Journal of Accounting Research

DOI: 10.1111/1475-679X.12429
Journal of Accounting Research
Vol. 60 No. 2 May 2022
Printed in U.S.A.

CHICAGO BOOTH

Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data

XI CHEN,* YANG HA (TONY) CHO,† YIWEI DOU,† AND BARUCH LEV†

Received 30 November 2020; accepted 24 January 2022

ABSTRACT

We use machine learning methods and high-dimensional detailed financial data to predict the direction of one-year-ahead earnings changes. Our models show significant out-of-sample predictive power: the area under the receiver operating characteristics curve ranges from 67.52% to 68.66%, significantly higher than the 50% of a random guess. The annual size-adjusted returns to hedge portfolios formed based on the prediction of our models range from 5.02% to 9.74%. Our models outperform two conventional models that use logistic regressions and small sets of accounting variables, and

*Department of Technology, Operations, and Statistics, Stern School of Business, New York University; †Department of Accounting, Stern School of Business, New York University
February 2022

Accepted by Christian Leuz. An earlier version of this paper circulated under the title "Fundamental Analysis of Detailed Financial Data: A Machine Learning Approach." We benefited from the comments of an anonymous reviewer, an anonymous associate editor, Aleksander Aleszczyk, Karthik Balakrishnan, Jeremy Bertomeu, Oliver Binz, Elizabeth Blankespoor, Mark Bradshaw, John Core, Robert Holthausen, Amy Hutton, Charles M.C. Lee, E. Jin Lee, Becky Lester, Miao Liu, Joshua Livnat, Michael Minnis, Miguel Minutti-Meza, Joseph Piotroski, K. Ramesh, Joshua Ronen, Christine Tan, Daniel Taylor, Siew Hong Teoh, Chenqi Zhu, participants at the 2021 *Journal of Accounting Research Conference* and the 2021 Transatlantic Doctoral Conference, and seminar participants at Boston College, Florida International University, New York University, Rice University, UC Irvine, and Stanford University. An online appendix to this paper can be downloaded at http://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements

professional analysts' forecasts. Analyses suggest that the outperformance relative to the conventional models stems from both nonlinear predictor interactions missed by regressions and the use of more detailed financial data by machine learning.

JEL codes: C53, G12, G17, M41

Keywords: direction of earnings changes; prediction; detailed financial data; XBRL; machine learning

1. Introduction

Developing corporate earnings prediction models is of significant importance to accounting researchers and investment practitioners. However, future earnings are difficult to forecast as they are related to numerous aspects of a company in a complex manner with little guidance from theoretical literature (Lev and Gu [2016], Monahan [2018]). Previous studies often use a small set of financial predictors and regression models. The former is unlikely to capture the high dimensional aspects relevant to future earnings; the latter cannot approximate the complex relations. To push the frontier of earnings prediction, we apply machine learning methods to a large set of detailed financial data to predict the direction of one-year-ahead earnings changes. We seek to investigate (1) the out-of-sample performance of our models and (2) performance differences between our models and conventional models as well as analysts' forecasts.

We examine the direction of earnings changes for several reasons. First, it is difficult to predict the level of future earnings and the amount of earnings changes (future earnings minus known current earnings), as extant studies find 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]). Second, Freeman et al. [1982, p. 643] argue that the variability in earnings changes is too large to be compared to the variability in expected earnings changes conditional on explanatory variables. They propose to reduce the variability in earnings changes by transforming the amount to the direction of earnings changes, predicting which is more achievable. Third, forecasting the sign of earnings changes is economically meaningful and actionable as extensive research constructs portfolios based on the direction of earnings changes (Ou and Penman [1989], Wahlen and Wieland [2011]).

We use two widely accepted machine learning methods based on decision trees: random forests and stochastic gradient boosting, which have recently achieved remarkable success in real-world applications (Zhou [2012], Mullainathan and Spiess [2017], Liu [2021]). Compared with regressions, these methods have three advantages. First, they can accommodate a far more expansive list of predictors to utilize more nuanced information in detailed financial data. For example, we can estimate machine learning models when the number of predictors is even greater than the number of observations,

whereas traditional regressions break down for such a scenario. Second, the machine learning algorithms cast a wide net in their specification search to allow complex associations between high-dimensional predictors and the predicted variable. Third, these algorithms are specialized for prediction tasks, rather than explanation tasks. They offer high out-of-sample predictive performance by using the "regularization" (e.g., using a number of decision trees in random forests) to mitigate overfitting.

To obtain detailed financial data in a machine-readable format, we use financial reports filed in eXtensible Business Reporting Language (XBRL). XBRL is an extensible markup language composed of a standard list of tags ("taxonomy") to describe business and financial information. Since 2012, all U.S. public companies must have XBRL tags on quantitative amounts in financial statements and footnotes of their 10-K reports. Commercial data aggregators have very limited coverage of these XBRL-tagged detailed financial data, particularly for footnote disclosures.

Our sample is composed of over 8,000 XBRL filings from 2012 to 2018. These filings contain more than 4,000 distinct 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., $4{,}000 \times 3$ for current values, lagged values, and percentage changes). For each year in the test period, 2015–2018, 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 the next year's earnings.

To evaluate model performance, we use the area under the receiver operating characteristics (ROC) curve (AUC) and 12-month size-adjusted excess returns to the hedge portfolios formed three months after the fiscal-year end based on *Pr* in the test period. While AUC is commonly used in classification problems, the excess returns offer an economic meaning for the prediction gains.[1] We find significant out-of-sample predictability of our models using machine learning and detailed financial data, concerning the direction of the next year's earnings changes. The AUC in the test period ranges from 67.52% to 68.66%, significantly higher than the 50% of a random guess. The annual size-adjusted returns to the hedge portfolios

[1] Holthausen and Larcker [1992] use logistic regressions and the same set of financial variables as Ou and Penman [1989] to directly predict the sign of future stock returns. Recent studies also apply machine learning to small sets of variables to directly predict future stock returns (Chinco et al. [2019], Rasekhschaffe and Jones [2019], Livnat and Singh [2021]). How to leverage machine learning methods to analyze a large set of detailed financial data in XBRL documents for direct return predictions presents an opportunity for future research.

are both economically and statistically significant, ranging from 5.02% to 9.74%.

We compare our models with three benchmarks. First, following Ou and Penman [1989], we estimate logistic regressions using their 65 financial variables, which represent a "kitchen sink" approach with a large number of predictors.[2] Our models significantly outperform the Ou and Penman [1989] model, which exhibits an AUC of only 61.79% and annual size-adjusted returns of 2.48% in the test period. We investigate the source of this superior performance by applying the same machine learning methods to 65 financial variables of Ou and Penman. These hybrid models exhibit an AUC of 66.63% to 66.87% and annual size-adjusted returns of 3.97% to 4.67%. They significantly outperform the original Ou and Penman regression model and marginally underperform those we build using both machine learning and detailed financial data. The results suggest that our models' superior performance relative to the Ou and Penman [1989] model stems primarily from nonlinear predictor interactions in machine learning, which are missed by regressions, and secondarily from the use of more detailed financial information.

Second, we estimate a logistic regression using variables from the DuPont decomposition by Nissim and Penman [2001]. These variables are identified by experts with an internal structure and could improve predictive performance. However, forecasts from this model exhibit an AUC of 57.96% and annual size-adjusted returns of 1.90% in the test period, still significantly lower than those of our models. Applying our machine learning methods to the DuPont variables yields an AUC of 61.15% to 61.51% and annual size-adjusted returns of 2.12% to 2.73%, higher than the original logistic model with the DuPont variables but lower than the models using both machine learning and detailed financial data. The results suggest that both nonlinear predictor interactions and more detailed financial information than the DuPont variables contribute to our models' outperformance.

Third, we compare our models with analysts' forecasts issued in the month following the portfolio formation when all the detailed financial information in 10-Ks is available and can be incorporated by these professionals. Our models significantly outperform analysts' earnings forecasts, which exhibit an AUC of 64.71% and annual size-adjusted returns of 3.93% in the test period. The results highlight the usefulness of machine learning and detailed financial information in earnings prediction, even in the presence of professional forecasters, who may have limited ability to process detailed financial data.

Although XBRL documents offer detailed financial data, there are several reasons the data quality could be compromised, affecting our

[2] Ou and Penman [1989] 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 1,575 (302) articles on Google Scholar (Web of Science) as of January 10, 2022.

predictions. First, XBRL documents are not audited, and errors in these documents are subject to only modified liability within two years after the initial adoption (a safe harbor provision in Rule 406T; SEC [2009]). Second, firms can choose to use a custom tag called an "extension" rather than a standard tag. 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]). Third, in addition to custom tags, firms also combine tags with dimensions to report information by category (e.g., for segment reporting). Regardless of whether custom tags and tags with dimensions are necessary, these items are typically firm-specific and thus not comparable across firms in a large sample. As a result, we are unable to use them in our models.

We conduct two sets of analyses to examine whether data quality issues temper the usefulness of detailed financial data in XBRL documents. First, we use Compustat as an alternative source of detailed financial information. 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 (e.g., footnote disclosures). Using the machine learning methods, we continue to find robust predictive power for the Compustat data (with an AUC of 67.39% to 69.40% and annual size-adjusted returns of 3.95% to 10.07%), similar to XBRL-tagged data. Thus, the benefit of additional financial details relative to Compustat is offset by the presence of errors, custom tags, and tags with dimensions in XBRL documents.

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 a firm-level measure of the prevalence of tags that cannot be used in our large-sample analysis and observe worse predictive performance in the subsample with more such tags. The results suggest that errors and the prevalence of tags that cannot be used in a large sample reduce the usefulness of detailed financial data in XBRL documents.

Finally, to provide more transparency on the inner workings of our models, we estimate each variable's importance by computing the decrease in the predictive performance 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 current earnings are still leading indicators among financial numbers for future earnings. We also classify the variables into six groups (the five financial statements and footnotes) and find that in aggregate, footnote disclosures contribute most to our models' predictive power, followed by the balance sheet, income statement, and cash flow statement. Comprehensive income statement and shareholders' equity statement contribute the least. Among footnote disclosures, the list of top 10 most important variables 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 predictors.

Our study makes three contributions. First and foremost, we contribute to the earnings prediction literature by applying machine learning algorithms to a large set of detailed financial data. Several concurrent papers employ machine learning methods to forecast future earnings using a small group of financial statement variables (Gerakos and Gramacy [2013], Anand et al. [2019], Hunt et al. [2019], Cao and You [2020], Binz et al. [2021]). Our detailed financial data offer more nuanced information than the small set of variables and unleash the power of machine learning, which can accommodate a far more expansive list of predictors in a nonlinear fashion.

Second, 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. [2021]). Barth et al. [2021] 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 disclosures (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—earnings prediction using detailed financial information.

Finally, we add to the XBRL literature. 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 that is important to regulators, practitioners, and academics.[3] Previous studies find 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

[3] Richardson et al. [2010, p. 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."

