Journal of Accounting Research

DOI: 10.1111/1475-679X.12292
Journal of Accounting Research
Vol. 58 No. 1 March 2020
Printed in U.S.A.

CHICAGO BOOTH

Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach

YANG BAO,* BIN KE,† BIN LI,‡ Y. JULIA YU,§ AND JIE ZHANG‖

Received 7 October 2015; accepted 1 October 2019

ABSTRACT

We develop a state-of-the-art fraud prediction model using a machine learning approach. We demonstrate the value of combining domain knowledge and machine learning methods in model building. We select our model input based on existing accounting theories, but we differ from prior accounting research by using raw accounting numbers rather than financial ratios.

*Antai College of Economics and Management, Shanghai Jiao Tong University; †Department of Accounting, NUS Business School, National University of Singapore; ‡Department of Finance, Economics and Management School, Wuhan University; §McIntire School of Commerce, University of Virginia; ‖School of Computer Engineering, Nanyang Technological University.

Accepted by Christian Leuz. We wish to thank an anonymous reviewer, Mark Cecchini, Luo Zuo, and workshop participants at the Singapore Tri-Uni Accounting Research Conference, the Inaugural Conference on Intelligent Information Retrieval in Accounting and Finance at CUHK (Shenzhen), and HKUST for helpful comments. Part of this research is funded by a Singapore Ministry of Education Tier 2 grant (No. MOE2012-T2-1-045). Yang Bao acknowledges the financial support from a NSFC grant (No. 71601116) and Shanghai Pujiang Program (No. 16PJC045). Ke Bin acknowledges the financial support from an MOE start-up grant (No. R-521-000-032-133). Bin Li acknowledges the financial support from National Natural Science Foundation of China (71971164, 91646206). An online appendix to this paper can be downloaded at http://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements. The codes and data used for our best model RUSBoost are available at the Github repository: https://github.com/JarFraud/FraudDetection.

We employ one of the most powerful machine learning methods, ensemble learning, rather than the commonly used method of logistic regression. To assess the performance of fraud prediction models, we introduce a new performance evaluation metric commonly used in ranking problems that is more appropriate for the fraud prediction task. Starting with an identical set of theory-motivated raw accounting numbers, we show that our new fraud prediction model outperforms two benchmark models by a large margin: the Dechow et al. logistic regression model based on financial ratios, and the Cecchini et al. support-vector-machine model with a financial kernel that maps raw accounting numbers into a broader set of ratios.

JEL codes: C53; M41

Keywords: fraud prediction; machine learning; ensemble learning

1. Introduction

Accounting fraud is a worldwide problem. If not detected and prevented on a timely basis, it can cause significant harm to the stakeholders of fraudulent firms (e.g., Enron and WorldCom) as well as the stakeholders of many nonfraudulent firms indirectly (Gleason, Jenkins, and Johnson [2008], Goldman, Peyer, and Stefanescu [2012], Hung, Wong, and Zhang [2015]). Unfortunately, accounting fraud is difficult to detect. Moreover, even if it is detected, serious damage has usually already been done (Dyck, Morse, and Zingales [2010]). Hence, efficient and effective methods of corporate accounting fraud detection would offer significant value to regulators, auditors, and investors.

The objective of this study is to develop a new accounting fraud prediction model *out of sample* by using readily available financial statement data from publicly traded U.S. firms. Following Cecchini et al. [2010], and Dechow et al. [2011], we use the detected material accounting misstatements disclosed in the SEC's Accounting and Auditing Enforcement Releases (AAERs) as our accounting fraud sample. Although there are useful nonfinancial predictors of accounting fraud (e.g., an executive's personal behavior), we use only readily available financial data for two reasons. First, fraud prediction models based on publicly available financial data can be applied to any publicly traded firm at low cost. Second, most of the fraud prediction models in the existing accounting literature also rely on publicly available financial data (e.g., Green and Choi [1997], Summers and Sweeney [1998], Beneish [1999], Cecchini et al. [2010], Dechow et al. [2011]). By limiting the predictors to financial data only, the performance of our fraud prediction models can be compared with the performance of such existing models.

There is a fairly large accounting literature on the determinants of accounting fraud (e.g., Entwistle and Lindsay [1994], Beasley [1996], Dechow, Sloan, and Sweeney [1996], Beneish [1997, 1999], Summers and Sweeney [1998], Efendi, Srivastava, and Swanson [2007], Brazel, Jones, and

Zimbelman [2009], Dechow et al. [2011], Schrand and Zechman [2012]), but the primary objective of most studies is to explain fraud *within sample* and often emphasize causal inference. Our objective is different: We wish to develop a model that can accurately predict accounting fraud *out of sample* (i.e., a prediction problem). Shmueli [2010] shows that the problems of causal inference and prediction, although related, are fundamentally different. Specifically, the objective of causal inference modeling is to minimize the bias resulting from model misspecification to obtain the most accurate representation of the underlying theory. In contrast, the objective of predictive modeling seeks to minimize out-of-sample prediction error, that is, the combination of the bias and estimation variance that results from using a sample to estimate model parameters.[1] Although causal inference represents the main stream of existing social science research, Kleinberg et al. [2015] show that there are many interesting prediction problems that are neglected in the extant business and economics literatures.

We use two types of fraud prediction models from the extant literature as benchmarks. The first is ratio-based logistic regression, commonly used in the accounting literature (e.g., Beneish [1997, 1999], Summers and Sweeney [1998], Dechow et al. [2011]). Such models typically use financial ratios as predictors; the ratios are often identified by human experts based on theories (e.g., the motivation-ability-opportunity framework from the criminology literature). Among these models, the model in Dechow et al. [2011] is generally regarded as the most comprehensive fraud prediction model in accounting literature. Accordingly, we adopt a similar logistic regression model as our first benchmark model (referred to as the Dechow et al. model). The second benchmark model is a fraud prediction model developed by Cecchini et al. [2010] based on a more advanced machine learning method (hereafter referred to as the Cecchini et al. model). Rather than using the financial ratios identified by human experts alone, Cecchini et al. [2010] develop a new fraud prediction model based on support vector machines (SVM) with a financial kernel that maps raw financial data into a broader set of ratios within the same year and changes in ratios across different years. Cecchini et al. [2010] find that the SVM with a financial kernel outperforms the traditional fraud prediction models in accounting, including the Dechow et al. model.[2]

Our proposed fraud prediction model differs from both of these benchmark models in two key ways. First, we use ensemble learning, a state-of-the-art machine learning paradigm, to predict fraud. Most prior fraud prediction research in accounting uses the logistic regression (see Dechow et al. [2011] for a review). Although ensemble learning has been successfully

[1] See the online appendix for a more detailed discussion on the differences between causal inference and prediction.

[2] It is important to note that the performance results of our Dechow et al. model and Cecchini et al. model are not directly comparable to those of Dechow et al. [2011] and Cecchini et al. [2010] because of a few crucial research design differences, explained in section 3.

applied in many other fields (see Zhou [2012] for a review), ours is the first study to apply the method to an accounting setting with a severe class imbalance problem (i.e., the rarity of fraud). Fernandez-Delgado et al. [2014] show that there is no universally best model across all data settings; hence, it is an empirical question whether the ensemble learning method can outperform the traditional fraud prediction methods in our special setting.

Second, our proposed model uses raw financial data items, taken directly from financial statements, as fraud predictors. Because raw financial data items are the most fundamental building blocks of the accounting system, it is interesting to explore whether they can be directly used in fraud prediction. Ex ante, it is unclear whether fraud prediction models based on raw financial data can outperform fraud prediction models based on human expert–identified financial ratios. On the one hand, fraud prediction models based on financial ratios could be more powerful because the ratios identified by human experts are often grounded in theories that offer sharp prediction on when corporate managers have incentives to engage in fraud. Because fraud prediction models based on raw financial data are not directly linked to theory, they may be less powerful. On the other hand, existing theories about the drivers of accounting fraud may well be incomplete, as accounting fraud is, by definition, conducted in secrecy and designed to be difficult to detect. Accordingly, converting raw accounting data into a limited number of financial ratios based on potentially incomplete behavioral theories could mean the loss of useful predictive information. In contrast, fraud prediction models that make use of raw financial data could be more powerful because they do not impose any ex ante structure on the raw data, instead letting them "speak for themselves." In addition, with the rapid advance of machine learning methods in computer science, fraud prediction models based on raw data can take on more flexible and complex functional forms. As a result, such fraud prediction models may be able to extract more useful information from raw data. Because of these conflicting trade-offs, we believe it is an empirical question on whether our proposed ensemble learning model based on raw data can outperform the two benchmark models based on financial ratios.

To compare the out-of-sample performance of different fraud prediction models, we adopt two distinctive performance evaluation metrics. First, we follow Larcker and Zakolyukina [2012] by using the area under the Receiver Operating Characteristics (ROC) curve (AUC) as a performance evaluation metric. The AUC is equivalent to the probability that a randomly chosen fraud observation will be ranked higher by a classifier than will a randomly chosen nonfraud observation (Fawcett [2006]). The AUC for random guesses is 0.50. Therefore, any reasonable fraud prediction model must have an AUC higher than 0.50.

Second, we introduce an alternative performance evaluation metric commonly used for ranking problems, referred to as Normalized Discounted Cumulative Gain at the position k (NDCG@k). Because of the infrequency

with which accounting fraud is identified by the SEC's AAERs, even the best-performing fraud prediction model (e.g., Cecchini et al. [2010]) would result in a large number of false positives that far exceed the number of true positives in a test period. Clearly, it is impractical for regulators or corporate monitors to investigate all predicted cases of fraud, given the limited resources available to fight such fraud (Ernst & Young [2010]). Even if one wished to investigate all predictions of fraud, the direct and indirect costs would be huge, while the benefit would be small (because the majority of the predicted fraud observations are false positives). Naturally, then, regulators and other monitors seek to investigate the smallest number of observations with the highest predicted likelihood of fraud. Accordingly, we also evaluate the out-of-sample performance of the different fraud prediction models using NDCG@k. Intuitively, NDCG@k assesses the ability of a fraud prediction model to identify actual fraud by picking the top k observations in a test year that have the highest predicted probability of fraud. In our study, we pick a k that equals the top 1% of the observations. We select a cutoff of 1% because typically less than 1% of the firms in a year are fraud per the SEC's AAERs. The values of NDCG@k are bounded between 0 and 1.0, with a higher value representing better model performance.

To put the performance evaluation of all fraud prediction models on an equal footing, we require that all models start with a common set of raw financial data. Although a company's three financial statements contain hundreds of readily available raw financial data, the two benchmark models only use a combined total of 28 raw financial data items derived from 14 expert-identified financial ratios. Hence, we use these 28 raw financial data items as the starting point for the performance evaluation of all fraud prediction models. The Dechow et al. model uses 14 financial ratios derived from the 28 raw data items, whereas the Cecchini et al. model uses a financial kernel that maps the 28 raw data items into a broader set of ratios.

Many of our accounting fraud cases span multiple consecutive years (such cases are referred to as serial fraud). As we explain in section 3.3, if cases of serial fraud span both the training and test periods, they could artificially inflate the performance of more flexible machine learning models such as ensemble learning. To deal with this concern, we recode all the fraud observations in the training period as nonfraud if a case of serial fraud spans both the training and test periods. In section 7.3, we show direct evidence that failing to correct the serial fraud problem would significantly inflate the performance of the ensemble learning model.

Our sample covers all publicly listed U.S. firms over the period 1991–2008. Our sample starts from 1991 because there is a significant shift in U.S. firms' fraudulent behavior as well as the nature of SEC enforcement starting around that time. Our sample ends in 2008 because the regulators reduced the enforcement of accounting fraud starting from around 2009, increasing the possibility that many accounting fraud cases remain undetected for the post-2008 period (see Rakoff [2014] and section 3); in addition, it takes time for the SEC to identify and prosecute the

alleged fraud cases, especially in the post-2008 period with reduced public resources for accounting fraud investigations. Nevertheless, our inferences are qualitatively similar if we choose a shorter period of 1991–2005 or a longer period of 1991–2011 or 1991–2014.

