges
AssetRetirementObligationAccretionExpense
AccretionExpense
AccretionExpenseIncludingAssetRetirementObligations
ProductWarrantyExpense
ProvisionForDoubtfulAccounts
GainLossOnSaleOfProperty
GainLossOnDispositionOfAssets
GainLossOnSaleOfPropertyPlantEquipment
GainLossOnDispositionOfIntangibleAssets
EquityMethodInvestmentRealizedGainLossOnDisposal
GainOrLossOnSaleOfStockInSubsidiary
GainLossOnSaleOfStockInSubsidiaryOrEquityMethodInvestee
GainLossOnSaleOfBusiness
GainLossOnSaleOfOtherAssets
TradingSecuritiesUnrealizedHoldingGainLoss
MarketableSecuritiesUnrealizedGainLossExcludingOtherThanTemporaryImpairments
TradingSecuritiesRealizedGainLoss
AvailableforsaleSecuritiesGrossRealizedGainLossExcludingOtherThanTemporaryImpairments
HeldtomaturitySecuritiesSoldSecurityRealizedGainLossExcludingOtherThanTemporaryImpairments
MarketableSecuritiesRealizedGainLossExcludingOtherThanTemporaryImpairments
MarketableSecuritiesGainLossExcludingOtherThanTemporaryImpairments
CostmethodInvestmentsRealizedGainLossExcludingOtherThanTemporaryImpairments
GainLossOnInvestmentsExcludingOtherThanTemporaryImpairments
GainLossOnInvestments
GainLossOnSecuritizationOfFinancialAssets
DisposalGroupNotDiscontinuedOperationGainLossOnDisposal
GainLossOnContractTermination
PublicUtilitiesAllowanceForFundsUsedDuringConstructionAdditions
ForeignCurrencyTransactionGainLossBeforeTax
AmortizationOfFinancingCosts
GainsLossesOnExtinguishmentOfDebt
IncomeLossFromEquityMethodInvestments
DiscontinuedOperationIncomeLossFromDiscontinuedOperationBeforeIncomeTax
DiscontinuedOperationTaxEffectOfDiscontinuedOperation
IncomeLossFromDiscontinuedOperationsNetOfTax
ProfitLoss
NetIncomeLossAttributableToRedeemableNoncontrollingInterest
NetIncomeLossAttributableToNoncontrollingInterest
NetIncomeLoss
IncomeLossIncludingPortionAttributableToNoncontrollingInterest
AmortizationOfLeaseIncentives
AmortizationOfMortgageServicingRightsMSRs
ProvisionForOtherCreditLosses
ProvisionForOtherLosses
ProvisionForLoanLeaseAndOtherLosses
GainLossOnSaleOfEquityInvestments
GainLossOnSaleOfDebtInvestments
GainLossOnSaleOfDerivatives
GainLossOnSaleOfMortgageLoans
MortgageServicingRightsMSRImpairmentRecovery
GainLossOnSalesOfLoansNet
AmortizationOfDeferredLoanOriginationFeesNet
GainLossOnSaleOfSecuritiesNet
GainLossOnSaleOfCapitalLeasesNet

GainLossOnSaleOfLeasedAssetsNetOperatingLeases
GainsLossesOnSalesOfInvestmentRealEstate
RealizedInvestmentGainsLosses
DeferredPolicyAcquisitionCostAmortizationExpense
AmortizationOfValueOfBusinessAcquiredVOBA
GainLossOnSalesOfAssetsAndAssetImpairmentCharges

Comprehensive Income Statement

OtherThanTemporaryImpairmentLossesInvestmentsPortionRecognizedInEarningsNet
OtherComprehensiveIncomeLossNetOfTax
ComprehensiveIncomeNetOfTaxAttributableToNoncontrollingInterest
OtherThanTemporaryImpairmentLossesInvestmentsPortionInOtherComprehensiveIncomeLossBeforeTaxPortion AttributableToParentAvailableforsaleSecurities
OtherThanTemporaryImpairmentLossesInvestmentsPortionInOtherComprehensiveIncomeLossBeforeTaxIncludingPortionAttributableToNoncontrollingInterestHeldtomaturitySecurities
OtherThanTemporaryImpairmentLossesInvestmentsPortionInOtherComprehensiveIncomeLossNetOfTaxPortion AttributableToParentAvailableforsaleSecurities
OtherThanTemporaryImpairmentLossesInvestmentsPortionInOtherComprehensiveIncomeLossNetOfTaxIncludingPortionAttributableToNoncontrollingInterestAvailableforsaleSecurities
OtherThanTemporaryImpairmentLossesInvestmentsPortionInOtherComprehensiveIncomeLossTaxPortionAttributableToParentAvailableforsaleSecurities
OtherThanTemporaryImpairmentLossesInvestmentsPortionInOtherComprehensiveIncomeLossTaxIncludingPortionAttributableToNoncontrollingInterestHeldtomaturitySecurities