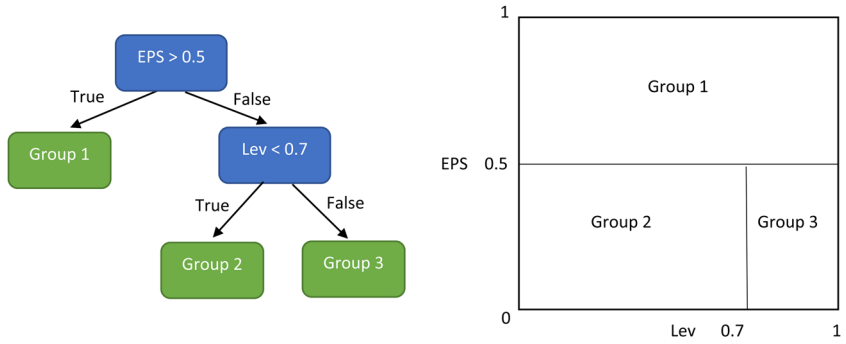


FIG. 1.—An example of a decision tree to predict an earnings increase. In this figure, the left panel presents an example of a decision tree to predict a one-year-ahead earnings increase. A blue box ("a node") represents a split and a green box ("a leaf") indicates a final partition. The right panel shows how the space of "EPS" and "Lev" is partitioned by this tree model. See subsection 2.1 for details.

research documenting errors and unnecessary extensions in XBRL filings (Debreceny et al. [2010, 2011]) and the associated adverse consequences in the capital markets (Li and Nwaeze [2015, 2018], Kirk et al. [2016]). Our paper provides direct evidence indicating that XBRL filings still inform investors' forecasts and investment decisions despite the data quality issues.

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, decision trees are built nonparametrically and designed to group observations with similar predictors. The tree "grows" in a sequence of steps. At each step, the sample 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 (i.e., an early stopping criterion) that can be selected adaptively using a validation sample (Hastie et al. [2009], Varian [2014]).

The left panel of figure 1 presents an example of a decision tree to forecast a one-year-ahead earnings increase. Suppose the tree has two predictors, "EPS" and "Lev" (i.e., earnings per share and leverage calculated as liabilities divided by total assets), and the associated threshold is the final output based on the training sample. The tree 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 at or below 0.5 are then further sorted by Lev: observations with Lev below 0.7 go into Group 2, while those with Lev at or 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. The mode of the one-year-ahead earnings increase indicator for each group ("the majority vote rule") from the training sample forms the binary forecast (i.e., $\hat{y} = 1$ for an earnings increase and 0 otherwise) for a new observation (from a validation sample or a test sample) that is assigned to that group based on its predictor value. 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 β_i denotes the mode of the earnings increase indicator for group i in the training sample and $1_{\{.\}}$ is set to 1 when the curly bracket statement is true, and 0 otherwise.[4]

A decision tree has four advantages. 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, as is necessary for traditional regression analysis. Second, a decision tree is invariant to monotonic transformations of predictors. Third, a decision tree can approximate a high degree of nonlinearities. Fourth, a decision 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 training data, as shown in figure 2. Once a random forests model is developed, each new observation (from a validation sample or a test sample) generates m predictions ($\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m$), with the final forecast being a simple average of them ($\widehat{Pr}$). Trees tend to overfit the individual bootstrap samples, which makes their individual predictions less effective. 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 each split. As a

[4] This example illustrates how to form a forecast of an earnings increase using a single decision tree model with a zero–one loss function. Nevertheless, we do not use the single decision tree model for our subsequent analyses.

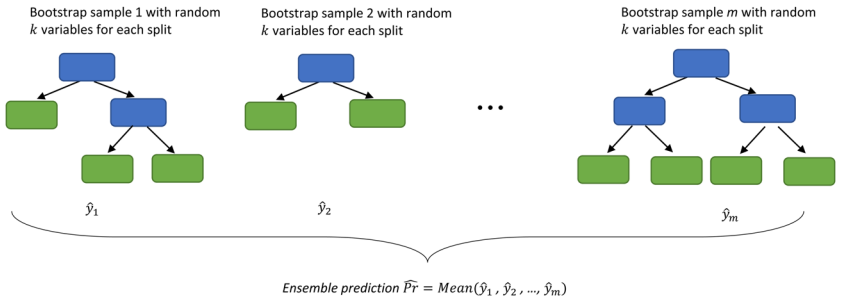


FIG. 2.—Random forests. This figure shows how an ensemble prediction of the probability of an earnings increase ($\widehat{Pr}$) is generated by random forests. See subsection 2.1 for details.

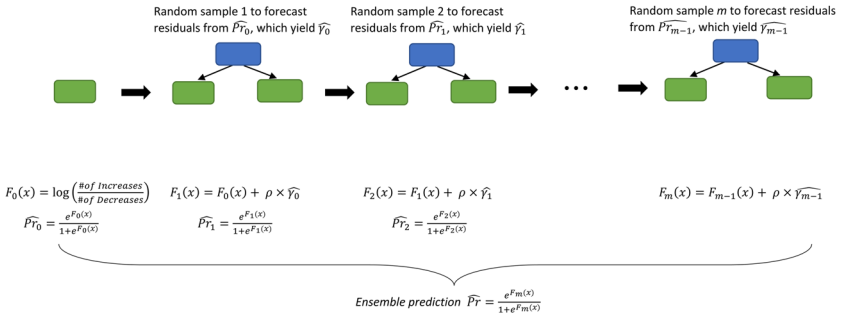


FIG. 3.—Stochastic gradient boosting. This figure shows how an ensemble prediction of the probability of an earnings increase ($\widehat{Pr}$) is generated by stochastic gradient boosting. See subsection 2.1 for details.

result, the dominant predictor may not be considered for some splits. 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. To develop a stochastic gradient boosting model using a training sample, we start by computing the initial log-odds $F_0(x) = \log(\frac{\#\text{of Increases}}{\#\text{of Decreases}})$, which can be converted to the initial prediction, $\widehat{Pr_0} = \frac{e^{F_0(x)}}{1+e^{F_0(x)}}$. We then fit a shallow tree (e.g., with depth $L = 1$) to the residuals from the initial prediction, $r_0 = y - \widehat{Pr_0}$, where $y = 1$ for an earnings increase and 0 otherwise. The residuals are converted to $\widehat{\gamma_0} = \frac{\sum r_0}{\sum[\widehat{Pr_0} \times (1-\widehat{Pr_0})]}$, where the summation is for each leaf, shrunken by a factor $\rho \in (0, 1)$ (i.e., the learning rate), and added to the initial log-odds to form an updated log-odds $F_1(x) = F_0(x) + \rho \times \widehat{\gamma_0}$ and an updated prediction $\widehat{Pr_1} = \frac{e^{F_1(x)}}{1+e^{F_1(x)}}$. 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. 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. Once a stochastic gradient boosting model is developed using a training sample, for a new observation (from a validation sample or a test sample), the output of this additive model of shallow trees yields a series of predictions (e.g., $\widehat{Pr_0}$, $\widehat{Pr_1}$, $\widehat{Pr_2}$) and the final ensemble prediction $\widehat{Pr} = \frac{e^{F_m(x)}}{1+e^{F_m(x)}}$.[5]

2.2 DETAILED FINANCIAL ACCOUNTING DATA IN XBRL

Although listed companies' financial reports are publicly available, arranging the detailed financial information in a machine-readable format for large-scale analysis is nontrivial. Although commercial data aggregators (e.g., Yahoo Finance and Compustat) collect many financial statement items, their coverage of footnote disclosures is very limited. To overcome this challenge, we take advantage of financial reports filed in XBRL.

The SEC mandate of 2009 ("Interactive Data to Improve Financial Reporting") required public companies to provide their financial reports in XBRL to the SEC and post them on their corporate Web sites.[6] The XBRL format disclosure is in addition to disclosure in the traditional electronic filing formats of ASCII or HTML. It provides a means to convert the information from human-readable formats (e.g., paper, PDF, and HTML) to a machine-readable format, comparable to the shift from paper maps to digital maps. The requirements begin for the first quarterly report for a period ending after (1) June 15, 2009, for large accelerated filers (with a public equity float over $5 billion), (2) June 15, 2010, for other large accelerated filers (with a public equity float over $700 million), and (3) 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.[7] Accordingly, 2012 is the earliest year with available XBRL-tagged financial statement and footnote items for all firms. 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 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

[5] For technical details of the two machine learning algorithms, see Breiman [2001] and Hastie et al. [2009, chapter 12] for random forests and Friedman [2002] and Hastie et al. [2009, chapter 10] for stochastic gradient boosting.

[6] 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 voluntarily filed in XBRL. See Bartley et al. [2011], Efendi et al. [2016], and Hsieh and Bedard [2018] for studies on the voluntary filing program.

[7] 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).

XBRL document. This point-in-time feature avoids issues related to data backfilling in capital market research.

XBRL U.S., a nonprofit 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 has updated it every year since 2011. The annual update occurs for reasons such as changes in accounting standards, technical corrections, and the actual use of tags.

Preparers must tag the quantitative items in the financial reports with the appropriate elements from the standard list. The appendix 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.[8] 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]).

Although 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% 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. Some tags include information by category using dimensions (e.g., for segment reporting or further disaggregation of property assets). Although the U.S.

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

GAAP taxonomy provides some standard dimensions, SEC [2016] reports that 50% of filers use custom dimensions, significantly compromising dimensional data comparability. The errors, custom tags, and tags with dimensions compromise the effective use of 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.[9] 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.[10] The Data Quality Committee of XBRL U.S. sets guidance and validation rules to prevent or detect inconsistencies or errors in XBRL documents. The committee also collects and publishes real-time errors in XBRL filings.[11] Third-party filing service companies such as XBRL Cloud also monitor for data quality issues in XBRL filings.[12] Using a sample of over 4,000 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's EDGAR Web site 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 Approach to Prediction

3.1 DATA

Table 1 shows our sample selection process. We first obtain XBRL 10-K and 10-K/A submissions between June 15, 2012 and March 31, 2018 from the SEC's Web site.[13] To take advantage of detailed footnote disclosures in XBRL, we restrict our sample to submissions with a reporting period ending on or after June 15, 2012. Recent studies demonstrate that earnings

[9] 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).

[10] 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.

[11] The errors can be found at https://xbrl.us/data-quality/filing-results/.

[12] See https://edgardashboard.xbrlcloud.com/edgar-dashboard/.

[13] 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 and Ou and Penman [1989].

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 nonzero total assets	8,358
(4) Retaining the most recent XBRL filings as of three months after the fiscal year end	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 nonzero total assets. Exploiting the point-in-time nature of XBRL-tagged financial data, we only retain the most recent filings as of three months after the fiscal year end.

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 I/B/E/S-reported EPS to compute annual earnings changes. After merging XBRL documents with pro forma earnings from I/B/E/S, we obtain 10,073 submissions.[14] We require that these companies have share price data from CRSP, yielding 8,381 submissions. Requiring nonzero 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 three months after the fiscal year end, resulting in a sample of 8,149 submissions.[15] Panels A and B of table 2 report the number of XBRL submissions by calendar period and by industry, respectively.[16] 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 distinct tag names (for both custom and standard tags) from the 8,149 submissions. Figure 4, panel A,

[14] When we use U.S. GAAP EPS 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 U.S. GAAP earnings being less informative about fundamentals than pro forma earnings (Bentley et al. [2018]; Bradshaw et al. [2018]).

[15] We keep only XBRL documents filed before the end date of three months after the fiscal year end. 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 end 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.

[16] 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 (see online appendix table A2).

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 Submissions
Food products	149
Beer and 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, nonmetallic, 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.