We first report the out-of-sample performance results for the two benchmark models. Using AUC as the performance evaluation metric, we find that the out-of-sample performance of the two benchmark models is significantly better than the performance of random guesses. The average AUC is 0.672 for the Dechow et al. model and 0.626 for the Cecchini et al. model. We find no evidence that the Cecchini et al. model outperforms the Dechow et al. model, suggesting that adopting more advanced machine learning methods, per se, does not necessarily translate into better performance. Inferences are qualitatively similar if we use NDCG@k as our performance evaluation metric. Specifically, the average value of NDCG@k is 0.028 for the Dechow et al. model and 0.020 for the Cecchini et al. model.

We next examine whether combining raw data with a flexible and powerful machine learning model, ensemble learning, can lead to better prediction performance than that exhibited by the two benchmark models. We find two key results. First, we find no clear evidence that a logistic regression model based on the 28 raw financial data items yields better prediction performance than the two benchmark models. The average AUC for the logistic regression model based on raw data is 0.690, higher than the average AUC for both benchmark models. However, the average NDCG@k for the logistic model based on the 28 raw data items is only 0.006, lower than the average NDCG@k for both benchmark models. Second, we find that an ensemble learning model based on the same 28 raw financial data items yields much better prediction performance than the two benchmark models. The average AUC and the average NDCG@k for the ensemble learning model are 0.725 and 0.049, respectively, representing a performance increase of 7.9% and 75%, respectively, relative to the performance of the better benchmark model, the Dechow et al. model. These performance differences are also economically significant: Using the NDCG@k approach (where $k = 1\%$), our best model, the ensemble learning model, identified a total of 16 fraud cases in the test period 2003–08 whereas the comparable figure is 9 for the Dechow et al. model and 7 for the Cecchini et al. model. These results suggest that the direct use of raw financial data, coupled with a more flexible and powerful machine learning method, can yield better fraud prediction.

We also examine whether an ensemble learning model based on the 14 financial ratios, or the combination of the 14 financial ratios and the 28 raw data items, can yield even better fraud prediction performance. Our results show no evidence that the above two alternative ensemble learning models outperform the ensemble learning model that makes use of the 28 raw data items alone. These results provide further evidence of the predictive benefit of combining raw financial data with the powerful ensemble learning method.

An interesting question is whether it is possible to improve the performance of our best fraud prediction model, ensemble learning, by adding additional raw financial data items readily available from existing commercial databases. To find out, we include 266 additional raw financial data items readily available from Compustat in the ensemble learning model. Using either AUC or NDCG@k as the performance evaluation metric, we find no evidence that including a long "laundry list" of raw financial data items helps improve the out-of-sample performance of the ensemble learning model. Although we cannot rule out the possibility that more data will produce better prediction models, this preliminary evidence suggests that dumping a large quantity of raw data items into the ensemble learning model without any theoretical guidance does not necessarily translate into higher out-of-sample prediction performance.

Our study joins a small but growing accounting literature that uses financial statement data to predict accounting fraud out of sample. Representative studies in this small literature include Cecchini et al. [2010] and Perols et al. [2017]. We have highlighted the differences between Cecchini et al. and our study above. Perols et al. [2017] and our study share both similarities and differences. Like us, Perols et al. [2017] use different statistical methods to deal with the unique challenges of fraud prediction, such as rarity of detected fraud, serial fraud, and abundance of fraud predictors. However, Perols et al. [2017] adopt a very stringent definition of accounting fraud and focus on large firms with analysts following. As a result, their final sample contains only 51 fraud firms over the period 1998–2005. More importantly, our study differs from these existing studies in several key aspects. First, we introduce a state-of-the-art and powerful machine learning method, ensemble learning. Our results suggest that ensemble learning, if properly used, is more powerful than logistic regression and SVM for the purposes of fraud prediction.

Second, we are the first study to assess the usefulness of using raw financial data, rather than ratios derived from raw financial data, for the purposes of fraud prediction. Our empirical investigations provide preliminary evidence that it is possible to produce more powerful fraud prediction models by carefully selecting—with the aid of theoretical guidance—a small set of raw financial data, then coupling that data with a powerful machine learning method. Our results also raise the exciting possibility that we can further improve fraud prediction by using additional readily available raw financial data guided by new theory.

Third, we introduce to the fraud prediction literature a new performance evaluation metric, NDCG@k. Compared with the commonly used performance evaluation metric AUC, which has been used in prior literature (e.g., Larcker and Zakolyukina [2012], Perols et al. [2017]), NDCG@k is more useful to regulators and other monitors. This is because regulators and other monitors often face significant resource constraints; therefore, they can only investigate a small number of alleged fraud cases. Because NDCG@k measures a model's prediction performance by picking

the top k firms with the highest predicted probability of fraud, it offers a simple decision rule for regulators and other monitors to identify the most suspicious firms for investigation.

The results from this study also have important implications for the ongoing accounting research that compares the usefulness of textual data versus quantitative data in predicting accounting frauds (e.g., Larcker and Zakolyukina [2012]). A typical benchmark model used in this line of research is Dechow et al. [2011]. Our results raise the bar for this line of text-mining research because we show that the commonly used Dechow et al. ratio-based logistic regression model significantly understates the value of financial data in fraud prediction.

2. *An Introduction to Ensemble Learning*

Ensemble learning, a primary paradigm of machine learning, has recently achieved notable success in many real-word applications (see pages 17–19 in Zhou [2012] for a review of applications of ensemble learning methods). Unlike conventional machine learning methods (e.g., SVM methods), which usually generate a single estimator, ensemble learning methods combine the predictions of a set of base estimators (e.g., decision trees) to improve the generalization ability and robustness. Previous studies (Zhou [2012]) have shown that ensembles usually outperform any single base estimator. However, due to the possibility of a class imbalance problem, conventional ensemble learning methods usually need to be combined with a sampling technique that balances the class distribution of training data, by either adding examples to the minority class (oversampling) or removing examples from the majority class (undersampling; Liu and Zhou [2013]). In this study, we employ one variation of ensemble learning called RUSBoost (Seiffert et al. [2010]). RUSBoost seeks to take advantage of both the efficient undersampling technique (Liu, Wu, and Zhou [2009]) and the current most influential ensemble algorithm, AdaBoost (Freund and Schapire [1997]). A review by Galar et al. [2012] finds that RUSBoost shows the best performance and is also more computationally efficient due to its simplicity (Seiffert et al. [2010]). The superiority of RUSBoost was further confirmed by Khoshgoftaar, Van Hulse, and Napolitano [2011].

In the paragraphs that follow, we first introduce the AdaBoost algorithm, then describe how it is combined with an undersampling technique in our empirical implementation of RUSBoost. The AdaBoost algorithm is one of the most important ensemble learning methods because of its solid theoretical foundation, strong predictive power, and simplicity (Wu et al. [2008]). Its basic idea is to "train" a sequence of weak classifiers (i.e., models that are only slightly better than random guesses, e.g., small decision trees) on repeatedly weighted samples. Specifically, in each iteration, the weights of the incorrectly classified observations will be increased, whereas the weights of correctly classified observations will be decreased. In this way, the weak classifiers in each iteration will be forced to concentrate on

the observations that were difficult to predict in the previous iterations. Finally, a strong classifier can be produced by taking the weighted average of all weak classifiers, where the weight is based on a weak classifier's classification error rate in the training sample. The weak classifiers with lower classification error rates will receive higher weights.

RUSBoost is a variant of AdaBoost that makes use of random undersampling (RUS) to address the problem of class imbalance learning (Seiffert et al. [2010]). It works in much the same way as AdaBoost, except that RUS is performed in each iteration to address the imbalance of fraudulent and nonfraudulent firms. Specifically, when training the weak classifier in each iteration, the RUS algorithm uses the full sample of fraudulent firms in the training period and a randomly generated subsample of nonfraudulent firms in the same training period.[3] RUSBoost estimates require selection of the ratio between the number of undersampled majority class observations (i.e., nonfraud) and the number of minority class observations (i.e., fraud). In this paper, we construct our RUSBoost model by setting this ratio at 1:1. That is, we simply sample the same number of fraud observations and nonfraudulent observations.

3. The Sample and Data

3.1 THE SAMPLE PERIOD

Our sample covers all publicly listed U.S. firms during the period 1991–2008. We start the sample in 1991 because there is a significant shift in U.S. firms' fraudulent behavior as well as the nature of SEC enforcement starting around that time. In their review of the history and evolution of the SEC enforcement program, Atkins and Bondi [2008] suggest that the purpose of the SEC's enforcement program shifted from remedial to punitive in the 1990s. Prior to 1990, the SEC's statutory purpose was to provide remedial relief for aggrieved investors and to deter future violations. But in the mid- to late 1980s, Congress passed a series of laws that expanded the SEC's powers and provided it with new penalizing authority. As a result of these laws, the SEC gained the power to seek or impose more punitive actions, such as (1) the ability to seek civil monetary penalties against persons and entities that may have violated federal securities laws, (2) the authority to bar directors and officers of public companies from serving in those capacities if they have violated federal antifraud provisions, and (3) the authority to issue administrative cease-and-desist orders, temporary restraining orders, and orders for disgorgement of ill-gotten profits to violators of federal securities laws (Atkins and Bondi [2008]).

---

[3] We fix the seed of the random number generator to zero to ensure the replicability of our reported results. This is a commonly used method for reproducing the experimental results in computer science. For more details, please refer to http://www.mathworks.com/help/matlab/math/generate-random-numbers-that-are-repeatable.html.

In addition, the use of stock options as a form of executive compensation rose dramatically during the 1990s, as did other similar pay-for-performance plans such as restricted stock and bonus plans tied to performance (Murphy [1999], Erickson, Hanlon, and Maydew [2006]). Consequently, we observe more frequent citations of insider trading as a possible motive for accounting fraud in the AAERs published in the 1990s than in the 1980s (Beasley, Carcello, and Hermanson [1999] and Beasley et al. [2010]).[4]

Finally, an analysis of the instances of accounting fraud included in the AAERs in the 1980s and 1990s by Beasley et al. ([1999] and [2010], respectively) reveals subtle changes, over time, in the nature of the fraud. In both time periods, the two most common techniques used to fraudulently misstate financial statement information involved overstating revenues and assets. However, misstatements through the understatement of expenses/liabilities became a more frequently used fraud technique during the 1990s (an increase from 18% of cases to 31% of cases).

We end the sample in 2008 because there is a noticeable shift in the regulators' enforcement of accounting fraud that approximately coincided with the 2008 financial crisis. Jed Rakoff [2014], the U.S. district judge for the Southern District of New York, which has the most experience in dealing with financial fraud of publicly listed firms, summarized the regime shift as follows. First, the FBI, which had more than 1,000 agents assigned to investigating financial frauds before 2001, shifted many of these agents to antiterrorism work after the 9/11 terrorist attack. Second, to deflect criticism from its failure to detect the Madoff fraud, which was discovered in 2008, the SEC shifted its focus from accounting fraud investigations to other Ponzi-like schemes that emerged in the wake of the financial crisis.[5] Third, the Department of Justice made a decision in 2009 to spread the investigation of financial fraud cases among numerous U.S. Attorney's Offices, many of which had little or no previous experience in investigating and prosecuting sophisticated financial frauds. At the same time, the U.S. Attorney's Office with the greatest expertise in these kinds of cases, the Southern District of New York, was just embarking on its prosecution of insider-trading cases arising from the Raj Rajaratnam tapes that absorbed a huge amount of the attention of the securities fraud unit of that office. Because of these significant downward shifts in the public enforcement of accounting fraud cases around the 2008 financial crisis, Judge Rakoff was worried that many

[4] However, there is mixed evidence from the extant academic literature on the effect of managerial equity compensation on accounting frauds (e.g., Erickson et al. [2006], Johnson, Ryan, and Tian [2009], Armstrong, Jagolinzer and Larcker [2010]).

