Detecting Accounting Frauds in Publicly Traded U.S. Firms: New Perspective and New Method

Yang Bao[1], **Bin Ke**[2], **Bin Li**[3], **Julia Yu**[4], **and Jie Zhang**[5]

October 7, 2015

We wish to thank Mark Cecchini and workshop participants at the Singapore Tri-Uni Accounting Research Conference, 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).

[1] Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, P.R. China 200030. Tel: +86 15821616047. Email: baoyang@sjtu.edu.cn.

[2] Department of Accounting, NUS Business School, National University of Singapore, Mochtar Riady Building, BIZ 1, # 07-53, 15 Kent Ridge Drive, Singapore 119245.Tel: +65 6601 3133. Fax: +65 6773 6493. Email: bizk@nus.edu.sg.

[3] Department of Finance, Economics and Management School, Wuhan University, Wuhan, P.R. China 430072. Tel: +86-27-6875 6536. Email: binli.whu@whu.edu.cn.

[4] Division of Accounting, Nanyang Business School, Nanyang Technological University, S3-01c-84, 50 Nanyang Avenue, Singapore 639798. Tel: +65 6790 6016. Email: juliayu@ntu.edu.sg.

[5] School of Computer Engineering, Nanyang Technological University, N4-02c-100, 50 Nanyang Avenue, Singapore 639798. Tel: +65 6790 6245. Email: zhangj@ntu.edu.sg.

Detecting Accounting Frauds in Publicly Traded U.S. Firms: New Perspective and New Method

ABSTRACT

We propose a new perspective and a new method to detect accounting frauds out of sample. We show that a logistic regression that directly uses raw accounting data as regressors outperforms the traditional logistic regression that uses expert-identified financial ratios. Using the same raw data as inputs, ensemble learning, a state-of-the-art machine learning method, further outperforms the logistic regression model. The ensemble method also outperforms a support vector machine (SVM) with a financial kernel that maps the same raw accounting data into a broader set of ratios and changes in ratios. Overall, our results suggest that the existing fraud prediction models haven't fully utilized the information from publicly available financial statement data. In addition, we show that it is possible to extract such useful information by adopting better fraud prediction models that rely on raw accounting data rather than financial ratios as model inputs.

1. Introduction

Accounting frauds are a worldwide problem. If not detected and prevented on a timely basis, frauds can cause significant harm to not only the stakeholders of fraudulent firms directly (e.g., Enron and WorldCom) but also the stakeholders of many non-fraudulent firms indirectly (Gleason, Jenkins, and Johnson [2008], Goldman, Peyer, and Stefanescu [2012], Hung, Wong, and Zhang [2015]).[1] Unfortunately, accounting frauds are rare and therefore difficult to detect. For example, the frequency of detected accounting frauds among publicly traded U.S. firms is typically less than one percent. Even if a fraud is detected, it is usually too late (Dyck, Morse, and Zingales [2005]) and the real damages are already done by the time of the fraud's disclosure. Hence, an important research question in academic research is to develop effective methods to detect corporate accounting frauds on a timely basis so that the extent of damages from such frauds can be minimized.

The objective of this study is to use a sample of publicly traded U.S. firms to develop a new accounting fraud prediction model using only publicly available financial data as inputs. While there are useful non-financial indicators of frauds (e.g., an executive's personal behavior), we use only financial data as inputs for two reasons. First, fraud detection models based on publicly available financial data can be readily applied to any publicly traded firm at low costs. Second, most prior accounting fraud research relies on publicly available financial data (e.g., Green and Choi [1997], Summers and Sweeney [1998], and Beneish [1999], Cecchini, Aytug, Koehler, and Pathak [2010], Dechow, Ge, Larson, and Sloan [2011]). Hence, the performance of our methods can be readily compared with the performance of traditional fraud detection methods that also use financial data.