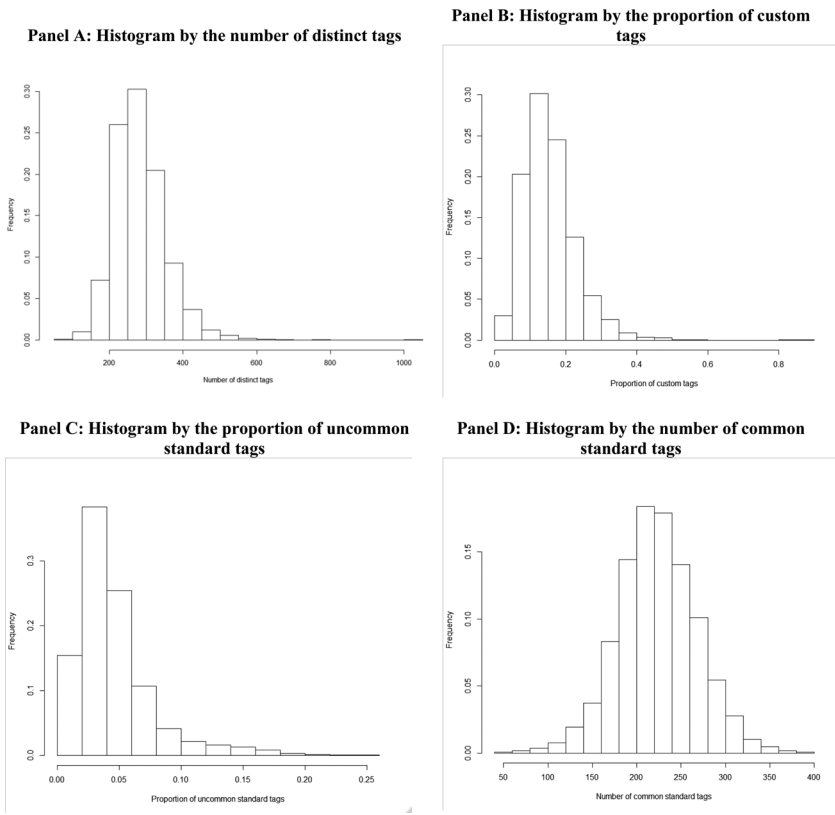


FIG. 4.—Tag distribution across XBRL submissions. Frequency refers to the proportion of XBRL documents in the 8,149 submissions. Panel A shows the histogram by the number of total distinct tags (including both custom and standard tags). Panel B shows the histogram by the proportion of custom tags, calculated as the number of distinct custom tags divided by the number of distinct tags. Panel C shows the histogram by the proportion of uncommon standard tags, calculated as the number of distinct uncommon standard tags divided by the number of distinct tags. Uncommon standard tags are standard tags that have not been used at least once in each year. Panel D 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 distinct common standard tags in subsequent analyses.

shows a histogram by the number of distinct tag names. More than 30% of submissions use 250 to 300 distinct tags, and an average submission uses 284 distinct tags. For each submission, we divide the number of distinct custom tags by the number of distinct tags (i.e., the proportion of custom tags) and plot a histogram by this variable in panel B. For about 30% of submissions, 15%–20% of distinct tags are extensions, and the average proportion of extensions is 15.5%. 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 2012–2018 sample period.[17] Panel C presents a histogram by the proportion of uncommon standard tags (i.e., the number of distinct uncommon standard tags divided by the number of distinct tags). Close to 40% of submissions contain 2%–4% uncommon standard tags, and the average proportion of these tags is 4.5%, 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 distinct common standard tags.[18] Panel D shows a histogram by the number of these tags. We also discard disaggregate items tagged with dimensions, as they are mostly firm-specific (see subsection 2.2) and thus cannot be used in a large sample.[19]

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. Companies use identical tags in a document to refer to items of different reporting periods. For instance, multiple items are identically tagged as "NetIncomeLoss" 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.[20]

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

---

[17] For example, a standard tag "UnrecognizedTaxBenefitsResultingInNetOperatingLossCarryforward" was deprecated in the 2014 U.S. GAAP taxonomy with the implementation of ASU 2013-11 about income taxes in 2014.

[18] Given the proportion of custom tags, the drastic drop in the number of distinct tag names (from 167,136 to 4,627) is due to the firm-specific nature of custom tags. For example, suppose there are 1,000 documents; each document contains 200 standard tags and 50 custom tags that other firms never use. The average proportion of custom tags across the 1,000 documents is 20% (= 50/250), but custom tags account for 50,000 (= 50 × 1,000) out of 50,200 (= 200 + 50 × 1,000) distinct tags. To preserve more standard tags, we also define uncommon standard tags as those that have not been used at least once in each year of the four-year rolling window (i.e., two years of a training sample, one year of a validation sample, and one year of a test sample). Using this alternative definition does not affect our inferences (see online appendix table A3).

[19] The exclusion of items tagged with dimensions causes the loss of some disaggregated detailed information about the primary financial statement items and thus can substantially diminish the potential predictive power of XBRL-tagged data. How to incorporate information tagged with dimensions is an interesting question for future research.

[20] Creating an indicator variable for missing values in each predictor, which will double the number of predictors, does not alter our inferences (see online appendix table A4). We also use the industry-year average to impute missing values and obtain similar results (see online appendix table A5). Dropping the percentage change predictors does not affect our inferences (see online appendix table A6). Also, adding the Fama and French 30 industry indicators to the models does not affect our inferences (see online appendix table A7). This is unsurprising as many items unique to certain industries (e.g., "CapitalizedSoftwareDevelopmentCostsForSoftwareSoldToCustomers") already capture the industry effects.

financial statement" (FASB [2018]). Using this map, we classify 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 natural fit (see online appendix table A8). 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., nonzero) 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 APPROACH TO PREDICTION

3.2.1. Direction of Earnings Changes. We examine the direction of earnings changes for several reasons. First, it is difficult to predict the level of future earnings, as extant studies find that earnings forecasts based on firm characteristics are not substantially more accurate than forecasts obtained from the random-walk model (Kothari [2001], Gerakos and Gramacy [2013], Li and Mohanram [2014]). This also means that it is challenging to forecast the amount of future earnings changes, which is equivalent to forecasting the level of future earnings minus known current earnings. Second, Freeman et al. [1982, p. 643] argue that the substantial variability in earnings changes relative to the amount of information in explanatory variables makes it difficult to form accurate forecasts of earnings changes. They propose to reduce the variability in earnings changes by considering the easier-to-predict direction rather than the amount of the changes. While losing some variation, this binary specification helps mitigate the concern about low out-of-sample performance from predicting the amount of earnings changes.[21] Third, forecasting the sign of earnings changes is economically meaningful and actionable, as previous studies construct portfolios based on the direction of earnings changes (Ou and Penman [1989], Ou [1990], Wahlen and Wieland [2011]).

Following the literature of predicting the direction of earnings changes (Freeman et al. [1982], Ou and Penman [1989]), we adjust for the

[21] Ou and Penman [1989, p. 298] argue: "There is a loss of information in the binary specification, but we were concerned that, given outliers common to accounting data, estimation with dollar magnitudes might produce parameter estimates that perform poorly in out-of-sample prediction and result in investment strategies that give undue weight to estimation errors." While focusing on the direction of earnings changes, we examine the predictability of the machine learning methods concerning the level of earnings and the amount of earnings changes in subsection 4.5.

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
$\text{LiabilitiesAndStockholdersEquity}_t$	8,138	17,290.44	648.93	2,245.22	7,730.65
$\text{RetainedEarningsAccumulatedDeficit}_t$	7,882	2,672.38	−71.66	195.64	1,402.41
$\text{CashAndCashEquivalentsAtCarryingValue}_t$	7,836	762.66	43.25	133.65	479.34
$\text{PropertyPlantAndEquipmentNet}_t$	7,641	2,638.33	47.47	219.90	1,035.82
$\text{StockholdersEquity}_t$	7,635	3,699.93	219.88	702.92	2,282.85
$\text{AccumulatedDepreciationDepletionAndAmortizationPropertyPlantAndEquipment}_t$	7,349	2,052.62	50.78	238.90	956.60
$\text{CommonStockSharesAuthorized}_t$	7,066	142,369.72	100.00	210.00	500.00
$\text{AccumulatedOtherComprehensiveIncomeLossNetOfTax}_t$	7,022	−297.63	−107.91	−8.58	−0.01
$\text{PropertyPlantAndEquipmentGross}_t$	6,935	4,659.43	108.59	485.09	1,982.24

(*Continued*)

TABLE 3—*(Continued)*

Panel C: Top 10 most populated current predictors (i.e., nonzero 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
IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFrom EquityMethodInvestments$_t$	4,794	493.18	4.64	71.61	329.66

Panel D: Top 10 most populated current predictors (i.e., nonzero 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

(*Continued*)

TABLE 3—*(Continued)*

Panel D: Top 10 most populated current predictors (i.e., nonzero values) from cash flow statement

Predictor	Frequency	Mean	Q1	Median	Q3
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., nonzero 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., nonzero values) from shareholders' equity statement

Predictor	Frequency	Mean	Q1	Median	Q3
CommonStockSharesOutstanding$_t$	5,529	97,794.45	31.99	60.39	146.21
AdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriod RecognitionValue$_t$	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

(*Continued*)

TABLE 3—*(Continued)*

Panel F: Top 10 most populated current predictors (i.e., nonzero values) from shareholders' equity statement

Predictor	Frequency	Mean	Q1	Median	Q3
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., nonzero 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 the top 10 most populated (i.e., nonzero) current predictors from balance sheet, income statement, cash flow statement, comprehensive income statement, shareholders' equity statement, and footnotes, respectively, and descriptive statistics for the predictor values. Frequency counts the number of XBRL filings with a nonzero predictor value. Except for per share items, all predictor values are in millions.

firm-specific trend by subtracting the average change in EPS over the past four years from the current EPS changes. An earnings increase/decrease is coded after taking out the drift term. This procedure helps us accomplish three goals. First, as earnings increases tend to outnumber earnings decreases, taking out the drift term mitigates the class imbalance issue (Freeman et al. [1982, p. 645], Japkowicz and Stephen [2002]). Second, as some earnings changes are anticipated due to drift, predicting the direction of de-trended earnings changes is more useful for investment decisions.[22] Third, the drift adjustment permits a direct comparison of our models with the literature (i.e., Ou and Penman [1989]). Nevertheless, our results are robust to using the sign of earnings changes without de-trending (see online appendix table A11).

3.2.2. Parameters. When choosing parameters to tune in a machine learning model, we need to balance prediction accuracy and computational cost. Although trying more values potentially improves prediction accuracy, the price paid for these improvements is increasingly more computational time. In many cases, using a large set of values for parameter tuning is computationally prohibitive. As such, we set these parameters following standard practice in machine learning when available or around default values offered by R packages.[23] Table 4 shows the parameters of the two machine learning methods. 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 2,000 trees with an increment of 100 and bootstrap 50% of

[22] Our final sample from 2012 to 2018 consists of 3,610 earnings increases and 4,539 earnings decreases. Without taking out the drift term, the numbers are 5,418 and 2,731, respectively. Note that this drift adjustment uses only past information and thus does not rely on any foresight. We also use an alternative way to remove the firm-specific trend by subtracting the lagged change in EPS, which overlaps the percentage change predictors in time, from the current change in EPS. No inference is affected (as shown in online appendix table A9). Another 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 prior research (e.g., Ou and Penman [1989]), we do not adopt this alternative strategy in our primary analyses, but report the results of using this alternative in online appendix table A10.