Shareholders' Equity Statement

TreasuryStockValue
StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest
PreferredStockRedemptionPremium
PreferredStockRedemptionDiscount
PreferredStockSharesOutstanding
CommonStockSharesOutstanding
CommonStockDividendsPerShareDeclared

## Table A4: Using the sign of analysts' forecast errors as a proxy for the direction of earnings changes

This table presents out-of-sample prediction performance when we use the sign of analysts' forecast errors as a proxy for the direction of earnings changes. Specifically, we compare actual earnings in fiscal year $t + 1$ with the consensus analyst forecast issued in the month following the earnings release for fiscal year $t$ to define an earnings increase/decrease. For each method of random forest and stochastic gradient boosting, two sets of probability thresholds are considered. In the first set of probability thresholds, $\widehat{Pr} > 0.5$ and $\widehat{Pr} \leq 0.5$, we assign stocks with predicted probability of an increase in next year's earnings greater than (less than or equal to) 0.5 to the long (short) position. In the second set of probability thresholds, $\widehat{Pr} \geq 0.6$ and $\widehat{Pr} \leq 0.4$, we assign stocks with predicted probability of an increase in next year's earnings greater than or equal to 0.6 (less than or equal to 0.4) to the long (short) position. The resulting 12-month size-adjusted return (SAR) is reported for each hedge portfolio. The AUC does not depend on the thresholds, but varies with the sample and the model. The bootstrap *p*-value is the proportion of 10,000 bootstrap AUCs that are below 50%. We use a bootstrap sample with the same size as the original sample to compute each bootstrap AUC.

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.5$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.5$ | $\widehat{Pr} \leq 0.4$ |
| AUC (%) | 57.38 | 58.64 | 56.57 | 58.13 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |
| Hedge portfolio SAR (%) | 2.52 | 11.20 | 4.24 | 8.81 |

## Table A5: Chosen parameter values

This table shows the chosen parameter values for the two machine learning methods for each test year.

| | Random forests | | | |
|---|---|---|---|---|
| | 2015 | 2016 | 2017 | 2018 |
| # of variables ($k$) | 114 | 116 | 115 | 120 |
| # of trees ($m$) | 1,000 | 1,000 | 1,000 | 1,000 |
| Min. # of obs. In a leaf ($b$) | 1 | 3 | 2 | 3 |
| | **Stochastic gradient boosting** | | | |
| | 2015 | 2016 | 2017 | 2018 |
| # of trees ($m$) | 800 | 700 | 500 | 700 |
| Learning rate ($\rho$) | 0.005 | 0.01 | 0.01 | 0.005 |
| Tree depth ($L$) | 4 | 3 | 3 | 4 |

## Table A6: Falsification tests for temporal changes in data quality using detailed financial data from Compustat

This table shows AUCs of two subsamples by period using detailed financial data from Compustat, which do not experience the same data quality changes as XBRL documents. The early period is 2015 and the late period is 2016-2018. The bootstrap *p*-value for AUC difference is the proportion of 10,000 bootstrap AUC differences that are below zero. We use a bootstrap sample with the same size as the original subsample to compute the bootstrap AUC for each subsample and the AUC difference between the two subsamples.

| | Random forests | | Stochastic gradient boosting | |
|---|---|---|---|---|
| Probability thresholds | | | | |
| Long | $\widehat{Pr} > 0.6$ | $\widehat{Pr} \geq 0.6$ | $\widehat{Pr} > 0.6$ | $\widehat{Pr} \geq 0.6$ |
| Short | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ | $\widehat{Pr} \leq 0.4$ |
| Period | Early | Late | Early | Late |
| AUC (%) | 69.56 | 69.68 | 67.49 | 67.93 |
| Bootstrap *p*-value for AUC vs. 50% | <0.01 | <0.01 | <0.01 | <0.01 |
| Bootstrap *p*-value for AUC difference | 0.4774 | | 0.4184 | |

## Table A7: Summary Statistics of Estimated Transaction Costs

We estimate the effective bid-ask spread using a Bayesian Gibbs sampler on a generalized Roll (1984) model, as proposed by Hasbrouck (2009). The summary statistics of the estimated transaction costs for all stocks from 2009 to 2019 are shown below.

| Year | Mean | Median | Std. Dev. |
|---|---|---|---|
| 2009 | 0.00805 | 0.00454 | 0.01058 |
| 2010 | 0.00441 | 0.00269 | 0.00532 |
| 2011 | 0.00429 | 0.00272 | 0.00534 |
| 2012 | 0.00412 | 0.00242 | 0.00544 |
| 2013 | 0.00340 | 0.00211 | 0.00396 |
| 2014 | 0.00316 | 0.00210 | 0.00324 |
| 2015 | 0.00377 | 0.00246 | 0.00392 |
| 2016 | 0.00391 | 0.00244 | 0.00447 |
| 2017 | 0.00333 | 0.00208 | 0.00365 |
| 2018 | 0.00371 | 0.00245 | 0.00366 |
| 2019 | 0.00335 | 0.00197 | 0.00392 |