[5] Andrew Ceresney, Co-Director of the SEC's Division of Enforcement, also made similar remarks in a public speech, saying "In the wake of the financial crisis, the SEC was very focused on financial crisis cases—cases involving CDOs, RMBS, Ponzi schemes, and other transactions that resulted in massive losses to investors. Consequently, we devoted fewer resources to accounting fraud. During this period, we have had fewer accounting fraud investigations" (Ceresney [2013]).

accounting fraud cases that occurred around and after the financial crisis would remain undetected forever.[6]

All of our fraud prediction models require a training period and a test period. To ensure the reliability of model training, we require the training period to exceed 10 years. In addition, we require a gap of 24 months between the financial results announcement of the last training year and the results announcement of a test year. We do this because Dyck, Morse, and Zingales [2010] find that it takes approximately 24 months, on average, for the initial disclosure of the fraud.[7] Therefore, we use the last six years of our sample 2003–08 as the test period. For example, the training period is 1991–2001 for test year 2003 and 1991–2003 for test year 2005.

3.2 THE FRAUD SAMPLE

Our accounting fraud sample comes from the SEC's AAERs provided by the University of California-Berkeley Center for Financial Reporting and Management (CFRM). Prior fraud research has typically used four popular databases: the CFRM database, the Government Accountability Office's (GAO) earnings restatement database, Audit Analytics' (AA) earnings restatement database, and the Stanford Securities Class Action Clearinghouse (SCAC) database of securities class action lawsuits filed since the passage of the Private Securities Litigation Reform Act of 1995. We chose the CFRM database for our empirical analyses for two key reasons. First, our research question requires the accurate identification of all cases of accounting fraud, and the findings of Karpoff et al. [2017] indicate that the CFRM database is the best for this purpose. Specifically, Karpoff et al. [2017] provide a comprehensive investigation of the advantages and disadvantages of the aforementioned four databases. Their findings suggest that no single database dominates and the choice of database depends on a researcher's specific research question. For example, the CFRM database would be a poor choice for researchers conducting an event study of the stock market's reaction to the initial revelation of an accounting fraud. However, Karpoff et al. [2017] find that CFRM ranks first if one wishes to identify a comprehensive list of fraud cases (see their table 8).

Second, we wish to compare the performance of our proposed models with the performance of the two benchmark models from Dechow et al. [2011] and Cecchini et al. [2010]. Because these benchmark models make

---

[6] Because fraud investigations typically take years to finish, there is a possibility that the investigations of many accounting frauds that occurred before 2009 could be also negatively affected by this regime shift. Hence, we also conduct a robustness check using a shorter test period 2003–05.

[7] Despite the 24-month gap between the training period and test period, we wish to acknowledge that, just like prior research, all of our fraud prediction models considered in this study may not be directly implemented by regulators or other monitors in real time. This is because some of the fraud observations in our training period may not be publicly known at the time of the model construction.

use of AAER data, it makes sense for us to use the same data source. However, we acknowledge that the CFRM database is subject to its own potential selection bias (Kedia and Rajgopal [2011], deHaan et al. [2015]), a limitation that readers should keep in mind when interpreting our results.[8]

The version of the CFRM database we obtained in March 2017 covers the period from May 17, 1982, to September 30, 2016. Because the CFRM has not updated its database ever since, we hand-collected additional fraud observations from the SEC website (https://www.sec.gov/divisions/enforce/friactions.shtml) for the period up to December 31, 2018 (AAER 4012). Table 1 shows the distribution of fraud in our sample by year over the fiscal years 1979–2014. The AAERs cover instances of accounting fraud that occurred as early as 1971, but there are only 13 fraudulent firm-years before 1979. Hence, we omit the years prior to 1979. We tabulate the fraud observations up to 2014 because it takes several years for the SEC to finish the investigations of alleged fraud cases (Karpoff et al. [2017]).

As shown in table 1, there were 1,171 detected fraudulent firm-years in total over 1979–2014, but the frequency of detected fraud is very low, typically less than 1% of all firms per year. The rarity of detected accounting fraud highlights the ongoing challenge of fraud prediction. In addition, the observed frequency of fraud declines almost monotonically over the test period 2003–2014. For instance, the average frequency of fraud is 0.94% for 2003–2005, 0.51% for 2006–2008, 0.46% for 2009–2011, and 0.21% for 2012–2014. If one assumes that the true incidence of fraud remains the same over 2003–2014,[9] the declining frequency of fraud in table 1 suggests an increase in undetected fraud over time, consistent with the downward shift in the regulators' enforcement of accounting fraud since the 2008 financial crisis discussed above. Another possible reason is that it takes longer for the regulators to finish many accounting fraud investigations because of the reduced resources allocated to accounting fraud investigations in the postcrisis period. Because we start with a very small number of detected fraud observations each year, the presence of a significant number of hidden fraud observations in a test year could completely alter inferences in

[8] One could combine the four fraud databases to derive a single comprehensive database of accounting frauds. However, as noted by Dechow et al. [2011] and Karpoff et al. [2017], the other three databases contain a lot of measurement error in terms of identifying the fraud cases and therefore building such a comprehensive database would be time consuming and beyond the scope of this study.

[9] In this regard, it is interesting to note the remarks by the SEC's Co-Director of the Division of Enforcement Andrew Ceresney [2013] in a public speech: "But I have my doubts about whether we have experienced such a drop in actual fraud in financial reporting as may be indicated by the numbers of investigations and cases we have filed. It may be that we do not have the same large-scale accounting frauds like Enron and Worldcom. But I find it hard to believe that we have so radically reduced the instances of accounting fraud simply due to reforms such as governance changes and certifications and other Sarbanes-Oxley innovations. The incentives are still there to manipulate financial statements, and the methods for doing so are still available. We have additional controls, but controls are not always effective at finding fraud."

TABLE 1
Distribution of Fraud Firms by Year over 1979–2014

Year	Total Number of Firms	Number of Fraud Firms	Percentage
1979	3,782	4	0.11
1980	4,010	10	0.25
1981	4,501	12	0.27
1982	4,718	19	0.40
1983	5,056	14	0.28
1984	5,100	16	0.31
1985	5,087	10	0.20
1986	5,234	21	0.40
1987	5,406	16	0.30
1988	5,129	19	0.37
1989	4,977	23	0.46
1990	4,893	18	0.37
1991	4,981	28	0.56
1992	5,209	28	0.54
1993	5,644	31	0.55
1994	5,966	24	0.40
1995	6,561	22	0.34
1996	7,095	34	0.48
1997	7,138	45	0.63
1998	7,042	56	0.80
1999	7,230	79	1.09
2000	7,153	92	1.29
2001	6,780	87	1.28
2002	6,475	81	1.25
2003	6,285	69	1.10
2004	6,218	60	0.96
2005	6,119	47	0.77
2006	6,130	35	0.57
2007	6,073	30	0.49
2008	5,817	27	0.46
2009	5,618	30	0.53
2010	5,585	26	0.47
2011	5,583	22	0.39
2012	5,814	21	0.36
2013	5,831	11	0.19
2014	5,786	4	0.07
Total	**206,026**	**1,171**	**0.57**

Table 1 shows the yearly fraud percentages after merging the AAERs as of December 31, 2018, with the COMPUSTAT fundamental data over the years 1979–2014.

out-of-sample tests. For this reason, we use the years 2003–2008 as our primary test sample, even though we also show results using the following alternative test samples: 2003–2005, 2003–2011, and 2003–2014.

3.3 SERIAL FRAUD

Accounting fraud may span multiple consecutive reporting periods, creating a situation of so-called "serial fraud." In our sample, the mean, median, and 90th percentile of the duration of the disclosed accounting fraud cases is two years, two years, and four years, respectively, suggesting that it is

common for a case of fraud to span multiple consecutive reporting periods. Such serial fraud may overstate the performance of the ensemble learning method if instances of fraudulent reporting span both the training and test periods. This is because ensemble learning is more flexible and powerful than the logistic regression model, and may therefore be better able to fit a fraudulent firm than a fraudulent firm-year.[10] Hence, enhanced performance of the ensemble learning method may result from the fact that both the training and test samples contain the same fraudulent firm; the ensemble learning model may not perform as well when the sample contains different firms. To deal with this concern, we break up those cases of serial fraud that span both the training and test periods. Because we have a small number of fraudulent firm-years relative to the number of nonfraudulent firm-years in any test year, we recode all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods. Although this approach helps us avoid the problems associated with serial fraud, it may also introduce measurement errors into the training data.

3.4 RAW FINANCIAL DATA

The list of raw financial data items is selected based on Cecchini et al. [2010] and Dechow et al. [2011]. Our initial list of raw financial data items is selected following Cecchini et al. [2010]. After reviewing a comprehensive list of academic papers (including Summers and Sweeney [1998], Beneish [1999], Dechow et al. [2011], and Green and Choi [1997]), Cecchini et al. [2010, table 3] identified an initial list of 40 raw financial data items used in prior fraud prediction research to construct the regression variables. Cecchini et al. retained a final list of 23 raw financial data items after imposing the requirement that no raw variable have more than 25% of its values missing. Following the same sample selection procedures, we obtain a list of 24 raw financial data items during our sample period 1991–2008. Table 2 shows the list of the initial 40 raw financial data items from Cecchini et al. [2010] in column 1 and our list of 24 raw financial data items in column 2.

We also identify an initial list of raw financial data items based on Dechow et al. [2011]. Because the aim of our study is to use data from readily available financial statements to predict accounting fraud, we do not simply replicate the Dechow et al. models in table 7. Specifically, we begin with the initial larger list of candidate regression variables identified by Dechow et al. [2011, table 3]. Dechow et al. [2011, table 3] suggest four types of fraud determinants: (1) "accruals quality-related variables": nine variables, which can be calculated from the numbers in annual financial statements such as balance sheets and income statements; (2) "performance variables": five variables that gauge a firm's financial performance on various

[10] We thank the anonymous referee for raising this point.

TABLE 2

List of Variables Selected in Replicating Dechow et al. [2011] and Cecchini et al. [2010]

	(1) 40 Raw Data Items from Cecchini et al. [2010]	(2) 24 Raw Data Items in Our Replication of Cecchini et al. [2010]	(3) 11 Financial Ratios Used in the Basic Dechow et al. [2011] Model
Balance sheet items			
Cash and short-term investments	Yes	Yes	Yes
Receivables, total	Yes	Yes	Yes
Receivables, estimated doubtful	Yes	–	–
Inventories, total	Yes	Yes	Yes
Short-term investments, total	Yes	Yes	Yes
Current assets, total	Yes	–	Yes
Property, plant and equipment, total	Yes	Yes	Yes
Investment and advances, other	Yes	Yes	Yes
Assets, total	Yes	Yes	Yes
Accounts payable, trade	–	–	Yes
Debt in current liabilities, total	Yes	Yes	Yes
Income taxes payable	Yes	Yes	Yes
Rental commitments minimum 1st year	Yes	–	–
Current liabilities, total	Yes	Yes	Yes
Long-term debt, total	Yes	Yes	Yes
Rental commitments minimum 2nd year	Yes	–	–
Rental commitments minimum 3rd year	Yes	–	–
Rental commitments minimum 4th year	Yes	–	–
Rental commitments minimum 5th year	Yes	–	–
Liabilities, total	Yes	Yes	Yes
Common/ordinary equity, total	Yes	Yes	Yes
Preferred/preference stock (capital), total	Yes	Yes	Yes
Retained earnings	Yes	Yes	–
Income statement items			
Sales/turnover (net)	Yes	Yes	Yes
Cost of goods sold	Yes	Yes	Yes
Depreciation and amortization	Yes	Yes	–
Depreciation expense (Schedule VI)	Yes	–	–
Selling, general, and administrative expense	Yes	–	–
Interest and related expense, total	Yes	Yes	–
Interest and related income, total	Yes	–	–
Income taxes, total	Yes	Yes	–
Income taxes, deferred	Yes	–	–
Income before extraordinary items	Yes	Yes	Yes
Net income (loss)	Yes	Yes	–
Cash flow statement items			
Long-term debt issuance	–	–	Yes
Sale of common and preferred stock	–	–	Yes
Financing activities net cash flow	Yes	–	–
Market value items			
Price close, annual, fiscal	Yes	Yes	Yes