[23] We use the package "randomForest" for random forests and "gbm" for stochastic gradient boosting (also see footnote 39 for loss functions used to compute variable importance). The default values in "randomForest" are 500 trees, a minimum of 1 observation in a leaf, and 1 for bagging. The default values in "gbm" are 100 trees, a learning rate of 0.1, a tree depth of 1, a minimum of 10 observations in a leaf, and 0.5 for bagging.

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,..., 2,000	500, 600, 700,…, 2,000
Learning rate (ρ)		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 (ρ) 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.

the sample for each tree.[24] 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 500–2,000 trees with an increment of 100 with three possible learning rates (0.005, 0.01, and 0.05).[25] We randomly pick 50% 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 1–4 and set the minimum number of observations in a leaf to 10. The best parameters are chosen using the sample splitting method.

3.2.1. Sample Splitting. In machine learning, the data are typically split into training, validation, and testing samples. Models are estimated using the training sample, selected via the validation/hold-out sample (i.e., tuning the parameters), and then applied to the test sample. A common approach to tuning the parameter is n-fold cross-validation. This method splits the sample randomly into n folds and fits the model by excluding one fold as a hold-out sample, which is then used for model evaluation; this procedure is repeated by rotating the excluded fold and selects parameters that maximize the average performance on hold-out samples. As our data set is a panel and predicting future earnings is intertemporal in nature, conducting the n-fold cross-validation is inappropriate: the random partition and rotation of the hold-out samples would imply using past events (e.g., an earnings increase in 2012) to evaluate a model estimated from some future data (e.g., earnings increases/decreases in 2014), inconsistent with

[24] We set the bagging parameter to 0.5 for random forests to enable comparison with stochastic gradient boosting. Nevertheless, using the bagging parameter of 1 yields a similar result (an AUC of 67.53).

[25] We set the minimum number of trees (500) higher than the default value (100) in "gbm" to make it comparable with random forests. Nevertheless, if we allow the machine to grow from 100 to 2,000 trees with an increment of 100, the chosen model always consists of more than 500 trees and exhibits similar predictive power (an AUC of 67.53).

our chronological earnings prediction task. As such, we split the training, validation, and test samples temporally. Specifically, 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. For each year in the test period from 2015 to 2018 (e.g., 2015), the models are trained in the second and third preceding years (e.g., 2012–2013) and validated in the preceding year (e.g., 2014) to tune the parameters as shown in table 4. The training sample always starts fresh in the second and third preceding years and thus has the benefit of using more recent information for model construction.

3.2.2. Model Performance. We use two metrics to evaluate out-of-sample performance. The first is 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]). While commonly used in classification problems, this measure offers little economic meaning for prediction gains. To better quantify these gains, we also use excess returns to easily implementable hedge portfolios as the second measure. Once we apply the estimated model to the year in the test period, we obtain the summary measure $\widehat{Pr}$, the predicted probability of an earnings increase in the next year. A hedge portfolio is then formed based on this measure. Specifically, each stock in the sample 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, from three months after the fiscal year-end, when all the predictor values are publicly available as the 10-K has been filed, to three months after the next fiscal year-end, when the next annual earnings have been released.[26] We measure excess returns using the size-adjusted returns (SAR), which are commonly used in accounting and finance studies (Piotroski and So [2012], Li and Mohanram [2019]). 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.

[26] On average, the next earnings announcement is 320 days after the portfolio formation date and 45 days before the portfolio end date, and the next 10-K filing is 331 days after the portfolio formation date and 34 days before the portfolio end date. For several reasons, we do not form the portfolio immediately after the 10-K filing and liquidate it several days after the next earnings announcement but before the next 10-K filing. First, the trading horizon varies across firms, making the excess returns less comparable. Second, this strategy requires more effort in predicting the next 10-K filing dates. Third, for 18% (9%) of firms, the earnings announcement and 10-K filing occurred on the same day (in two consecutive days), making it challenging to implement such a strategy. Nevertheless, we acknowledge that the returns of our strategy contain noise to the extent that stock prices incorporate some information in the next 10-K filing by the portfolio end date.

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} (see online appendix table A12).

4. Results

4.1 OUT-OF-SAMPLE PERFORMANCE OF MACHINE LEARNING MODELS

Table 5 reports the out-of-sample prediction performance in the 2015–2018 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 $\widehat{Pr} \leq 0.4$).[27] The distribution of observations by year (in table 1, panel A) and the rolling windows for machine learning (in subsection 3.2.3) can be used to compute the training and validation sample size for each year in the test period. There are 5,520 observations in the entire test period. Among predicted increases, 60.05%–65.64% are actually earnings increases. We also observe that 61.9%–67.5% of observations are correctly predicted. Specifically, 28.12%–33.07% (86.69%–94.64%) of earnings increases (decreases) are correctly predicted, suggesting that the symmetric cutoffs are too stringent (lenient) for earnings increases (decreases). For example, using 0.4 as the cutoff ($\widehat{Pr} > 0.4$ and $\widehat{Pr} \leq 0.4$), we observe that 61.64% (63.01%) of earnings increases (decreases) are correctly predicted by random forests, and 57.13% (66.98%) of earnings increases (decreases) are correctly predicted by stochastic gradient boosting. We next turn to the AUC, which does not rely on specific cutoffs.

The AUC ranges from 67.52% to 68.66% depending on methods (random forests or stochastic gradient boosting) and samples (the full sample or the sample with $\widehat{Pr} \geq 0.6$ and $\widehat{Pr} \leq 0.4$), significantly higher than 50% of a random guess.[28] Following Carpenter and Bithell [2000], we construct

[27] The chosen parameter values for each method are reported in online appendix table A13. 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 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).

[28] A random walk model (with drift) predicts an equal probability of an earnings increase and decrease (net of drift), which represents only one point (0.5, 0.5) on the ROC curve of a random guess (the diagonal line; Fawcett [2006, p. 863]). We also consider two simple classifiers, although they have not been developed in the literature: (1) predicting an earnings increase (net of drift) next year if there is an earnings increase (net of drift) in the current year and (2) predicting an earnings increase (net of drift) next year if there is an earnings decrease (net of drift) in the current year. The two classifiers exhibit an AUC of 46.54% and 53.46%, respectively, lower than our models'. Unsurprisingly, the sum of the two AUCs equals 1 as (2) is a flip of (1). It is also consistent with the intuition that the drift adjustment makes

TABLE 5
Out-of-Sample Prediction Performance

Panel A: Machine learning models

	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
% of predicted increases that are actual earnings increases	60.10	65.64	60.05	64.50
% correctly predicted	61.90	67.50	62.26	66.90
% of actual earnings increases correctly predicted	33.07	28.12	31.03	29.28
% of actual earnings decreases correctly predicted	86.69	94.64	89.12	94.58
AUC (%)	67.52	68.62	67.54	68.66
Bootstrap *p*-value for AUC versus 50%	<0.01	<0.01	<0.01	<0.01
SAR (%)	5.02	9.43	6.57	9.74
Bootstrap *p*-value for SAR versus 0%	<0.01	<0.01	<0.01	<0.01

(*Continued*)

TABLE 5—*(Continued)*

Panel B: Benchmark models

			Compared with			
			XBRL/RF		XBRL/SGB	
	AUC (%)	SAR (%)	Bootstrap *p*-Value for Difference in:		Bootstrap *p*-Value for Difference in:	
			AUC	SAR	AUC	SAR
1. OP/Logit	61.79	2.48	<0.01	<0.01	<0.01	<0.01
1a. OP/RF	66.63	4.67	0.064	<0.01		
1b. OP/SGB	66.87	3.91			0.108	<0.01
2. DuPont/Logit	57.96	1.90	<0.01	<0.01	<0.01	<0.01
2a. DuPont/RF	61.51	2.73	<0.01	<0.01		
2b. DuPont/SGB	61.15	2.12			<0.01	<0.01
3. Analysts' forecasts	65.09	3.38	<0.01	<0.01	<0.01	<0.01

Panel A presents a summary of prediction performance using XBRL-tagged financial data and different probability cutoffs. Panel B presents a summary of prediction performance for different benchmark models and the comparison between them and our machine learning models (XBRL/RF and XBRL/SGB). See subsections 4.1 and 4.2 for details.

a bootstrap p-value for the difference between our AUCs and 50%. 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%. All the p-values are less than 0.01, indicating that our models' predictive power is unlikely to be a random outcome.[29] The results suggest that the machine learning models extract meaningful signals from the detailed financial data.

Table 5 also 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. 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% for random forests and 6.57% for stochastic gradient boosting. These returns account for 38.7% and 50.7% of returns from a strategy with perfect foresight of the direction of one-year-ahead earnings changes (i.e., 12.97%).

To assess 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 0.01 for both returns (5.02% and 6.57% for random forests and stochastic gradient boosting, respectively) 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% for random forests and 9.74% for stochastic gradient boosting.[30]

---

the first classifier not work well. If we do not take out drift from earnings changes, the two classifiers exhibit an AUC of 56.33% and 43.67%, respectively.

[29] 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.

[30] In additional analyses reported in the online appendix, we observe the following patterns of the excess returns. (1) The returns increase monotonically as $\widehat{Pr}$ moves from below 0.1 to above 0.8 (online appendix table A14) and are concentrated in the long positions (online appendix figure A1). (2) The excess returns persist after accounting for transaction costs estimated as the effective bid–ask spread following Novy-Marx and Velikov [2016] (online appendix table A15). (3) The excess returns are robust to using two alternative metrics (ROE and EBIT per share) to measure the direction of earnings changes (online appendix table A16, panel A), excluding microcaps (online appendix table A16, panel B), accounting for the five risk factors of Fama and French [2015] (market returns in excess of the one-month T-bill rate and returns on size, book-to-market, profitability, and investment portfolios; online appendix table A16, panel C). (4) No inference is affected if we use size-and-book-to-market-adjusted returns (Green et al. [2011]; online appendix table A17 and figure A2).

4.2 THREE BENCHMARKS

To learn more about the out-of-sample performance of our models, we compare them with three benchmarks. The first two are conventional regression models, with one using a "kitchen sink" approach and the other using the DuPont Analysis to select predictors. The last benchmark comes from forecasts of professional analysts.

4.2.1. Ou and Penman's [1989] Model. Following Ou and Penman [1989], we estimate logistic regressions using 65 accounting variables, which typically appear in financial statement analysis textbooks.[31] Their model represents a "kitchen sink" approach with a large number of predictors. 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. Their variable selection approach consists of three steps. First, they run a univariate logistic regression for each of the 65 variables and retain only variables that load significantly at the 10% level. Second, a logistic regression is estimated using all remaining variables simultaneously. All variables with coefficients that are not significant at the 10% level are dropped. In the final stage, for the remaining variables, they delete the variables that do not load significantly at the 10% level stepwise until all explanatory variables have statistically significant coefficients at the 10% level. This stepwise elimination helps remove redundant variables and improve model fitting. We strictly follow each step of their approach and refer to the final logistic model as OP/Logit. To better understand the differences between our random forests and stochastic gradient boosting models (XBRL/RF and XBRL/SGB, respectively) 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.