We depart from the existing fraud detection research in several important ways. First, we directly use raw accounting data from the financial statements to predict frauds. The

[1] See Schiesel [2002] for an interesting account of how WorldCom's accounting fraud affected the business strategies of AT&T (WorldCom's main competitor).

extant accounting fraud detection literature typically uses financial ratios identified by accounting experts as inputs (see Dechow et al. [2011] for a review). Ratio analysis based on the DuPont model is an integral part of the traditional accounting valuation analysis. Hence, it is natural for accounting experts to use intuitive financial ratios such as change in ROA as fraud determinants. However, this traditional ratios-based fraud detection approach suffers from two important limitations. The first limitation is that the extant fraud detection models haven't adopted a systematic framework to build fraud detection models and hence there is no guarantee that accounting experts have identified all relevant financial ratios useful for fraud prediction. The second limitation is that the conversion of raw accounting data into a limited number of financial ratios could result in a loss of useful information. To our best knowledge, we are the first study to examine whether raw accounting data are more useful than financial ratios for the purposes of fraud prediction.

Second, we use ensemble learning, one of the state-of-the-art paradigms in machine learning, to predict frauds. Most prior fraud detection research in accounting uses the logistic regression (see Dechow et al. [2011]). While ensemble learning has been successfully applied in many other fields (see Zhou [2012] for a review), we are the first study to apply the ensemble method to a setting with a severe class imbalance problem (i.e., the rarity of frauds). Hence, it is an empirical question whether the ensemble method can outperform the traditional fraud detection methods in our setting.

Third, our primary objective is to predict accounting frauds *out of sample*. While there is a fairly large accounting literature on fraud detection, the primary objective of most empirical studies is to explain frauds within sample and emphasize causal inferences. While related, predicting frauds out of sample is a fundamentally different task. Shmueli [2010] explains that the focus in explanatory modeling is on minimizing the bias resulting from model misspecification to obtain the most accurate representation of the underlying theory. In

contrast, predictive modeling seeks to minimize the combination of the bias and estimation variance which results from using a sample to estimate model parameters. Because of the bias and estimation variance tradeoff, a better specified fraud detection model with a lower bias doesn't necessarily yield better fraud prediction out of sample.

To compare the out-of-sample performance of different fraud detection 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 or 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 a randomly chosen nonfraud observation (Fawcett [2006]).The AUC for random guesses is 0.50 and therefore any reasonable fraud detection model must have an AUC higher than 0.50.

Because the frequency of accounting frauds sanctioned by the SEC's AAERs is typically small, even the best performing fraud detection model from prior research (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 to investigate all the predicted frauds given the limited resources available to fight fraud (Ernst & Young [2010]). Even if one wishes to investigate all the predicted fraud observations, 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. Hence, there is a natural demand for investigating only a small number of observations with the highest predicted fraud probabilities. Accordingly, we also examine the out-of-sample performance of the different fraud detection models using an alternative performance evaluation metric commonly used for ranking problems, referred to as Normalized Discounted Cumulative Gain at the position k (NDCG@k). Intuitively, NDCG@k assesses the ability of a fraud detection model to identify true frauds by picking only 1% of the observations in a test year that have the highest predicted probability of fraud

(i.e., k=the number of top 1% observations). We select a cutoff of 1% because the average frequency of accounting frauds sanctioned by the SEC's AAERs is typically less than 1% each year.

For the purposes of the out-of-sample performance evaluation comparison, we adopt two types of benchmark fraud detection models from the extant literature. The first type is the ratios-based logistic regression model commonly used in the accounting literature. Because Dechow et al. [2011] is the most recent comprehensive fraud detection study in the accounting literature, we adopt their logistic regression model based on 11 financial ratios commonly used in the extant literature as our first type of benchmark model (referred to as the basic Dechow et al. model).[2] The second type of benchmark model is a fraud detection model developed by Cecchini et al. [2010] based on more advanced machine learning. Rather than using the financial ratios identified by accounting experts alone, Cecchini et al. [2010] develop a new fraud detection model based on support vector machines (SVM) with a financial kernel that maps raw accounting data into a broader set of ratios and changes in ratios. Cecchini et al. [2010] find that the SVM with a financial kernel outperforms the traditional fraud detection models in accounting, including the ratios-based logistic regression model in Dechow et al. [2011].