(Continued)

TABLE 2—*Continued*

	(1) 40 Raw Data Items from Cecchini et al. [2010]	(2) 24 Raw Data Items in Our Replication of Cecchini et al. [2010]	(3) 11 Financial Ratios Used in the Basic Dechow et al. [2011] Model
Price close, annual, calendar	Yes	–	–
Common shares outstanding	Yes	Yes	Yes
Other disclosure items			
Employees	Yes	–	–
Order backlog	Yes	–	–
Pension plans, anticipated long-term rate of return on plan assets	Yes	–	–

Column 1 lists the initial 40 raw data items selected by Cecchini et al. [2010] and column 2 lists the 24 raw data items retained in our final analysis after deleting variables with more than 25% missing values. Column 3 shows the 23 raw data items used to calculate 11 financial ratios from Dechow et al. [2011]. We combine the raw data items in columns 2 and 3 to obtain a total of 28 unique raw data items as model input for our fraud prediction models.

dimensions; (3) two "nonfinancial variables" and four "off-balance-sheet variables," which can be calculated using annual report disclosures; and (4) eight "market-related incentives" variables, which can be computed using either annual report disclosures, or stock price data, or both. We include all the variables under the category of "accruals quality-related variables" except for the four discretionary accrual measures (i.e., "*modified jones discretionary accruals*," "*performance-matched discretionary accruals*," "*mean-adjusted absolute value of DD residuals*," and "*studentized DD residuals*") because we wish to use variables that can be easily calculated from financial statements.[11] We include all five variables under the category of "performance variables" except for "*deferred tax expense*" because the variable "*income taxes, deferred*" required to calculate this ratio has more than 25% of its values missing in our sample period (notably, Dechow et al. did not include this variable in their subsequent regression analyses; see their tables 7 and 9). We exclude all nonfinancial variables and off-balance-sheet variables. We keep only "*actual issuance*" and "*book-to-market*" under the category "market-related incentives" because the raw financial data for both variables are readily available in Compustat. Accordingly, our replication of the Dechow et al. fraud prediction model (referred to as the basic Dechow et al. model) contains 11 financial ratios from Dechow et al. [2011, table 3].

The last column of table 2 displays the 23 raw financial data items required to compute the 11 financial ratios derived from Dechow et al. [2011]. Although there is a significant overlap in the raw financial data items between our replication of Cecchini et al. [2010] in column 2 and

[11] It is important to note that Dechow et al. [2011] did not include these four discretionary accrual measures in their subsequent regression models either.

Dechow et al. [2011] in column 3, there are a few key differences.[12] First, the Dechow et al. raw variable list in column 3 contains four raw financial data items (i.e., "*current assets, total*"; "*accounts payable, trade*"; "*long-term debt issuance*"; and "*sale of common and preferred stock*") that are excluded from the Cecchini et al. raw data list in column 2. These four raw financial data items are used to construct four financial ratios ("*change in cash margin*," "*actual issuance*," "*RSST accruals*," and "*WC accruals*") used in the the basic Dechow et al. model. Second, the Dechow et al. raw variable list in column 3 excludes five raw financial data items included in the Cecchini et al. list in column 2: "*retained earnings*"; "*depreciation and amortization*"; "*interest and related expense, total*"; "*income taxes, total*"; and "*net income (loss).*" These raw financial data items are used to construct three financial ratios: "*depreciation index*" in Beneish [1999], and "*retained earnings over total assets*" and "*EBIT*" in Summers and Sweeney [1998]. Finally, Cecchini et al. include "*net income (loss)*" as a normalizing factor because Cecchini et al. use ratios and year-over-year changes in ratios.

Because of the differences in the two lists of raw financial data items, we combine the raw data items from both lists to obtain a final sample of 28 raw financial data items for all of our subsequent fraud prediction models. In particular, our Dechow et al. model contains 14 financial ratios (the 11 ratios from the basic Dechow et al. model plus the three extra ratios from Cecchini et al., noted above).[13] Our Cecchini et al. model constructs the financial kernel using 28 raw financial data items.[14]

4. Performance Evaluation

A common approach to evaluating the out-of-sample performance of a classification model is to perform an *n*-fold cross validation (Efron and Tibshirani [1994], Witten and Frank [2005], Hastie, Tibshirani, and Friedman [2009]). Because our fraud data are intertemporal in nature, performing the standard *n*-fold cross validation is inappropriate. Therefore, as noted in section 3, we use the last few years of our sample period (i.e., 2003–08) as the test period and all the earlier years as the training period.

4.1 OUT-OF-SAMPLE PERFORMANCE EVALUATION METRIC 1: AUC

Because our fraud prediction task can be cast as a binary classification problem (fraud versus nonfraud), we can measure fraud prediction performance using the evaluation metrics for classification problems.

---

[12] Dechow et al. [2011] use "*Property, Plant, and Equipment, Net*" whereas Cecchini et al. [2010] use "*Property, Plant, and Equipment, Gross.*" We do not believe this is a significant difference and therefore treat these two items as equivalent.

[13] The performance of the basic Dechow et al. model based on the 11 financial ratios is slightly inferior to the performance of the Dechow et al. model based on the 14 financial ratios (see table A1 of the online appendix).

[14] As the Cecchini et al. model requires lagged values of the 28 raw data items, it actually uses more information than the Dechow et al. model and our proposed models.

One standard classification performance metric is accuracy, defined as $\frac{TP+TN}{TP+FN+FP+TN}$, where TP (true positive) is the number of fraudulent firm-years that are correctly classified as fraud; FN (false negative) is the number of fraudulent firm-years that are misclassified as nonfraud; TN (true negative) is the number of nonfraudulent firm-years that are correctly classified as nonfraud; and FP (false positive) is the number of nonfraudulent firm-years that are misclassified as fraud. Unfortunately, this standard classification performance metric is not appropriate in our scenario due to the imbalanced nature of our fraud versus nonfraud data (recall that the fraud percentage in our sample is less than 1% each year). For example, the naïve strategy of classifying all firm-years as nonfraud in our sample would lead to accuracy of better than 99% based on the standard classification performance metric. However, such seemingly high-performance fraud prediction models are of little value in our scenario because our chief aim is to accurately detect as many fraudulent firm-years as possible without misclassifying too many nonfraudulent firm-years. That is, we care about both the true negative rate (i.e., *specificity,* as defined below) and the true positive rate (i.e., *sensitivity,* as defined below).

To gauge the performance of a fraud prediction model properly, one could use balanced accuracy (BAC) as an alternative performance evaluation metric (He and Ma [2013]). BAC is defined as the average of the fraud prediction accuracy within fraudulent observations, and the nonfraud prediction accuracy within nonfraudulent observations. Specifically, $\text{BAC} = \frac{1}{2} \times (\textit{Sensitivity} + \textit{Specificity})$, where $\textit{Sensitivity} = \frac{TP}{TP+FN}$ and $\textit{Specificity} = \frac{TN}{TN+FP}$. Larcker and Zakolyukina [2012] note two important limitations of BAC as a performance evaluation metric. First, BAC is constructed based on a specific predicted fraud probability threshold of a given classifier, and the threshold is usually automatically determined by the classifier to maximize the BAC. In other words, by setting a different classifier threshold, one would obtain a different BAC value. In the absence of any knowledge of the costs of misclassifying false positives versus the costs of misclassifying false negatives, one could not determine the optimal predicted fraud probability threshold for the purposes of classifying fraud and nonfraud. Second, measures such as *Sensitivity* are very sensitive to the relative frequency of positive and negative instances in the sample (i.e., data imbalance).

To avoid these two limitations, we follow Larcker and Zakolyukina [2012] by using the AUC as our out-of-sample performance evaluation metric. An ROC curve is a two-dimensional depiction of a classifier's performance that combines the true positive rate (i.e., *sensitivity*) and the false positive rate (i.e., $1 - \textit{specificity}$) in one graph (Fawcett [2006]). The BAC defined above represents only one point in the ROC curve.

It is possible to reduce the performance of a fraud prediction model to a single scalar by computing the AUC. Because the AUC is a portion of the area of the unit square, its value will always lie between 0 and 1.0.

Because a random guess produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC of less than 0.5. As discussed in Fawcett [2006], the AUC is equivalent to the probability that a randomly chosen positive instance (i.e., a true fraud) will be ranked higher by a classifier than a randomly chosen negative instance (i.e., a nonfraud).

4.2 OUT-OF-SAMPLE PERFORMANCE EVALUATION METRIC 2: NDCG@K

The fraud prediction task can also be thought of as a ranking problem. Specifically, we can limit the out-of-sample performance evaluation to only a small number (i.e., k as defined below) of firm-years with the highest predicted probability of fraud. In this scenario, the performance of a fraud prediction model can be measured by the following performance evaluation metric for ranking problems: NDCG@k. NDCG@k is a widely used metric for evaluating ranking algorithms such as web search engine algorithms and recommendation algorithms (Järvelin and Kekäläinen [2002]), and has been theoretically proven to be effective (Wang et al. [2013]).

Formally, the Discounted Cumulative Gain at the position k (DCG@k) is defined as follows: $\text{DCG@k} = \sum_{i=1}^{k} (2^{rel_i} - 1)/\log_2 (i+1)$, where rel_i equals 1 if the ith observation in the ranking list is a true fraud, and 0 otherwise. The value k represents the k number of firm-years in a test period that have the highest predicted probability of fraud (referred to as the "ranking list"). In our subsequent empirical analyses, we select k so that the number of firm-years in the ranking list represents 1% of all the firm-years in a test year. We select a cutoff of 1% because the average frequency of accounting fraud punished by the SEC's AAERs is typically less than 1% of all firms in a year.

DCG@k rests on two key assumptions: (1) a true fraud observation is scored higher (i.e., $(2^{rel_i} - 1) = 1$) than a nonfraud observation (i.e., $(2^{rel_i} - 1) = 0$); and (2) a true fraud observation is scored higher if it is ranked higher in the ranking list. That is, a higher ranked observation (i.e., with a smaller i) will be weighted more highly by the position discount, that is, the denominator $\log_2 (i+1)$.

NDCG@k is DCG@k normalized by the ideal DCG@k; that is, $\text{NDCG@k} = \frac{DCG@k}{ideal\ DCG@k}$, where the ideal DCG@k is the DCG@k value when all the true instances of fraud are ranked at the top of the ranking list. Hence, the values of NDCG@k are bounded between 0 and 1.0, and a higher value represents a model's better ranking performance.

To illustrate the benefits of our second performance metric, NDCG@k, more intuitively, we also report the following two performance metrics using only the top 1% of firms with the highest predicted fraud probabilities in a test year: (i) $Sensitivity = \frac{TP}{TP+FN}$, where TP is the number of fraudulent firms contained in the top 1% of firms with the highest predicted fraud probabilities in a test year, and FN is the number of fraudulent firms that are misclassified as nonfraudulent firms in the bottom 99% of the

observations in a test year. The sum of TP and FN is the total number of fraudulent firms in a test year. (ii) $Precision = \frac{TP}{TP+FP}$, where TP is defined above and FP is the number of nonfraudulent firms that are misclassified as fraudulent firms in the top 1% of firms with the highest predicted fraud probabilities in a test year. The sum of TP and FP is the total number of firms in the top 1% of firms with the highest predicted fraud probabilities in a given test year.