As shown in table 5, panel B, our models significantly outperform OP/Logit by a large margin. The XBRL/RF (XBRL/SGB) model exhibits an AUC of 67.52% (67.54%), compared with 61.79% for OP/Logit. We also observe an AUC of 66.63% (66.87%) for the OP/RF (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

[31] 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].

versus OP/SGB, for which the *p*-value is 0.108. Thus, the improvements in predictive power are unlikely to be random outcomes.[32]

We observe a similar pattern in hedge portfolio returns. The strategy of applying machine learning methods to the 65 variables generates size-adjusted returns of 4.67% for OP/RF and 3.91% for OP/SGB. These returns are significantly higher than 2.48% from Ou and Penman's original strategy (OP/Logit), and marginally lower than 5.02% for XBRL/RF and 6.57% for XBRL/SGB. 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 return 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 0.01 for all pairs. Overall, 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.

4.2.2. DuPont Analysis. Unlike the "kitchen sink" approach in Ou and Penman [1989], Nissim and Penman [2001] derive eight drivers of earnings based on the DuPont Analysis. The eight drivers are sales profit margin, asset turnover, other items divided by net operating assets, financial leverage, net borrowing cost, return on net operating assets, operating liability leverage, and minority interest sharing (Nissim and Penman [2001, p. 118]). These expert-identified variables could improve predictive performance.

For each year in the 2015–2018 test period, we use the past three years to estimate a logistic model using levels of and changes in the eight variables to predict the direction of one-year-ahead earnings changes (net of drift).[33] The estimated coefficients are then applied to the test year. The resulting forecasts from this analysis exhibit an AUC of 57.96% and annual size-adjusted returns of 1.90%, significantly lower than those of our models. Nissim and Penman [2001] acknowledge that while these eight drivers map

[32] The results for OP/RF versus XBRL/RF and OP/SGB versus XBRL/SGB should be interpreted with caution, as none of them is statistically significant if the significance threshold of 5% is used. 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 versus OP/RF and XBRL/SGB versus OP/SGB in 2015, which is unsurprising, given the data quality issues in the early years in the training sample (2012–2013) for 2015.

[33] Soliman [2008] estimates a similar model with the linear combination of levels of and changes in DuPont components to explain future returns on net operating assets.

into current earnings, their predictive power depends on their time-series behavior, which is not explicitly developed.

As the logistic model does not accommodate the multiplicative nature of the DuPont Analysis (e.g., profit margin and asset turnover), we apply the machine learning methods to the levels of and changes in the eight drivers (DuPont/RF and DuPont/SGB). DuPont/RF (DuPont/SGB) exhibits an AUC of 61.51% (61.15%) and size-adjusted returns of 2.73% (2.12%), which are significantly higher than those of DuPont/Logit with all bootstrap p-values <0.01 and lower than those of XBRL/RF (XBRL/SGB) with all bootstrap p-values <0.01. The results highlight the usefulness of both machine learning and detailed financial information in earnings prediction.

4.2.3. Analysts' Forecasts. Professional analysts have access to machine learning, detailed financial information from 10-K filings, and other sources of information. Their forecasts could dominate our models and thus serve as the third benchmark. We take analysts' forecasts in the month following the portfolio formation when all the detailed financial information in 10-Ks is available and compare each forecast with the realized earnings net of drift in fiscal year t (i.e., subtracting drift from the difference between a forecast and realized earnings) to determine whether an analyst forecasts an earnings increase or decrease for the covered firm. The analysts' prediction is the proportion of forecasts that indicate an earnings increase. Table 5, panel B, shows an AUC of 65.09% for analysts' prediction of an earnings increase, significantly lower than our models (bootstrap p-value <0.01). A hedge portfolio with a long (short) position for stocks with the analysts' prediction of an earnings increase (decrease) >0.5 (≤ 0.5) yields size-adjusted returns of 3.38%, lower than those from our models.[34] The results highlight the usefulness of machine learning and detailed financial information in earnings prediction, even in the presence of professional forecasters, who may have limited ability to process detailed financial data. Figure 5, panel A, reports the ROC curves for our machine learning models and the benchmarks. Consistent with the results in table 5, the figure shows that our models dominate all of them.

---

[34] If we do not subtract drift from the difference between analysts' forecasts and realized earnings to determine the analysts' prediction, the AUC and size-adjusted returns are 64.71% and 3.93%, respectively. If we compare the consensus (i.e., median) analyst forecast with the realized earnings in fiscal year t to define the analysts' prediction, the AUC and size-adjusted returns are 63.62% and 2.49%, respectively. The decreased predictive power is unsurprising because the median forecast misses information from the entire forecast distribution (e.g., Q1 and Q3). We also examine whether consensus analyst 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% for random forests and 7.01% for stochastic gradient boosting, higher than the original returns (5.02% and 6.57%, respectively).

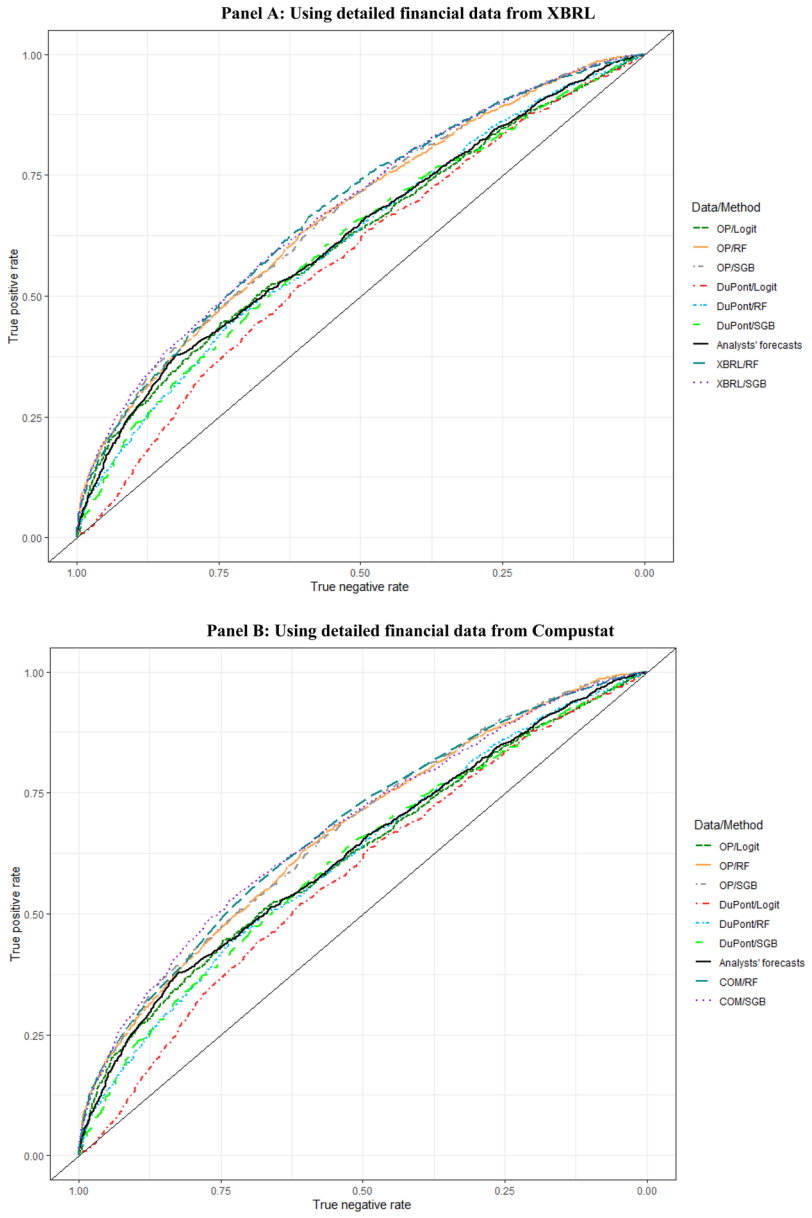


FIG. 5.—Comparison of out-of-sample ROC curves for different data/method combinations. Panels A and B present out-of-sample ROC curves for different data/method combinations. The data consist of the Ou and Penman [1989] variables (OP), DuPont variables, our XBRL-tagged items (XBRL), and all Compustat items (COM). The employed methods are logistic regression (Logit), random forests (RF), and stochastic gradient boosting (SGB). Analysts' forecasts are taken from I/B/E/S.

4.3 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.[35] 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 per-share basis). Table 6, panel A, reports the predictive power of detailed financial information from Compustat similar to XBRL-tagged data. Figure 5, panel B, reports the ROC curves for our machine learning models using the Compustat data. This figure and table 6, panel B, show that our models with the Compustat data continue to outperform the three benchmarks. The results also suggest that the benefit of additional financial details relative to Compustat is offset by the presence of errors, custom tags, and tags with dimensions in XBRL documents.

4.4 VARIATION IN DATA QUALITY

We conduct two additional tests based on variation in XBRL data quality. First, as errors in XBRL documents are subject to only modified 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 A, our models exhibit higher AUCs 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.[36] The average annual size-adjusted returns are also higher in the late period. The bootstrap p-value for the SAR difference between the two periods is less than 0.01 for both models. The results suggest that data quality issues in early adoption years decrease the usefulness of detailed financial data in XBRL documents.

Second, we use the proportion of distinct custom and uncommon standard tags in an XBRL submission as an inverse measure of data quality (or the "completeness" of data used in our models) 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 re-

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

[36] 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 A18).

TABLE 6
Using Detailed Financial Data from Compustat

Panel A: Machine learning models using Compustat data

	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
% of predicted increases that are actual earnings increases	60.53	65.37	62.37	65.18
% correctly predicted	62.10	67.51	63.66	66.78
% of actual earnings increases correctly predicted	35.66	35.53	42.67	43.07
% of actual earnings decreases correctly predicted	84.84	91.57	81.70	85.89
AUC (%)	67.50	69.40	67.39	68.66
Bootstrap *p*-value for AUC versus 50%	<0.01	<0.01	<0.01	<0.01
SAR (%)	5.12	9.74	3.95	10.07
Bootstrap *p*-value for SAR versus 0%	<0.01	<0.01	<0.01	<0.01

(*Continued*)

TABLE 6—*(Continued)*

Panel B: Benchmark models

			Compared with			
			COM/RF Bootstrap *p*-Value for Difference in:		COM/SGB Bootstrap *p*-Value for Difference in:	
	AUC (%)	SAR (%)	AUC	SAR	AUC	SAR
1. OP/Logit	61.79	2.48	<0.01	<0.01	<0.01	<0.01
1a. OP/RF	66.63	4.67	0.070	<0.01		
1b. OP/SGB	66.87	3.91			0.210	<0.01
2 DuPont/Logit	57.96	1.90	<0.01	<0.01	<0.01	<0.01
2a. DuPont/RF	61.51	2.73	<0.01	<0.01		
2b. DuPont/SGB	61.15	2.12			<0.01	<0.01
3 Analysts' forecasts	65.09	3.38	<0.01	<0.01	<0.01	<0.01

Panel A presents a summary of prediction performance using all Compustat items and different probability cutoffs. Panel B presents a summary of prediction performance for different benchmark models and the comparison between them and our machine learning models (COM/RF and COM/SGB). See subsection 4.3 for details.