Relative to the first performance evaluation metric, AUC, NDCG@k avoids the significant direct and indirect costs of investigating a large number of cases of predicted fraud, the majority of which are likely to be false positives due to the severe data imbalance. Because the average frequency of accounting fraud punished by the SEC's AAERs is less than 1%, even the best performing fraud prediction model (e.g., Cecchini et al. [2010]) would identify a large number of false positives. For example, table 7 of Cecchini et al. [2010] reports that their SVM with a financial kernel correctly classifies 80% of the fraud observations and 90.6% of the nonfraudulent observations in the out-of-sample test period, the best among the competing models considered in their study. However, applying the Cecchini et al. model to our test period 2003–08 would result in too many false positives. Specifically, fraud occurred in only 237 of the 30,883 firm-years analyzed by the Cecchini et al. model during the test period 2003–2008. The Cecchini et al. method, however, would mislabel 2,881 $((1 - 90.6\%) \times (30{,}883 - 237))$ nonfraudulent observations as fraud—a serious overestimate of the number of actual cases of fraud in the test period. Clearly, it is impractical to investigate all predicted instances of fraud. Even if one wished to investigate all the predicted fraud observations, the direct and indirect costs of doing so would be prohibitively high. At the same time, the benefits would be limited, because the majority of predicted fraudulent observations are false positives. NDCG@k avoids this problem by limiting the investigation to no more than a given number k of firm-years with the highest predicted fraud probability in the test period.

5. *The Out-of-Sample Performance of the Benchmark Models*

Before we discuss the results of our benchmark models, we wish to remind readers that unless explicitly stated otherwise, all of our performance evaluations are conducted after correcting for the potential bias resulting from serial fraud cases that span both the training and test periods (see section 3.3). With the exception of Perols [2011] and Perols et al. [2017], no existing fraud prediction studies consider the serial fraud issue. For this reason, the performance evaluation results of our benchmark models cannot be directly compared with those of existing published papers.

5.1 THE DECHOW ET AL. MODEL

Table 3 reports the out-of-sample performance evaluation results for the Dechow et al. model for the test period 2003–2008, using the evaluation

TABLE 3
The Out-of-Sample Performance Evaluation Metrics for the Test Period 2003–08

			Performance Metrics Averaged over the Test Period 2003–2008			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
14 financial ratios	1)	logit	0.672	0.028	3.99%	2.63%
			(0.167)	(0.479)		
28 raw financial data items	2)	SVM-FK	0.626	0.020	2.53%	1.92%
			(0.012)	(0.171)		
	3)	Logit	0.690	0.006	0.73%	0.85%
			(0.211)	(0.041)		
	4)	RUSBoost	0.725	0.049	4.88%	4.48%

This table shows fraud prediction models' performance comparison using the following performance metrics averaged over the test period 2003–08:

(1) *Area under the receiver operating characteristics (ROC) curve (AUC).* AUC is the area under the receiver operating characteristics (ROC) curve that combines the true positive rate (i.e., *Sensitivity*) and the false positive rate (i.e., 1 − *Specificity*) in one graph. The ROC curve is the standard technique for visualizing and selecting classifiers.

(2) *Normalized Discounted Cumulative Gain at the Position k (NDCG@k), where k is the number of top 1% firms in a test year.* The DCG@k is defined as follows: $DCG@k = \sum_{i=1}^{k} (2^{rel_i} - 1)/\log_2 (i+1)$. NDCG@k is the DCG@k normalized by the ideal DCG@k, that is, NDCG@k $= \frac{DCG@k}{ideal\ DCG@k}$, where the ideal DCG@k is the DCG@k value when all the true frauds are ranked at the top of the ranking list.

(3) *The values of Sensitivity by classifying the top 1% firms with the highest predicted fraud probabilities in a test year as frauds.* Specifically, *Sensitivity* $= \frac{TP}{TP+FN}$, where TP is the number of true frauds contained in the top 1% predicted frauds in a test year and FN is the number of true frauds that are misclassified as nonfrauds in the bottom 99% of the observations in a test year. The sum of TP and FN is the number of total true frauds in a test year.

(4) *The values of Precision by classifying the top 1% firms with the highest predicted fraud probabilities in a test year as frauds.* Specifically, *Precision* $= \frac{TP}{TP+FP}$, where TP is the number of true frauds contained in the top 1% predicted frauds in a test year and FP is the number of false frauds that are misclassified as frauds in the top 1% predicted frauds in a test year. The sum of TP and FP is the number of top 1% firms in a test year.

(5) The *p*-values in the parentheses are based on a two-tailed *t*-test for our RUSBoost model vs. another model.

metrics of AUC and NDCG@k. Table 3 also reports the pairwise *t*-tests of our ensemble learning model versus other models in parentheses. However, because we have only six observations for such *t*-tests (low power), we do not rely on them in the following inferences and instead we focus on the magnitude of the differences in the two performance evaluation metrics between the different fraud detection models.

Consistent with Dechow et al. [2011], the average AUC for the three test years is 0.672, much higher than 0.50, the AUC cutoff for random guesses. The average NDCG@k is 0.028. The average value of *sensitivity* (defined in table 3) is 3.99%, meaning that 3.99% of all true cases of fraud in the population are captured in the top 1% of observations with the highest predicted fraud probabilities. Similarly, the average value of *precision* (defined in table 3) is 2.63%, meaning that only 2.63% of the top 1% of observations with the highest predicted fraud probabilities are actual cases of fraud, much higher than the unconditional mean fraud frequency of 0.73% for the full test sample of 2003–2008.

5.2 THE CECCHINI ET AL. MODEL

Cecchini et al. [2010] developed an innovative SVM method (hereafter referred to as support vector machines with a financial kernel [SVM-FK]) based on a financial kernel that maps raw financial data into a list of predefined ratios. The Cecchini et al. list is broader than the ratio list typically used in the accounting fraud prediction literature.[15] Cecchini et al. [2010, table 7] show that their SVM-FK significantly outperforms several representative fraud prediction models in accounting, including that of Dechow et al. [2011].

In this section, we replicate the Cecchini et al. SVM-FK method using our sample data. Our replication improves upon Cecchini et al. [2010] by avoiding two look-ahead biases. First, to address the class imbalance issue, SVM-FK employs the cost-sensitive SVM by adjusting the model parameter $C^{+1} : C^{-1}$ (i.e., the ratio of the cost of misclassifying fraud and nonfraud). When searching for the optimal parameter $C^{+1} : C^{-1}$ to maximize the value of AUC, Cecchini et al. [2010] directly perform the search using the test sample rather than a holdout validation sample. For this reason, the Cecchini et al. implementation procedures are subject to a look-ahead bias (which they acknowledge; Cecchini et al. [2010, on page 1156]). We avoid this limitation by performing a grid search in a holdout validation sample. Specifically, we train the SVM-FK model using 1991–1999 and validate the model using 2000–2001 for the test years 2003–2008. We use two years instead of one year for validation because of the low frequency of fraud in a typical year.[16] After determining the optimal parameter $C^{+1} : C^{-1}$ (20 in our sample), we train the model and test its performance accordingly. Specifically, we use the training period 1991–2001 for test year 2003, 1991–2002 for test year 2004 and so on.

Second, we differ from Cecchini et al. [2010] in that we use all firm-years in a test period to perform the out-of-sample performance evaluation. Cecchini et al. [2010] perform the model training, model validation, and out-of-sample model evaluation only after obtaining a set of fraud firm-years and all matched nonfraud firm-years within the same industry year. Because SVM-FK models are extremely time-consuming to train and validate for large data sets, it is appropriate to use a smaller matched sample

---

[15] Specifically, the SVM-FK method maps a firm's raw financial data in two consecutive years (referred to as year 1 and year 2, respectively) into the following six types of predefined ratios, representing both intrayear ratios and year-over-year changes in ratios: $\phi(u) = \left(\frac{u_{1,i}}{u_{1,j}}, \frac{u_{1,j}}{u_{1,i}}, \frac{u_{2,i}}{u_{2,j}}, \frac{u_{2,j}}{u_{2,i}}, \frac{u_{1,i}u_{2,j}}{u_{1,j}u_{2,i}}, \frac{u_{1,j}u_{2,i}}{u_{1,i}u_{2,j}}\right)$, $i, j = 1, \ldots, n$, $i < j$, where $u_{1,i}$ is raw financial data item i in year 1 and n represents the total number of raw financial data items. Because we have a total of 28 raw financial data items (i.e., $n = 28$), the above transformation function would result in a total of 2,268 ratios. Following Cecchini et al. [2010], we also add year as a control variable. We thank Mark Cecchini for sharing with us the final data set used in Cecchini et al. [2010].

[16] The logistic regression method doesn't require a similar grid search because there are no parameters to tune.

of fraud and nonfraud during the training period. However, it is problematic to use only the matched fraudulent and nonfraudulent firm-years in a test year to evaluate the out-of-sample performance of the SVM-FK model, as doing so may invite look-ahead bias, making real-time implementation problematic. Specifically, because it takes an average of two years for the initial disclosure of accounting fraud (Dyck, Morse, and Zingales [2010]), a relevant decision maker (e.g., a regulator or investor) does not know at the time of prediction whether a company's financial statements in an industry year are fraudulent or not in a test year. Therefore, the decision maker cannot match a fraudulent firm with a nonfraudulent firm in the test year. Hence, a more appropriate out-of-sample performance evaluation approach is to evaluate the SVM-FK out-of-sample performance using the *entire population* of firm-years in the test period. For this reason, our replication of Cecchini et al. [2010] uses a matched sample of fraudulent and nonfraudulent firm-years for training and validation, but uses the entire population of firm-years in the test period 2003–08 when assessing the SVM-FK model's out-of-sample performance.

These distinctions appear to be critical in assessing the out-of-sample performance evaluation of the SVM-FK model. Specifically, untabulated results show that the sample of matched fraudulent and nonfraudulent firms constitutes only 22.61% (6,984/30,883) of the population of fraud and nonfraud firms in the test period 2003–2008. Two hundred thirty-seven (0.77%) of the 30,883 observations in the full sample of test period 2003–2008 are true instances of fraud. In contrast, for the matched sample of fraud and nonfraud based on industry and year in the test period 2003–2008, 237 (3.39%) of the 6,984 observations are actual cases of fraud. Table A1 of the online appendix shows that the average AUC for the SVM-FK model is 0.673 using a matched sample of fraudulent and nonfraudulent observations in the test period 2003–2008. Notably, this number drops significantly when the full population of firm-years in 2003–2008 is used. Specifically, as shown in table 3, the average AUC, after correcting for the two look-ahead biases, is only 0.626, even lower than the average AUC for the Dechow et al. model. Using the NDCG@k as an alternative evaluation criterion, we find that the average value of NDCG@k for our replication of the Cecchini et al. SVM-FK method is only 0.020, also lower than the average NDCG@k from the Dechow et al. model. For the top 1% of fraudulent firm-years predicted by the SVM-FK model in the test period 2003–08, the average values of *sensitivity* and *precision* are 2.53% and 1.92%, respectively. Using either AUC or NDCG@k as the performance evaluation metric, we conclude that, overall, the performance of the Cecchini et al. model is weaker than the performance of the Dechow et al. model.

6. *The Out-of-Sample Performance of Our Proposed Model*

We next examine whether it is possible to improve the performance of fraud prediction by using raw financial data coupled with a more

powerful machine learning method, ensemble learning. To see the value of combining raw data with ensemble learning more clearly, we break down this evaluation into three steps. First, we examine prediction performance by continuing to use the logistic regression but changing the model input from the 14 financial ratios to the 28 raw financial data items. Second, we examine the fraud prediction performance of our proposed model by using both the 28 raw data items and the ensemble learning method. Third, we examine the fraud prediction performance by using the ensemble learning method but changing the model input from the 28 raw data items to the 14 financial ratios or the combination of the 14 financial ratios and the 28 raw data items.

6.1 PREDICTING FRAUD USING RAW FINANCIAL DATA AND LOGIT

Table 3 shows the prediction performance of the logistic regression model based on the 28 raw financial data items. To minimize the effect of scale differences across different raw data items, we normalize each firm-year observation's input vector (i.e., the list of raw data items) such that the normalized vector has a unit length; that is, $x' = \frac{x}{||x||}$, where the divisions are element-wise. For example, a vector of $(1,\ 2)$ is normalized to $(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$.[17] The average AUC for the logistic regression model based on the 28 raw financial data items is 0.690, higher than the average AUC for both the Dechow et al. model and the Cecchini et al. model. However, the mean value of NDCG@k for the logistic regression based on the 28 raw data items is only 0.006, lower than the mean values of NDCG@k for the Dechow et al. model and the Cecchini et al. model. Overall, the results are mixed on whether using raw data alone without adopting a more advanced machine learning method can improve prediction performance.

6.2 PREDICTING FRAUD USING RAW DATA AND ENSEMBLE LEARNING

Limiting ourselves to the same 28 raw financial data items, we next examine whether it is possible to improve out-of-sample fraud prediction performance by using the more advanced data mining method, ensemble learning.[18] Although every algorithm has some parameters that require fine-tuning, it is worth noting that AdaBoost (and variants such as RUSBoost) is considered to be one of the best out-of-the-box classification algorithms because it performs well without fine tuning of parameters. RUSBoost has two main parameters to tune: the number of decision trees and the complexity of the trees. We tune these parameters by performing

[17] Although uncommon in the accounting literature, this scaling method is widely used in the machine learning literature (Han, Kamber, and Pei [2006]). See also http://en.wikipedia.org/wiki/Feature_scaling for an intuitive introduction to the scaling of features (i.e., raw data items in our case).

[18] Reported in table A1 of the online appendix, we also tried the SVM method based on the 28 raw financial data items and found no evidence that it outperforms our ensemble learning method discussed below.

a holdout validation in the same way as in SVM-FK. Specifically, we train the RUSBoost model using 1991–1999 and validate the model using 2000–2001. We set the number of trees to 3,000 because we find that the performance of RUSBoost models tends to become stable after 3,000 "grown" trees. Theoretically, there are several parameters that can be used to control the complexity of trees, such as the depth of trees or the minimum required number of samples at a tree leaf (i.e., "minleaf"). We choose to tune the parameter "minleaf" rather than "depth of trees" because the practical implementation of decision tree in Matlab does not support the parameter "depth of trees". We set the parameter "minleaf" to 5, which is the best value in the range (1, 200) for maximizing both AUC and NDCG@k performance metrics in the validation period 2000–2001.

In addition to the aforementioned two important parameters, we also consider two additional parameters: "learning rate" and "RatioToSmallest." "Learning rate" shrinks the contribution of each base model (i.e., decision tree in our case). The value of "learning rate" is between 0 and 1, and we set it to 0.1 because this choice generally enables the ensemble model to converge to a better solution. The other parameter, "RatioToSmallest" is a specific parameter of RUSBoost that specifies the ratio between the number of nonfraud cases and fraud cases when performing RUS. "RatioToSmallest" is usually set at 1:1. Therefore, we set the parameter "RatioToSmallest" at 1:1.

Table 3 reports the out-of-sample performance of the ensemble learning model based on the raw financial data over the test period 2003–2008. The average AUC for the ensemble learning method is 0.725, much larger than the average AUC for the better benchmark model, the Dechow et al. model (0.672). Using NDCG@k as an alternative evaluation criterion, we find that the average value of NDCG@k for the ensemble learning method is 0.049, larger than the average value of NDCG@k for the Dechow et al. model (0.028). For the top 1% of predicted fraudulent firms in the test period 2003–2008, the average values of *sensitivity* and *precision* for the ensemble learning model are 4.88% and 4.48%, respectively. In contrast, the corresponding values for the Dechow et al. model are 3.99% and 2.63%, respectively. Overall, these results suggest that our proposed model, which combines raw financial data with the more powerful ensemble learning method, outperforms the two benchmark models.

To better appreciate the economic significance in the differential performance of the various prediction models for the test period 2003–2008, we also calculate the number of true fraud observations identified using the NDCG@k approach, where $k = 1\%$ (see table A2 of the online appendix). We find that our best model, the ensemble learning model, identified a total of 16 fraud cases in the test period 2003–2008. In contrast, the comparable figure is 9 for the Dechow et al. model and 7 for the Cecchini et al. model. These results suggest that the differences in the performance of the ensemble learning model versus the two benchmark models are also economically significant.

TABLE 4

The Out-of-Sample Performance Evaluation Metrics: The Ensemble Learning Method Based on Different Sets of Input Variables

			Performance Metrics Averaged over the Test Period 2003–2008			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
28 raw financial data items	1)	RUSBoost (from table 3)	0.725	0.049	4.88%	4.48%
14 financial ratios	2)	RUSBoost	0.659	0.017	2.03%	1.69%
14 financial ratios + 28 raw financial data items	3)	RUSBoost	0.696	0.035	3.19%	2.54%
All 294 raw financial data items	4)	RUSBoost	0.692	0.015	1.92%	1.41%

This table shows the performance of the ensemble learning method based on different sets of input variables. As a benchmark, we also tabulate the results for the ensemble learning method based on the 28 raw data items alone from table 3. All performance metrics are averaged over the test period 2003–2008. See table 3 for the definitions of performance metrics.

6.3 RATIOS VERSUS RAW DATA USING THE ENSEMBLE LEARNING METHOD

In this section, we further examine whether we can improve the performance of the ensemble learning method by using the 14 ratios alone, or by using both the 28 raw data items and the 14 ratios together. Table 4 reports the out-of-sample performance statistics for these two alternative models. We find no evidence that the two alternative ensemble learning models outperform the ensemble learning model based on the 28 raw data items alone. This evidence is consistent with our conjecture that once we have considered the 28 raw data items coupled with a flexible and powerful machine learning method, the ratios constructed from the same raw data items are no longer incrementally useful in predicting fraud.

7. *Supplemental Analyses*

7.1 ALTERNATIVE TEST PERIODS

As noted in section 3.2, the observed frequency of fraud declines almost monotonically over 2003–2014 and one suspected reason for this decline is the presence of undetected fraud, especially in the postfinancial crisis period. Hence, we limit our test period to 2003–2008 in table 3. To show the robustness of main results, we replicate the results in table 3 using the following alternative test samples: 2003–2005, 2003–2011, and 2003–2014. Because of the stated reasons in section 3.2, it is reasonable to assume that the problem of undetected fraud grows over time. Hence, the performance evaluation results using longer test periods should be less reliable.

The results are reported in table 5. Panels A–C report the results for test years 2003–2005, 2003–2011, and 2003–2014, respectively. There are several key findings. First, for all the test periods, the ensemble learning model continues to perform the best using either AUC or NDCG@k, suggesting the robustness of the ensemble learning model. Second, the performance of the ensemble learning model declines almost monotonically from 2003–2005 to 2003–2008, 2003–2011, and 2003–2014. The performance of the ensemble learning model based on raw data is the most impressive for the earliest test period 2003–2005: The average values of AUC and NDCG@k are 0.753 and 0.085, respectively, in contrast to the values of 0.725 and 0.049 for the test period 2003–2008 in table 3. These results suggest that the presence of undetected fraud observations may have significantly reduced the performance of the ensemble learning model.[19] Third, there is weak evidence that the performance of the Dechow et al. model based on AUC (but not NDCG@k) increases from the test period 2003–2005 to 2003–2014, which is counterintuitive, but the performance of the Cecchini et al. model shows no clear pattern over time.

7.2 DO MORE RAW DATA HELP?

The empirical analyses so far only use 28 raw financial data items derived from accounting ratios that have been identified by human experts to be important in explaining accounting fraud. These 28 raw data items represent only a small fraction of the available data items from the three key corporate financial statements. Hence, we next ask whether it is possible to improve the performance of the ensemble learning method further by including more raw data items, without any guidance from theory, from the three financial statements. Although it is beyond the scope of this study to perform a systematic analysis of this research question, we take a "brute force" approach by including in the ensemble model all the available raw financial statement data items from the Compustat Fundamental file that satisfy relevant sample selection conditions. We identified a total of 266 additional raw data items from the three financial statements that are applicable to COMPUSTAT fundamental annual industrial format companies during the sample period 1991–2008. As the ensemble learning method can handle observations with missing values, we do not impose any missing value restrictions for the enlarged raw data set.

---

[19] We attempt to provide a rough estimate of the rate of undetected fraud for the postcrisis period 2009–2014 using our ensemble learning model in Table 3. To do so, we make three assumptions: (1) the reported fraud frequency for the test period 2003–2008 represents the true fraud rate; (2) the true fraud rate is the same for the periods 2003–2008 and 2009–2014, which may not be true (e.g., Wang, Winton, and Yu [2010]); and (3) if all true fraud cases are observed, the performance of the ensemble learning model is identical for the test periods 2003–2008 vs. 2009–2014. Because the ensemble learning model's average *reported* precision using the NDCG@k approach is 4.48% (see table 3) for the test period 2003–2008 and 0.93% (untabulated) for the test period 2009–2014, we infer that the estimated rate of undetected fraud for the period 2009–2014 would be 3.55% (i.e., 4.48%–0.93%).

TABLE 5

The Out-of-Sample Performance Evaluation Metrics for the Test Years 2003–2005, 2003–2011, and 2003–2014

Panel A: The out-of-sample performance evaluation metrics for the test period 2003–2005

			Performance Metrics Averaged over the Test Period 2003–2005			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
14 financial ratios	1)	Logit	0.649	0.012	1.37%	1.29%
			(0.041)	(0.153)		
28 raw financial data items	2)	SVM-FK	0.637	0.024	2.28%	2.53%
			(0.029)	(0.035)		
	3)	Logit	0.685	0.012	1.45%	1.69%
			(0.091)	(0.054)		
	4)	RUSBoost	0.753	0.085	7.64%	7.83%

Panel B: The out-of-sample performance evaluation metrics for the test period 2003–2011

			Performance Metrics Averaged over the Test Period 2003–2011			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
14 financial ratios	1)	Logit	0.672	0.024	3.49%	2.23%
			(0.143)	(0.395)		
28 raw financial data items	2)	SVM-FK	0.647	0.025	3.07%	1.98%
			(0.045)	(0.364)		
	3)	Logit	0.702	0.012	1.87%	1.19%
			(0.732)	(0.074)		
	4)	RUSBoost	0.710	0.040	4.40%	3.60%

Panel C: The out-of-sample performance evaluation metrics for the test period 2003–2014

			Performance Metrics Averaged over the Test Period 2003–2014			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
14 financial ratios	1)	Logit	0.702	0.023	3.45%	1.86%
			(0.509)	(0.616)		
28 raw financial data items	2)	SVM-FK	0.628	0.019	2.30%	1.48%
			(0.006)	(0.356)		
	3)	Logit	0.709	0.011	1.84%	1.04%
			(0.649)	(0.125)		
	4)	RUSBoost	0.717	0.030	3.30%	2.70%

Panels A–C show fraud prediction models' performance comparison using the following performance metrics averaged over the test years 2003–2005, 2003–2011, and 2003–2014, respectively. See table 3 for the definitions of performance metrics. The *p*-values in the parentheses are based on two-tailed *t*-tests for our RUSBoost model vs. other models.

TABLE 6

The Out-of-Sample Performance Evaluation Metrics for the Test Period 2003–2008: Ignore the Serial Fraud Problem

			Performance Metrics Averaged over the Test Period 2003–2008			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
14 financial ratios	1)	Logit	0.674	0.029	3.99%	2.63%
			(0.011)	(0.035)		
28 raw financial data items	2)	SVM-FK	0.661	0.025	2.90%	2.24%
			(0.001)	(0.008)		
	3)	Logit	0.708	0.002	0.24%	0.28%
			(0.005)	(0.006)		
	4)	RUSBoost	0.801	0.158	13.56%	10.74%

This table shows fraud prediction models' performance comparison based on a sample that ignores the serial fraud problem. All performance metrics are averaged over the test period 2003–2008. See table 3 for the definitions of performance metrics. The *p*-values in the parentheses are based on two-tailed *t*-tests paired with our RUSBoost model.