TABLE 7
Additional Analyses for Data Quality

Panel A: 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 diff. in AUC	0.049		0.307	
SAR (%)	2.11	11.87	3.98	11.66
Bootstrap *p*-value for SAR vs. 0%	0.178	<0.01	<0.01	<0.01
Bootstrap *p*-value for diff. in SAR	<0.01		<0.01	

Panel B: 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 diff. in AUC	<0.01		<0.01	
SAR (%)	8.80	8.87	8.67	9.79
Bootstrap *p*-value for SAR vs. 0%	<0.01	<0.01	<0.01	<0.01
Bootstrap *p*-value for diff. in SAR	0.014		<0.01	

Panel A shows AUCs and 12-month size-adjusted returns (SAR) of two subsamples by period. The early period is 2015 and the late period is 2016-2018. Panel B shows AUCs and 12-month size-adjusted returns (SAR) 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. See subsection 4.4 for details.

moved before the analysis.[37] We split the test sample by the year median and report the AUC of each subsample in table 7, panel B. Our models exhibit higher AUCs for firms with high data quality than other firms. The bootstrap *p*-value for the AUC difference between the two subsamples is less than 0.01 for both methods. The average annual size-adjusted returns are also higher for firms with high data quality. The bootstrap *p*-value for the SAR difference between the two subsamples is 0.014 for random forests and less than 0.01 for stochastic gradient boosting. The results suggest that low data quality reduces the predictive power of detailed financial data in XBRL documents.

[37] We do not claim that the use of the extension per se is inappropriate, but rather that the quality of data used in the modeling is lower if the extensions cannot easily be incorporated. As such, this measure should not be interpreted as reflecting firms' data quality in general.

4.5 PREDICTING THE LEVEL OF EARNINGS AND THE AMOUNT OF EARNINGS CHANGES

We examine the direction of earnings changes in primary analyses because it is difficult to predict the level of future earnings and the amount of earnings changes (future earnings minus known current earnings) and for other reasons discussed in subsection 3.2.1. Nevertheless, in this section, we use the two machine learning methods to predict the level of earnings and the amount of earnings changes following the same rolling windows as in subsection 3.2.3. Consistent with prior research, we observe a low out-of-sample R^2 of 5.3% (6%) for random forests (stochastic gradient boosting) in predicting the level of one-year-ahead earnings, lower than the out-of-sample R^2 of 7.5% for a simple random-walk model.[38] We also observe a low out-of-sample R^2 of 8% (5.8%) for random forests (stochastic gradient boosting) in predicting the amount of one-year-ahead earnings changes. A hedge portfolio with a long (short) position for stocks with the predicted earnings changes greater than (less than or equal to) 0 yields size-adjusted returns of 0.10% (p-value = 0.47) for random forests and 0.11% (p-value = 0.46) for stochastic gradient boosting. The results suggest that focusing on the direction of earnings changes facilitates the success of our machine learning models.

5. Understanding the Inner Workings of the Machine Learning Models

In this section, we seek to understand the inner working of the machine learning models. While quantifying each predictor's importance to the predictive power, we caution that the reader should not interpret the importance as the causal influence of the predictor. Rather, it is used to provide transparency for the items responsible for inferences drawn by the machine learning models.

---

[38] Following Gerakos and Gramacy [2013], we apply random forests to their 24 financial variables to predict the level of earnings. As they use the root-mean-squared error adjusted by the Consumer Price Index (adjusted RMSE) to evaluate model performance, we compute this metric for this model (GG/RF_L) and our models (XBRL/RF_L and XBRL/SGB_L), where "L" denotes the level of earnings. We observe an adjusted RMSE of 2.424 for GG/RF_L, similar to 2.409 as reported in their table 2. Our models yield an adjusted RMSE of 1.913 for XBRL/RF_L and 1.766 for XBRL/SGB_L, lower than GG/RF_L. One benefit of predicting the direction of earnings changes is that the direction is actionable. In contrast, a level of earnings would have to be combined with an expectation proxy. Using the consensus (i.e., median) analyst forecast in the month following the portfolio formation as such a proxy, we find that a hedge portfolio with a long (short) position for stocks with the predicted earnings minus the consensus greater than (less than or equal to) zero yields size-adjusted returns of 0.49% (p-value = 0.29) for random forests and –0.24% (p-value = 0.40) for stochastic gradient boosting.

We estimate each variable's importance by computing the predictive performance decrease when that variable is randomly shuffled.[39] As 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. 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 8, panel A, presents the top 10 most important variables. Most of them pertain to earnings, such as "NetIncomeLoss" and "EarningsPerShareBasic." The results suggest that earnings are still the most critical metrics among all the financial items in 10-K filings. 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 expenses in the top 70 list for random forests. 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.[40]

Our models contain many potentially multicollinear financial items. Although this is not an issue for building a predictive model, we may underestimate the individual importance of multicollinear predictors. To address this issue, we follow the Pedregosa et al. [2011] method by (1) performing hierarchical clustering on the predictors' Spearman rank-order correlations, (2) imposing a threshold of 0.8 to obtain 8,993 clusters, and (3) randomly picking one predictor from each cluster to be kept in the model.[41] We observe an AUC of 67.46% (67.06%) for the new random forests (stochastic gradient boosting) model, close to that of the original model. More importantly, we estimate the importance of the randomly chosen predictor in the new model, assign the importance to all the other predictors in the same cluster, and compute the correlation of importance

[39] We use the function "varImp(, type = 1, scale = FALSE)" for random forests and function "summary(, method = permutation.test.gbm, normalize = FALSE)" for stochastic gradient boosting. The former computes the reduction in the average prediction error rate across all trees and the latter calculates the decrease in the binomial deviance loss function (see equation 10.22 in Hastie et al. [2009] when $K = 2$), when each variable is randomly shuffled.

[40] 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.

[41] A limitation of this approach is that the reduction in dimensionality (from 13,881 predictors to 8,993 clusters) is not substantial as we impose a high threshold of 0.8.

TABLE 8
Importance of Predictors

Panel A: Top 10 most important predictors

Random Forests	Stochastic Gradient Boosting
$OperatingIncomeLoss_{t-1}$	$RetainedEarningsAccumulatedDeficit_{t-1}$
$NetIncomeLoss_t$	$\%\Delta LiabilitiesCurrent$
$ComprehensiveIncomeNetOfTax_{t-1}$	$EarningsPerShareBasicAndDiluted_t$
$ComprehensiveIncomeNetOfTax_t$	$OperatingIncomeLoss_{t-1}$
$OperatingIncomeLoss_t$	$EmployeeServiceShareBasedCompensationTaxBenefitFromCompensationExpense_t$
$NetIncomeLoss_{t-1}$	$IncomeTaxReconciliationChangeInDeferredTaxAssetsValuationAllowance_t$
$EarningsPerShareBasic_{t-1}$	$NetIncomeLoss_t$
$EarningsPerShareDiluted_{t-1}$	$\%\Delta StockholdersEquity$
$EarningsPerShareDiluted_t$	$TreasuryStockValue_{t-1}$
$IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethodInvestments_{t-1}$	$ShortTermInvestments_{t-1}$

(*Continued*)

TABLE 8—*(Continued)*

Panel B: Top 10 most important predictors in footnotes

Random Forests	Stochastic Gradient Boosting
$DeferredTaxAssetsValuationAllowance_{t}$	$EmployeeServiceShareBasedCompensationTaxBenefitFromCompensationExpense_{t}$
$IncomeLossFromContinuingOperationsBeforeIncomeTaxesDomestic_{t}$	$IncomeTaxReconciliationChangeInDeferredTaxAssetsValuationAllowance_{t}$
$IncomeTaxReconciliationIncomeTaxExpenseBenefitAtFederalStatutoryIncomeTaxRate_{t}$	$\%\Delta AllocatedShareBasedCompensationExpense$
$IncomeLossFromContinuingOperationsBeforeIncomeTaxesDomestic_{t-1}$	$UndistributedEarningsOfForeignSubsidiaries_{t}$
$DeferredTaxAssetsValuationAllowance_{t-1}$	$CurrentForeignTaxExpenseBenefit_{t}$
$IncomeTaxReconciliationIncomeTaxExpenseBenefitAtFederalStatutoryIncomeTaxRate_{t-1}$	$\%\Delta ShareBasedCompensationArrangementByShareBasedPaymentAwardOptionsOutstandingNumber$
$CurrentIncomeTaxExpenseBenefit_{t}$	$\%\Delta DeferredTaxAssetsNetNoncurrent$
$CurrentFederalTaxExpenseBenefit_{t}$	$\%\Delta CapitalLeasesLesseeBalanceSheetAssetsByMajorClassAccumulatedDeprecation$
$CurrentFederalTaxExpenseBenefit_{t-1}$	$\%\Delta ShareBasedCompensationArrangementByShareBasedPaymentAwardOptionsOutstandingWeightedAverageExercisePrice$
$CurrentIncomeTaxExpenseBenefit_{t-1}$	$UnrecognizedTaxBenefitsDecreasesResultingFromSettlementsWithTaxingAuthorities_{t}$

Panel A provides a list of the top 10 most important predictors for random forests and stochastic gradient boosting. Panel B presents a list of the top 10 most important predictors in footnotes for random forests and stochastic gradient boosting. The importance of a predictor is computed as the decrease in the predictive performance when that variable is randomly shuffled.

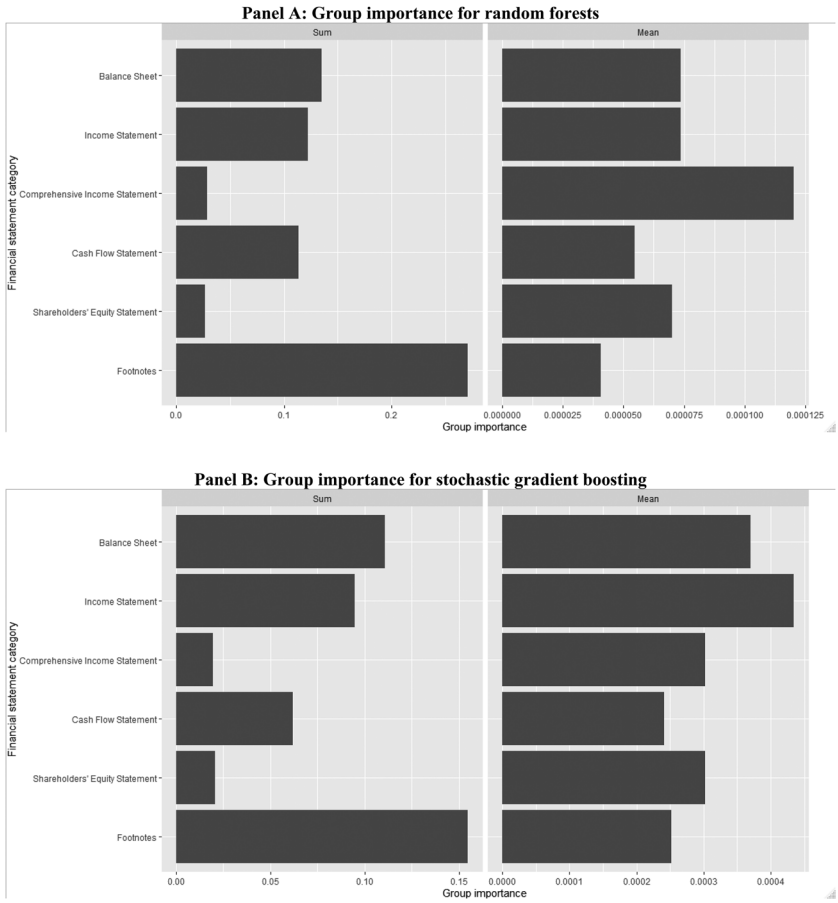


FIG. 6.—Group importance. Panels A and B show the importance of predictors grouped by financial statement category for random forests and stochastic gradient boosting, respectively. Each predictor is classified into balance sheet, income statement, comprehensive income statement, cash flow statement, shareholders' equity statement, or footnotes. The importance of a predictor is computed as the decrease in the predictive performance when that variable is randomly shuffled. The sum and mean of the predictor importance grouped by financial statement category are reported.

values between the new model and the original one across all predictors (i.e., $N = 13{,}881$). The two important values are highly correlated with a coefficient of 0.93 (0.80) for random forests (stochastic gradient boosting). The results suggest that multicollinearity is unlikely to influence the ranking of predictor importance significantly.