The last row of table 4 reports the results. The average value of AUC is lower for this more comprehensive ensemble learning model than for the ensemble method based on the 28 raw data items from table 3 (0.692 vs. 0.725). The average value of NDCG@k is also much lower for this more comprehensive ensemble learning model than for the ensemble learning method based on the 28 raw data items (0.015 vs. 0.049). These results suggest that including a large number of raw financial data items in our best fraud prediction model without any clear theoretical guidance does not help improve the model's performance. This finding suggests that theoretical guidance is still important in model input selection, even when one adopts powerful machine learning methods such as ensemble learning.

7.3 SERIAL FRAUD

The results reported so far are based on a sample that is free of the problem of serial fraud. Because most prior fraud prediction studies do not show the impact of serial fraud on fraud prediction performance, in this section, we also replicate all of our fraud prediction models without considering the serial fraud problem for the same test period 2003–2008. The results are reported in table 6. Consistent with our conjecture, the performance of the ensemble learning model in table 6 improves most significantly relative to the performance of the same model in table 3. This finding suggests the importance of dealing with serial fraud in fraud prediction models. However, the model performance rankings shown in table 6 are qualitatively similar to those in table 3. Hence, our overall inferences are unaffected by the serial fraud problem.

7.4 UNCOVERING THE BLACK BOX BEHIND THE PERFORMANCE OF THE ENSEMBLE LEARNING METHOD

A well-known disadvantage of many machine learning methods (e.g., Neural Network) is the lack of transparency with regard to the inner working of such models. Whereas individual decision trees can be interpreted easily by simply visualizing the tree structure, ensemble learning methods comprise hundreds of trees and thus cannot be easily interpreted by visual inspection of the individual trees. Fortunately, some techniques have been proposed to help shed light on the significant performance drivers of ensemble models by estimating the importance of various features (in the context of this study, raw data items) in fraud prediction (Tuv et al. [2009]). In this paper, we use the "predictorImportance" function implemented in MATLAB to estimate the importance of the 28 raw financial data items used in our ensemble learning model.[20] Specifically, individual decision trees intrinsically perform feature selection by selecting appropriate split points. This information can be used to measure the importance of each feature: The change of impurity (a measure of the quality of the split) due to the split on a feature indicates its importance (Breiman et al. [1984]). This notion of importance can be extended to decision tree ensembles by simply averaging the feature importance of each tree.

Panel A of table 7 reports the descriptive statistics on the average importance of 28 raw data items for our ensemble learning model for the six test years. The 28 raw data items are sorted from high to low based on the average feature importance. The top 10 most important features (or raw data items) that contribute to the ensemble learning model's superior performance are listed below, starting with the most important: (1) c*ommon shares outstanding*; (2) *current assets, total*; (3) *sale of common and preferred stock*; (4) *property, plant, and equipment, total*; (5) *account payable, trade*; (6) *cash and short-term investments*; (7) *price close, annual, fiscal*; (8) *retained earnings*; (9) *inventories, total*; (10) *common/ordinary equity, total.* It is interesting to note that "*common shares outstanding*" and "*price close, annual, fiscal*" provide significant information for fraud prediction. We suspect that "*common shares outstanding*" is useful in predicting accounting fraud simply because misstating firms often issue common equity (Dechow, Sloan, and Sweeney [1995], Dechow et al. [2011]). Similarly, "*price close, annual, fiscal*" provides valuable information about the likelihood of fraud because of informed trading during the misstatement period by both company employees and other related stakeholders (Summers and Sweeney [1998], Kedia and Philippon [2007], Agrawal and Cooper [2015]).

To benchmark the performance of the ensemble learning model, panel B of table 7 also lists the specific financial statement accounts affected by

[20] For a more detailed description of the function "predictorImportance", please refer to http://www.mathworks.com/help/stats/compactclassificationensemble.predictorimportance.html.

TABLE 7

The Importance of the 28 Raw Financial Data Items Used in the RUSBoost Model in Table 3

Panel A: The importance of the 28 raw data items in the RUSBoost model

Rank	28 Raw Data Items	Average Value of a Feature's Importance for the Test Period 2003–2008	Related Account Categories from Panel B
1	*Common shares outstanding*	1.448	*inc_exp_se*
2	*Current assets, total*	1.362	*asset*
3	*Sale of common and preferred stock*	1.324	*inc_exp_se*
4	*Property, plant and equipment, total*	1.289	*asset*
5	*Account payable, trade*	1.237	*pay*
6	*Cash and short-term investments*	1.210	*asset*
7	*Price close, annual, fiscal*	1.146	*inc_exp_se*
8	*Retained earnings*	1.119	*inc_exp_se; rev; cogs*
9	*Inventories, total*	1.064	*inv*
10	*Common/ordinary equity, total*	1.054	*inc_exp_se; res*
11	*Debt in current liabilities, total*	1.042	*liab*
12	*Depreciation and amortization*	1.041	*inc_exp_se*
13	*Receivables, total*	0.982	*rec*
14	*Cost of goods sold*	0.968	*cogs*
15	*Assets, total*	0.873	*asset*
16	*Long-term debt issuance*	0.801	*liab*
17	*Income before extraordinary items*	0.788	*inc_exp_se*
18	*Long-term debt, total*	0.784	*liab*
19	*Interest and related expense, total*	0.731	*inc_exp_se*
20	*Income taxes, total*	0.712	*inc_exp_se*
21	*Current liabilities, total*	0.704	*liab*
22	*Sales/turnover (net)*	0.625	*rev*
23	*Income taxes payable*	0.592	*liab*
24	*Investment and advances, other*	0.588	*asset*
25	*Liabilities, total*	0.546	*liab*
26	*Short-term investments, total*	0.476	*mkt_sec; asset*
27	*Net income (loss)*	0.371	*inc_exp_se; rev; cogs*
28	*Preferred/preference stock (capital), total*	0.247	*inc_exp_se*

Panel B: The specific account categories affected by the detected accounting frauds over the test period 2003–2008

Rank	Account Category from Dechow et al. [2011]	Definition from Dechow et al. [2011]	Frequency	Related Top-10 Raw Data Items from Panel A
1	*inc_exp_se*	Equals 1 if misstatement affected net income, hence, shareholder equity but could not be classified in any specific income, expense or equity accounts below in this table, 0 otherwise	169	*Common shares outstanding; sale of common and preferred stock; price close, annual, fiscal; retained earnings; common/ordinary equity, total*

(Continued)

TABLE 7—*Continued*

Panel B: The specific account categories affected by the detected accounting frauds over the test period 2003–2008

Rank	Account Category from Dechow et al. [2011]	Definition from Dechow et al. [2011]	Frequency	Related Top-10 Raw Data Items from Panel A
2	*Rev*	Equals 1 if misstatement affected revenues, 0 otherwise	127	*Retained earnings*[a]
3	*asset*	Equals 1 if misstatement affected an asset account that could not be classified in a separate individual asset account in this table, 0 otherwise	63	*Current assets, total; property, plant and equipment, total; cash and short-term investments; inventories, total*
4	*rec*	Equals 1 if misstatement affected accounts receivable, 0 otherwise	46	
5	*inv*	Equals 1 if misstatement affected inventory, 0 otherwise	29	*Inventories, total*
5	*res*	Equals 1 if misstatement affected reserves accounts, 0 otherwise. Dechow et al. [2011]	29	*Common/ordinary equity, total*
7	*liab*	Equals 1 if misstatement affected liabilities, 0 otherwise	26	*Account payable, trade*
7	*cogs*	Equals 1 if misstatement affected cost of goods sold, 0 otherwise	26	*Retained earnings*
9	*pay*	Equals 1 if misstatement affected accounts payable, 0 otherwise	17	*Account payable, trade*
10	*mkt_sec*	Equals 1 if misstatement affected marketable securities, 0 otherwise	0	
10	*debt*	Equals 1 if misstatement affected allowance for bad debts, 0 otherwise	0	
Total			532	

Panel A shows the feature importance of the 28 raw data items used in the RUSBoost model in table 3. The reported values are the estimates of feature importance computed using the "predictorImportance" function in MATLAB (multiplied by 10,000). Panel B shows the specific account categories affected by the misstatements identified by the AAERs over the period 2003–2008. These data were hand-collected from the AAERs by Dechow et al. [2011]. There are 268 misstatement firm-years in 2003–2008 involving 532 specific account categories. Note a single misstatement firm-year could affect more than one specific account.

[a]Although "*rev*" and "*cogs*" are not equity accounts, they are closed to "*Retained Earnings*" every year. Hence, we still assign "*rev*" and "*cogs*" to "*Retained Earnings*".

the misstatements reported by the AAERs over the same period 2003–2008, sorted by the frequency with which an account category is affected by accounting fraud. Then, we map the top-10 raw accounting data items in panel A to these individual accounts in panel B. We directly obtained the data used in panel B, including the 11 individual account categories, from Dechow et al. [2011], who in turn collected these data from the AAERs. There is a substantial overlap between our top-10 raw data items in the last column and the Dechow et al. top-ranked individual accounts in the second column. This evidence suggests that our ensemble learning method has been relatively effective in identifying the specific accounts most frequently affected by the misstatements. It is also interesting to note that the most frequent account category affected by the misstatements per Dechow et al. [2011] is "*inc_exp_se*," which is a miscellaneous category for the affected accounts that could not be classified as a distinctive income, expense, or equity account in panel B, table 7. The corresponding raw data items for "*inc_exp_se*" are "*common shares outstanding*"; "*sale of common and preferred stock*"; "*price close, annual, fiscal*"; "*retained earnings*"; and "*common/ordinary equity, total*"; interestingly, these raw data items are some of the most significant predictors of fraud, as shown in panel A, table 7.

8. Conclusion

Accounting fraud is extremely difficult to detect. Hence, an important area of accounting research is the development of effective methods for detecting corporate accounting fraud on a timely basis, thereby limiting the extent of fraud-related damage. The objective of this study is to develop a new out-of-sample fraud prediction model based on a sample of publicly traded U.S. firms over the period 1991–2008. To preserve the intertemporal nature of fraud prediction, we use the last six years of our sample period, 2003–2008, as the out-of-sample test period, and the years prior as the training period. We also show the robustness of our results using three alternative test periods, 2003–2005, 2003–2011, and 2003–2014. To mitigate potential look-ahead bias, we also require a minimum gap of 24 months between the financial results announcement for the last training period and the financial results announcement of a test year. We do this because it takes an average of 24 months for a fraud to be disclosed (Dyck, Morse, and Zingales [2010]).

In keeping with existing research, we use only readily available financial data as input in fraud prediction. However, we depart from most existing research in accounting in several important ways. First, we predict fraud out-of-sample rather than explain fraud determinants within sample. Second, we use raw financial data from financial statements to predict fraud. In contrast, the extant research typically uses financial ratios identified by human experts to predict fraud. Third, we use ensemble learning, one of the state-of-the-art paradigms in machine learning, for fraud prediction rather than the commonly used logit regression. Finally, we introduce a novel

approach, commonly used for ranking problems, to assessing the performance of fraud prediction models, referred to as NDCG@k.

Although there are numerous raw financial data items in financial statements, we limit our empirical analyses to the 28 raw data items from Cecchini et al. [2010] and Dechow et al. [2011] in order to compare the performance of our proposed fraud prediction model with that of more traditional fraud prediction methods (whose fraud predictors are also derived from the same set of raw financial data). The 28 raw data items have been identified by financial experts based on prior fraud research.

We adopt two benchmark fraud prediction models. First, we follow Dechow et al. [2011] by using a logistic fraud prediction model based on 14 financial ratios derived from the 28 raw data items, referred to as the Dechow et al. model. Our second benchmark model is the fraud prediction model developed by Cecchini et al. based on SVM-FK that maps the 28 raw financial data items into a broader set of financial ratios and changes in financial ratios.

We find that the out-of-sample performance of both benchmark models is better than random guesses, but the performance of the Dechow et al. model is better than the performance of the Cecchini et al. model. More importantly, we find that the ensemble learning model based on the 28 raw financial data items directly outperforms the two benchmark models. However, we find no evidence that a simple logistic regression model based on raw data, or an ensemble learning model based on the 14 financial ratios, or the combination of the 14 financial ratios and the 28 raw data items, outperforms the ensemble learning model based on the 28 raw data items only. These results suggest that the Dechow et al. model based on the financial ratios identified by human experts has not fully utilized the valuable information contained within the raw financial data. In addition, we show that it is possible to extract more useful predictive information from the raw data by constructing a more advanced machine learning model designed to make use of such data.

Because the 28 raw data items represent only a small fraction of the hundreds of possible raw financial data items that emerge from the accounting system, we also examine whether including all eligible, readily available raw financial data items from the three financial statements further improves the performance of our ensemble learning model. We find no such evidence, suggesting that simply adding more raw data items, without any supplemental theory, is not sufficient. However, we do not rule out the possibility that better fraud prediction models could be developed by performing a more systematic and theory-driven selection of model input from the hundreds of readily available raw financial data items.

Our findings are relevant to a growing accounting literature that attempts to harvest the textual data found in corporate filings for the purposes of predicting fraud or firm performance (e.g., Li [2010], Larcker and Zakolyukina [2012], Lo, Ramos, and Rogo [2017]). To demonstrate the usefulness of such textual data, a commonly used benchmark is a list

of quantitative variables derived from raw financial data. One interesting question future researchers may explore is whether the usefulness of textual data continues to hold if the information from the readily available raw financial data is more efficiently extracted using advanced data mining techniques.

REFERENCES

AGRAWAL A., AND T. COOPER. "Insider Trading Before Accounting Scandals." *Journal of Corporate Finance* 34 (2015): 169–90.

ARMSTRONG, C. S.; A. D. JAGOLINZER; AND D. F. LARCKER. "Chief Executive Officer Equity Incentives and Accounting Irregularities." *Journal of Accounting Research* 48 (2010): 225–71.

ATKINS, P. S., AND B. J. BONDI. "Evaluating the Mission: A Critical Review of the History and Evolution of the SEC Enforcement Program." *Fordham Journal of Corporate and Financial Law* 13 (2008): 367–417.

BEASLEY, M. "An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud." *The Accounting Review* 71 (1996): 443–65.

BEASLEY, M. S.; J. V. CARCELLO; AND D. R. HERMANSON. "Fraudulent Financial Reporting: 1987–1997: An Analysis of U.S. Public Companies." Sponsored by the Committee of Sponsoring Organizations of the Treadway Commission (COSO), 1999.

BEASLEY, M. S.; J. V. CARCELLO; D. R. HERMANSON; AND T. L. NEAL. "Fraudulent Financial Reporting: 1998–2007: An Analysis of U.S. Public Companies." Sponsored by the Committee of Sponsoring Organizations of the Treadway Commission (COSO), 2010.

BENEISH, M. D. "Detecting GAAP Violation: Implications for Assessing Earnings Management Among Firms with Extreme Financial Performance." *Journal of Accounting and Public Policy* 16 (1997): 271–309.

BENEISH, M. D. "The Detection of Earnings Manipulation." *Financial Analysts Journal* 55 (1999): 24–36.

BRAZEL, J. F.; K. L. JONES; AND M. F. ZIMBELMAN. "Using Nonfinancial Measures to Assess Fraud Risk." *Journal of Accounting Research* 47 (2009): 1135–66.

BREIMAN, L.; J. FRIEDMAN; C. J. STONE; AND R. A. OLSHEN. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

CECCHINI, M.; H. AYTUG; G. J. KOEHLER; AND P. PATHAK. "Detecting Management Fraud in Public Companies." *Management Science* 56 (2010): 1146–60.

CERESNEY, A. Co-Director of the Division of Enforcement, "Financial Reporting and Accounting Fraud." Speech at American Law Institute Continuing Legal Education, Washington, D.C. Sept. 19, 2013.

DECHOW, P. M.; W. GE; C. R. LARSON; AND R. G. SLOAN. "Predicting Material Accounting Misstatements." *Contemporary Accounting Research* 28 (2011): 17–82.

DECHOW, P. M.; R. G. SLOAN; AND A. P. SWEENEY. "Detecting Earnings Management." *The Accounting Review* 70 (1995): 193–226.

DECHOW, P. M.; R. G. SLOAN; AND A. P. SWEENEY. "Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC." *Contemporary Accounting Research* 13 (1996): 1–36.

DEHAAN, E; S. KEDIA; K. KOH; AND S. RAJGOPAL. "The Revolving Door and the SEC's Enforcement Outcomes: Initial Evidence from Civil Litigation." *Journal of Accounting and Economics* 60 (2015): 65–96.

DYCK, A.; A. MORSE; AND L. ZINGALES. "Who Blows the Whistle on Corporate Fraud?" *Journal of Finance* LXV (2010): 2213–53.

EFENDI, J.; A. SRIVASTAVA; AND E. P. SWANSON. "Why do Corporate Managers Misstate Financial Statements? The Role of Option Compensation and Other Factors." *Journal of Financial Economics* 85 (2007): 667–708.

EFRON, B., AND R. J. TIBSHIRANI. *An Introduction to the Bootstrap*. London, UK: CRC Press, 1994.

ENTWISTLE, G., AND D. LINDSAY. "An Archival Study of the Existence, Cause, and Discovery of Income-Affecting Financial Statement Misstatements." *Contemporary Accounting Research* 11 (1994): 271–96.

ERICKSON, M.; M. HANLON; AND E. L. MAYDEW. "Is There a Link between Executive Equity Incentives and Accounting Fraud?" *Journal of Accounting Research* 44 (2006): 113–43.

ERNST & YOUNG. "Driving Ethical Growth—New Markets, New Challenges." 11th Global Fraud Survey, 2010. Available at http://www.ey.com/Publication/vwLUAssets/Driving_ethical_growth_new_markets,_new_challenges:_11th_Global_Fraud_Survey/$FILE/EY_11th_Global_Fraud_Survey.

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

FERNANDEZ-DELGADO, M.; E. CERNADAS; S. BARRO; AND D. AMORIM. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 15 (2014): 3133–81.

FREUND, Y., AND R. SCHAPIRE. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55 (1997): 119–39.

GALAR, M.; A. FERNANDEZ; E. BARRENECHEA; H. BUSTINCE; AND F. HERRERA. "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches." *IEEE Transactions on Systems, Man, and Cybernetics, Part C–Applications and Reviews* 42 (2012): 463–84.

GLEASON, C.; N. T. JENKINS; AND W. B. JOHNSON. "The Contagion Effects of Accounting Restatements." *The Accounting Review* 83 (2008): 83–110.

GOLDMAN, E.; U. PEYER; AND I. STEFANESCU. "Financial Misrepresentation and Its Impact on Rivals." *Financial Management* 41 (2012): 915–45.

GREEN, P., AND J. H. CHOI. "Assessing the Risk of Management Fraud Through Neural Network Technology." *Auditing: A Journal of Practice & Theory* 16 (1997): 14–29.

HAN, J. W.; M. KAMBER; AND J. PEI. *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann, 2006.

HASTIE, T.; R. TIBSHIRANI; AND J. H. FRIEDMAN. *The Elements of Statistical Learning*. New York: Springer, 2009.

HE, H., AND Y. MA. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ: Wiley, 2013.

HUNG, M.; T. J. WONG; AND F. ZHANG. "The Value of Political Ties Versus Market Credibility: Evidence from Corporate Scandals in China." *Contemporary Accounting Research*, 32 (2015): 1641–75.

JÄRVELIN K., AND J. KEKÄLÄINEN. "Cumulated Gain-Based Evaluation of IR Techniques." *ACM Transactions on Information Systems* 20 (2002): 422–46.

JOHNSON, S. A.; H. E. RYAN; AND Y. S. TIAN. "Managerial Incentives and Corporate Fraud: The Sources of Incentives Matter." *Review of Finance* 13 (2009): 115–45.

KARPOFF, J. M.; A. KOESTER; D. S. LEE; AND G. S. MARTIN. "Proxies and Databases in Financial Misconduct Research." *The Accounting Review* 92 (2017): 129–63.

KEDIA, S., AND T. PHILIPPON. "The Economics of Fraudulent Accounting." *The Review of Financial Studies* 22 (2007): 2169–99.

KEDIA, S., AND S. RAJGOPAL. "Do the SEC's Enforcement Preferences Affect Corporate Misconduct?" *Journal of Accounting and Economics* 51 (2011): 259–78.

KHOSHGOFTAAR, T. M.; J. VAN HULSE; AND A. NAPOLITANO. "Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data." *IEEE Transactions on Systems, Man, and Cybernetics, Part A–Systems and Humans* 4 (2011): 552–68.

KLEINBERG, J.; J. LUDWIG; S. MULLAINATHAN; AND Z. OBERMEYER. "Prediction Policy Problems." *American Economic Review: Papers & Proceedings*, 105 (2015): 491–95.

LARCKER, D., AND A. A. ZAKOLYUKINA. "Detecting Deceptive Discussion in Conference Calls." *Journal of Accounting Research* 50 (2012): 495–540.

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.

LIU, X.-Y.; WU, J.; AND ZHOU, Z.-H. "Exploratory Undersampling for Class-Imbalance Learning." *IEEE Transactions on Systems, Man, and Cybernetics, Part B–Cybernetics* 39 (2009): 539–50.

LIU, X.-Y., AND Z.-H. ZHOU. "Ensemble Learning Methods for Class Imbalance Learning," in *Imbalanced Learning: Foundations, Algorithms, and Applications*, edited by H. He and Y. Ma. Hoboken, NJ: John Wiley & Sons Inc., 2013: 61–82.

LO, K.; F. RAMOS; AND R. ROGO. "Earnings Management and Annual Report Readability." *Journal of Accounting Research* 63 (2017): 1–25.

MURPHY, K. J. "Executive Compensation." *Handbook of Labor Economics* 3 (1999): 2485–563.

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

PEROLS, J. L.; R. M. BOWEN; C. ZIMMERMANN; AND B. SAMBA. "Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction." *The Accounting Review* 92 (2017): 221–45.

RAKOFF, J. S. "The Financial Crisis: Why Have No High-Level Executives Been Prosecuted?" The New York Review of Books, January 9, 2014.

SCHRAND, C. M., AND S. L. C. ZECHMAN. "Executive Overconfidence and the Slippery Slope to Financial Misreporting." *Journal of Accounting and Economics* 53 (2012): 311–29.

SEIFFERT, C.; T. M. KHOSHGOFTAAR; J. VAN HULSE; AND A. NAPOLITANO. "Rusboost: A Hybrid Approach to Alleviating Class Imbalance." *IEEE Transactions on Systems, Man, and Cybernetics, Part A–Systems and Humans* 40 (2010): 185–97.

SHMUELI, G. "To Explain or to Predict." *Statistical Science* 25 (2010): 289–310.

SUMMERS, S. L., AND J. T. SWEENEY. "Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis." *The Accounting Review* 73 (1998): 131–46.

TUV, E.; A. BORISOV; G. RUNGER; AND K. TORKKOLA. "Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination." *The Journal of Machine Learning Research* 10 (2009): 1341–66.

WANG, T. Y.; A. WINTON; AND X. YU. "Corporate Fraud and Business Conditions: Evidence from IPOs." *The Journal of Finance*, 65 (2010): 2255–92.

WANG Y.; L. WANG; Y. LI; D. HE; W. CHEN; AND T.-Y. LIU. "A Theoretical Analysis of NDCG Ranking Measures." In *Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

WITTEN, I. H., AND E. FRANK. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann, 2005.

WU, X.; V. KUMAR; J. R. QUINLAN; J. GHOSH; Q. YANG; H. MOTODA; G. J. MCLACHLAN; A. NG; B. LIU; P. S. YU; Z-H ZHOU; M. STEINBACH; D. J. HAND; AND D. STEINBERG. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14 (2008): 1–37.

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