Figure 6 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]).[42] 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 6 also reports the mean of variable importance within each category. We observe that items from financial statements on average play a stronger role than footnote items, but the latter's importance is still considerable.

Table 8, panel B, shows the top 10 important variables in footnotes (see online appendix table A19 for each category's top 10 most important variables). 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 association between tax items and $\widehat{Pr}$, we construct partial dependence plots (Hastie et al. [2009]).[43] Figure 7, panel A, shows a nonlinear negative association between the valuation allowance for deferred tax assets (the top predictor from footnotes) and $\widehat{Pr}$ under random forests, consistent with Miller and Skinner [1998]. We also observe an interaction effect in panel B: $\widehat{Pr}$ becomes higher when the valuation allowance is lower and lagged operating income (the top predictor under random forests) is higher. The results suggest that the valuation allowance provides additional details on operating income growth by revealing management's assessment of future taxable income. Panel C presents a nonlinear negative association between tax benefits related to the exercise of employee stock options (Hanlon and Shevlin [2002]; the top predictor from footnotes) and $\widehat{Pr}$ under stochastic gradient boosting. Panel D shows an interaction effect: $\widehat{Pr}$ becomes lower when both the tax benefits and lagged retained earnings

[42] This comparison is within XBRL-tagged data. It is not inconsistent with the previous finding that detailed financial data from XBRL documents (including both financial statement and footnote items) are marginally more useful than the 65 variables from Compustat under machine learning (e.g., XBRL/RF vs. OP/RF). Compared with the 65 variables, footnotes carry additional information, but financial statement items in XBRL documents suffer from data quality issues. As such, the usefulness of combined financial statement and footnote items is only marginally higher than that of the 65 variables.

[43] 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.

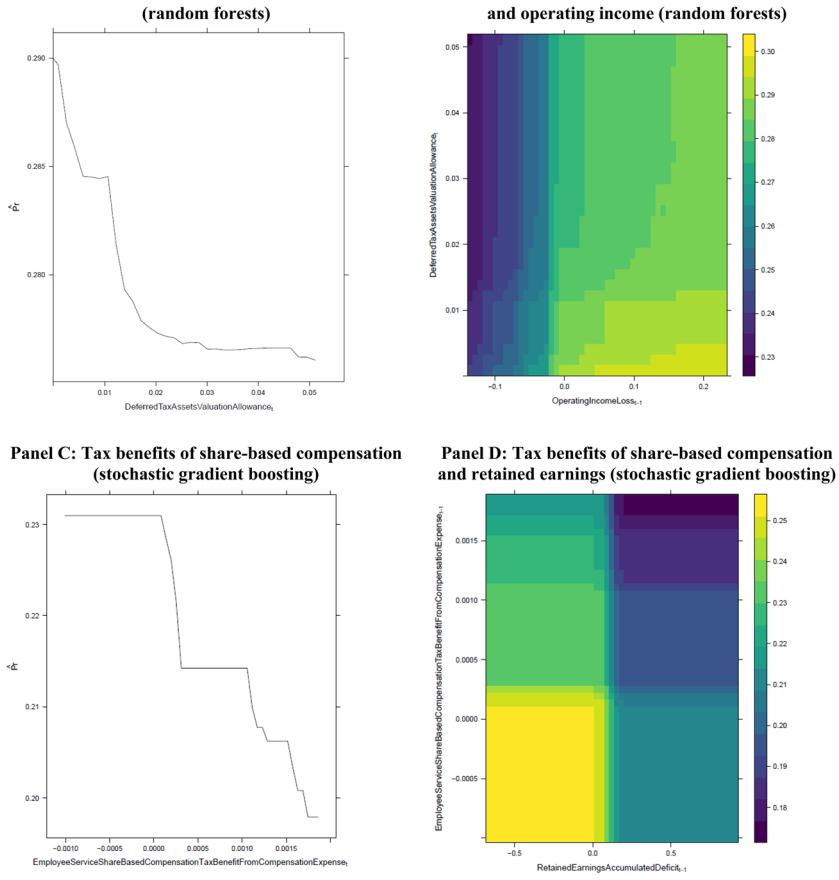


FIG. 7.—Partial dependence plots. Panels A and C show one-way partial dependence plots and panels B and D show two-way partial dependence plots. In a one-way partial dependence plot, for each value of a predictor (in the *x*-axis), we force all observations in the training sample to assume that value 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.

(the top 1 predictor under stochastic gradient boosting) are higher. The results are consistent with the Bartov and Mohanram [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.

6. Conclusion

We apply machine learning to a large set of detailed financial information aimed at predicting the direction of future earnings changes. Leveraging ensemble 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 ranges from 67.52% to 68.66% and is significantly higher than that of a random guess, which is only 50%. The annual size-adjusted returns to hedge portfolios formed based on this measure range from 5.02% to 9.74%, indicating economically significant predictive gains.

Our models using machine learning and a large set of detailed financial information outperform two conventional models that use logistic regressions and small sets of accounting variables, and professional analysts' forecasts. Analyses suggest that the outperformance relative to the conventional models stems from both nonlinear predictor interactions missed by regressions and the use of more detailed financial data. The results highlight the usefulness of machine learning and detailed financial information in predicting the direction of earnings changes.

REFERENCES

Anand, V.; R. Brunner; K. Ikegwu; and T. Sougiannis. "Predicting Profitability Using Machine Learning." Working paper, University of Illinois at Urbana-Champaign, 2019. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3466478.

Bao, Y.; B. Ke; B. Li; J. Yu; and J. Zhang. "Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach." *Journal of Accounting Research* 58 (2020): 199–235.

Barth, M.; K. Li; and C. Mcclure. "Evolution in Value Relevance of Accounting Information." Working paper, Stanford University, 2021. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2933197.

Bartley, J.; A. Chen; and E. Taylor. "A Comparison of XBRL Filings to Corporate 10-Ks—Evidence from the Voluntary Filing Program." *Accounting Horizon* 25 (2011): 227–45.

Bartov, E., and P. Mohanram. "Private Information, Earnings Manipulations, and Executive Stock-option Exercises." *The Accounting Review* 79 (2004): 889–920.

Bentley, J.; T. Christensen; K. Gee; and B. Whipple. "Disentangling Managers' and Analysts' Non-GAAP Reporting." *Journal of Accounting Research* 56 (2018): 1039–81.

Bertomeu, J.; E. Cheynel; E. Floyd; and W. Pan. "Using Machine Learning to Detect Misstatements." *Review of Accounting Studies* 26 (2021): 468–519.

Bhattacharya, N.; Y. Cho; and J. Kim. "Leveling the Playing Field Between Large and Small Institutions: Evidence from the SEC's XBRL Mandate." *The Accounting Review* 93 (2018): 51–71.

Binz, O.; K. Schipper; and K. Standridge. "What can Analysts Learn from Artificial Intelligence about Fundamental Analysis?" Working paper, INSEAD, Duke University, 2021. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3745078.

Blankespoor, E. "The Impact of Information Processing Costs on Firm Disclosure Choice: Evidence from the XBRL Mandate." *Journal of Accounting Research* 57 (2019): 919–67.

Blankespoor, E.; B. Miller; and H. White. "Initial Evidence on the Market Impact of the XBRL Mandate." *Review of Accounting Studies* 19 (2014): 1468–503.

Boritz, E., and W. No. "Assurance on XBRL-Related Documents: The Case of United Technologies Corporation." *Journal of Information Systems* 23 (2009): 49–78.

Bradshaw, M.; T. Christensen; K. Gee; and B. Whipple. "Analysts' GAAP Earnings Forecasts and their Implications for Accounting Research." *Journal of Accounting and Economics* 66 (2018): 46–66.

Breiman, L. "Random Forests." *Machine Learning* 45 (2001): 5–32.

Brown, L. "Influential Accounting Articles, Individuals, Ph.D. Granting Institutions and Faculties: A Citational Analysis." *Accounting, Organization and Society* 21 (1996): 723–54.

Cao, K., and H. You. "Fundamental Analysis via Machine Learning." Working paper, Hong Kong University of Science and Technology, 2020. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3706532.

Carpenter, J., and J. Bithell. "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians." *Statistics in Medicine* 19 (2000): 1141–64.

Cecchini, M.; H. Aytug; G. Koehler; and P. Pathak. "Detecting Management Fraud in Public Companies." *Management Science* 56 (2010): 1146–60.

Chinco, A.; A. Clark-Joseph; and M. Ye. "Sparse Signals in the Cross-Section of Returns." *Journal of Finance* 74 (2019): 449–92.

Chychyla, R., and A. Kogan. "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 (2015): 37–72.

Cohen, L.; C. Malloy; and Q. Nguyen. "Lazy Prices." *Journal of Finance* 75 (2020): 1371–415.

De Franco, G.; M.H. Wong; and Y. Zhou. "Accounting Adjustments and the Valuation of Financial Statement Note Information in 10-K Filings." *The Accounting Review* 86 (2011): 1577–604.

Debreceny, R.; S. Farewell; M. Piechocki; C. Felden; and A. Graning. "Does It Add Up? Early Evidence on the Data Quality of XBRL Filings to the SEC." *Journal of Accounting and Public Policy* 29 (2010): 296–306.

Debreceny, R.; S. Farewell; M. Piechocki; C. Felden; A. Graning; and A. D'eri. "Flex or Break? Extensions in XBRL Disclosures to the SEC." *Accounting Horizon* 25 (2011): 631–57.

Ding, K.; B. Lev; X. Peng; T. Sun; and M. Vasarhelyi. "Machine Learning Improves Accounting Estimates: Evidence from Insurance Payments." *Review of Accounting Studies* 25 (2020): 1098–134.

Dong, Y.; O. Li; Y. Lin; and C. Ni. "Does Information Processing Cost Affect Firm-Specific Information Acquisition? Evidence from XBRL Adoption." *Journal of Financial and Quantitative Analysis* 51 (2016): 435–62.

Du, H.; M. Vasarhelyi; and X. Zheng. "XBRL Mandate: Thousands of Filing Errors and So What?" *Journal of Information Systems* 27 (2013): 61–78.

Dyer, T.; M. Lang; and L. Stice-Lawrence. "The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation." *Journal of Accounting and Economics* 64 (2017): 221–45.

Efendi, J.; J. Park; and C. Subramaniam. "Does the XBRL Reporting Format Provide Incremental Information Value? A Study Using XBRL Disclosures during the Voluntary Filing Program." *Abacus* 52 (2016): 259–85.

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

Fasb. "SEC Reporting Taxonomy Technical Guide." 2018. Available at https://www.fasb.org/cs/ContentServer?d=Touch&c=Document_C&pagename=FASB%2FDocument_C%2FDocumentPage&cid=1176169716122.

Fawcett, T. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (2006): 861–74.

Frankel, R.; J. Jennings; and J. Lee. "Using Unstructured and Qualitative Disclosures to Explain Accruals." *Journal of Accounting and Economics* 62 (2016): 209–27.

Freeman, R.; J. Ohlson; and S. Penman. "Book Rate-of-Return and Prediction of Earnings Changes: An Empirical Investigation." *Journal of Accounting Research* 20 (1982): 639–53.

Freyberger, J.; N. Andreas; and M. Weber. "Dissecting Characteristics Nonparametrically." *Review of Financial Studies* 33 (2020): 2326–77.

FRIEDMAN, J. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38 (2002): 367–78.

GERAKOS, J., and R. GRAMACY. "Regression-Based Earnings Forecasts." *Chicago Booth Research Paper No. 12-26*, 2013.

GREEN, J.; J. HAND; and M. SOLIMAN. "Going, Going, Gone? The Apparent Demise of the Accruals Anomaly." *Management Science* 57 (2011): 797–816.

GU, S.; B. KELLY; and D. XIU. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies* 33 (2020): 2223–73.

HANLON, M. "The Persistence and Pricing of Earnings, Accruals, and Cash Flows When Firms Have Large Book-Tax Differences." *The Accounting Review* 80 (2005): 137–66.

HANLON, M., and T. SHEVLIN. "Accounting for Tax Benefits of Employee Stock Options and Implications for Research." *Accounting Horizon* 16 (2002): 1–16.

HARRIS, T., and S. MORSFIELD. "An Evaluation of the Current State and Future of XBRL and Interactive Data for Investors and Analysts." White Paper, Columbia University, 2012.

HASTIE, T.; R. TIBSHIRANI; and J. FRIEDMAN. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edition. New York: Springer, 2009.

HOLTHAUSEN, R., and D. LARCKER. "The Prediction of Stock Returns Using Financial Statement Information." *Journal of Accounting and Economics* 15 (1992): 373–411.

HOU, K.; C. XUE; and L. ZHANG. "Replicating Anomalies." *Review of Financial Studies* 33 (2020): 2019–133.

HSIEH, T.; and J. BEDARD. "Impact of XBRL on Voluntary Adopters' Financial Reporting Quality and Cost of Equity Capital." *Journal of Emerging Technologies in Accounting* 15 (2018): 45–65.

HUNT, J.; J. MYERS; and L. MYERS. "Improving Earnings Predictions with Machine Learning." Working paper, Mississippi State University, 2019.

JAPKOWICZ, N., and S. STEPHEN. "The Class Imbalance Problem: A Systematic Study." *Intelligent Data Analysis* 6 (2002): 429–49.

KIM, J.; B. LI; and Z. LIU. "Information-Processing Costs and Breadth of Ownership." *Contemporary Accounting Research* 36 (2019a): 2408–36.

KIM, J.; J. KIM; and J. LIM. "Does XBRL Adoption Constrain Earnings Management? Early Evidence from Mandated US Filers." *Contemporary Accounting Research* 36 (2019b): 2610–34.

KIRK, M.; J. VINCENT; and D. WILLIAMS. "From Print to Practice: XBRL Extension Use and Analyst Forecast Properties." Working paper, University of Florida, and University of Illinois, 2016. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2826159.

KOTHARI, S.P. "Capital Markets Research in Accounting." *Journal of Accounting and Economics* 31 (2001): 105–231.

LEV, B., and F. GU. *The End of Accounting and the Path Forward for Investors and Managers*. Hoboken, NJ: John Wiley & Sons Inc., 2016.

LEV, B., and D. NISSIM. "Taxable Income, Future Earnings, and Equity Values." *The Accounting Review* 79 (2004): 1039–74.

LI, F. "The Information Content of Forward-Looking Statements in Corporate Filings: A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (2010): 1049–102.

LI, K., and P. MOHANRAM. "Evaluating Cross-Sectional Forecasting Models for Implied Cost of Capital." *Review of Accounting Studies* 19 (2014): 1152–85.

LI, K., and P. MOHANRAM. "Fundamental Analysis: Combining the Search for Quality with the Search for Value." *Contemporary Accounting Research* 36 (2019): 1263–98.

LI, S., and E. NWAEZE. "The Association between Extensions in XBRL Disclosures and Financial Information Environment." *Journal of Information Systems* 29 (2015): 73–99.

LI, S., and E. NWAEZE. "Impact of Extensions in XBRL Disclosure on Analysts' Forecast Behavior." *Accounting Horizon* 32 (2018): 57–79.

LIU, M. "Assessing Human Information Processing in Lending Decisions: A Machine Learning Approach." Working paper, Boston University, 2021. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3483699.

LIVNAT, J., and J. SINGH. "Machine Learning Algorithms to Classify Future Returns Using Structured and Unstructured Data." *Journal of Investing* 30 (2021): 62–78.

Miller, G., and D. Skinner. "Determinants of the Valuation Allowance for Deferred Tax Assets Under SFAS No. 109." *The Accounting Review* 73 (1998): 213–33.

Monahan, S. "Financial Statement Analysis and Earnings Forecasting." *Foundations and Trends in Accounting* 12 (2018): 105–215.

Mullainathan, S., and J. Spiess. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2017): 87–106.

Nissim, D., and S. Penman. "Ratio Analysis and Equity Valuation: From Research to Practice." *Review of Accounting Studies* 6 (2001): 109–54.

Novy-Marx, R., and M. Velikov. "A Taxonomy of Anomalies and Their Trading Costs." *Review of Financial Studies* 29 (2016): 104–47.

Ou, J. "The Information Content of Nonearnings Accounting Numbers as Earnings Predictors." *Journal of Accounting Research* 28 (1990): 144–63.

Ou, J. and S. Penman. "Financial Statement Analysis and the Prediction of Stock Returns." *Journal of Accounting and Economics* 11 (1989): 295–329.

Pedregosa, F.; G. Varoquaux; A. Gramfort; V. Michel; B. Thirion; O. Grisel; M. Blondel; P. Prettenhofer; R. Weiss; V. Dubourg; J. Vanderplas; A. Passos; D. Cournapeau; M. Brucher; M. Perrot; and E. Duchesnay. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825–30. Available at https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html

Perols, J. "Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms." *Auditing: A Journal of Practice & Theory* 30 (2011): 19–50.

Piotroski, J., and E. So. "Identifying Expectation Errors in Value/Glamour Strategies: A Fundamental Analysis Approach." *Review of Financial Studies* 25 (2012): 2841–75.

Plumlee, R.D., and M. Plumlee. "Assurance on XBRL for Financial Reporting." *Accounting Horizon* 22 (2008): 353–68.

Rasekhschaffe, K., and R. Jones. "Machine Learning for Stock Selection." *Financial Analysts Journal* 75 (2019): 70–88.

Richardson, S.; I. Tuna; and P. Wysocki. "Accounting Anomalies and Fundamental Analysis: A Review of Recent Research Advances." *Journal of Accounting and Economics* 50 (2010): 410–54.

SEC. "Interactive Data to Improve Financial Reporting. Final Rule." 2009. Available at https://www.sec.gov/rules/final/2009/33-9002.pdf

SEC. "Staff Observations of Custom Axis Tags." 2016. Available at https://www.sec.gov/structureddata/reportspubs/osd_assessment_custom-axis-tags.html

Shumway, T. "The Delisting Bias in CRSP Data." *Journal of Finance* 52 (1997): 327–40.

Shumway, T., and V. Warther. "The Delisting Bias in CRSP's NASDAQ Data and Its Implications for the Size Effect." *Journal of Finance* 54 (1999): 2361–79.

Soliman, M. "The Use of DuPont Analysis by Market Participants." *The Accounting Review* 83 (2008): 823–53.

Thomas, J., and F. Zhang. "Tax Expense Momentum." *Journal of Accounting Research* 49 (2011): 791–821.

Varian, H. "Big data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2014): 3–28.

Wahlen, J., and M. Wieland. "Can Financial Statement Analysis Beat Consensus Analysts' Recommendations?" *Review of Accounting Studies* 16 (2011): 89–115.

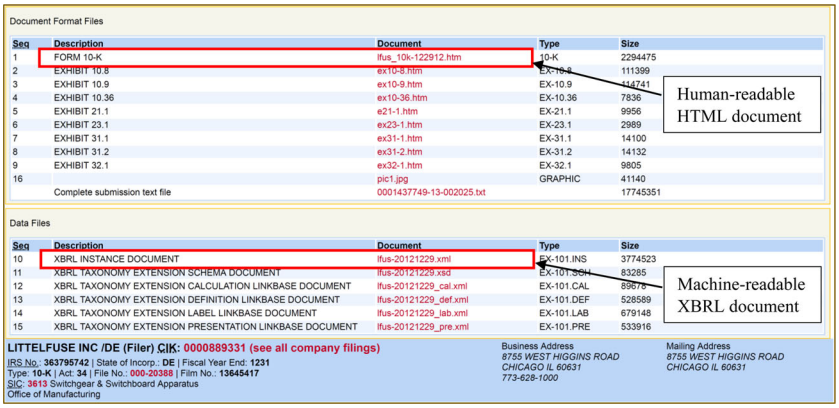
Zhou, Z. *Ensemble Learning Methods: Foundations and Algorithms.* Boca Raton, FL: CRC Press, 2012.

APPENDIX

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 Web site for Littelfuse, an electronic manufacturer.



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">23540400</us-gaap:CashAndCashEquivalentsAtCarryingValue>
<us-gaap:CashAndCashEquivalentsAtCarryingValue unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">16401600</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">10055900</us-gaap:AccountsReceivableNetCurrent>
<us-gaap:AccountsReceivableNetCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">9208800</us-gaap:AccountsReceivableNetCurrent>
<us-gaap:InventoryNet unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">7558000</us-gaap:InventoryNet>
<us-gaap:InventoryNet unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">7557500</us-gaap:InventoryNet>
<us-gaap:DeferredTaxAssetsNetCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">1189000</us-gaap:DeferredTaxAssetsNetCurrent>
<us-gaap:DeferredTaxAssetsNetCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">1189500</us-gaap:DeferredTaxAssetsNetCurrent>
<us-gaap:PrepaidExpenseAndOtherAssetsCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">1653200</us-gaap:PrepaidExpenseAndOtherAssetsCurrent>
<us-gaap:PrepaidExpenseAndOtherAssetsCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">1421900</us-gaap:PrepaidExpenseAndOtherAssetsCurrent>
<us-gaap:AssetsHeldForSaleCurrent unitRef="usd" contextRef="c0_AsOf29Dec2012" decimals="-3">550000</us-gaap:AssetsHeldForSaleCurrent>
<us-gaap:AssetsHeldForSaleCurrent unitRef="usd" contextRef="c1_AsOf31Dec2011" decimals="-3">659200</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 thous

	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>
```