We perform all of our empirical analyses using the sample period 1991-2005. Our sample starts in 1991 because there appear significant changes in both the determinants and detection of accounting frauds around 1990 and therefore it may not be appropriate to use the pre-1991 period to predict frauds after 1990. Our sample ends in 2005 because the SEC's Accounting and Auditing Enforcement Releases (AAERs) available to us end in September 2010 while there is an average of five-year gap between a fraud occurrence and the AAER publication date. Hence, a significant portion of the accounting frauds that occurred over

[2] It is important to note that we don't fully replicate Dechow et al.'s [2011] models which contain both financial ratios and non-financial variables. Hence, the results from this study are not directly comparable to those from Dechow et al. [2011].

2006-2010 are likely still unreported by the AAERs as of the end of 2010. Following Cecchini et al. [2010] and Dechow et al. [2011], we use the last three years 2003-2005 as the out-of-sample test period.

To make the performance comparison fair across the different models, we start with a common set of raw accounting data items. Specifically, we follow Cecchini et al. [2010] by selecting 24 raw accounting data items from the financial statements for our fraud detection models. Cecchini et al. [2010] selected the 24 raw accounting data items after reviewing a comprehensive list of existing academic papers on fraud prediction. The set of raw accounting data items used in the basic Dechow et al. model significantly overlaps with but is not identical to the set of 24 raw accounting data items used by Cecchini et al. [2010]. In order to give the ratios-based logistic regression model a fair chance, we also estimate a modified Dechow et al. model based on 11 expert identified financial ratios from the basic Dechow et al. model plus three additional expert identified financial ratios derived from Cecchini et al.'s 24 raw accounting data items not used in the basic Dechow et al. model. As a result, the modified Dechow et al. model contains 14 financial ratios derived from 28 raw accounting data items.

We first report the performance results of the two types of benchmark models. Using the first performance evaluation metric AUC, we find that the out-of-sample performance of the two types of benchmark models is significantly better than the performance of random guesses. The average AUC is 0.638 for the basic Dechow et al. model, 0.654 for the modified Dechow et al. model, and 0.740 for the Cecchini et al. model. Consistent with Cecchini et al. [2010], Cecchini et al.'s model significantly outperforms both the basic and modified Dechow et al. models in terms of AUC in our sample. However, the performance of all the benchmark models fares poorly using the second performance evaluation metric NDCG@k. Specifically, the top 1% of the observations with the highest predicted fraud probability in the

test years 2003-2005 can only capture less than 3% of the true frauds in the population. In addition, less than 2% of the top 1% of the observations with the highest predicted fraud probability are true frauds for all the benchmark models.

We find evidence that raw accounting data are more useful than expert identified financial ratios in predicting frauds out of sample. Specifically, a simple logistic regression that directly uses the 24 raw accounting data items as regressors outperform both the basic and modified Dechow et al. models by a significant margin in terms of AUC (0.747 versus 0.638 and 0.654 respectively or an increase of 15-17%). Furthermore, the AUC performance of our simple logistic regression based on the 24 raw accounting data items is on par with that of Cecchini et al.'s more advanced and complex SVM-FK method based on the same 24 raw accounting data. These results suggest that the financial ratios identified by accounting experts and used in the traditional ratios-based logistic regression models have not fully utilized the information from the raw accounting data. However, using the second performance evaluation metric NDCG@k, we find no evidence of significant difference in the performance of the logistic regression model based on the 24 raw accounting data items versus the two types of benchmark models.

More importantly, we find that the ensemble method using the 24 raw accounting data items has an average AUC of approximately 82%, significantly higher than the average AUC of all the benchmark models as well as the logistic regression model using the 24 raw accounting data items The ensemble method outperforms the next best model, the logistic regression model using the same 24 raw accounting data items, by more than 9% in AUC (0.822 versus 0.747), and outperforms the basic Dechow et al. model by more than 28% (0.822 versus 0.638). More significantly, we find that the ensemble method significantly outperforms all the other fraud detection models in terms of NDCG@k. Specifically, we find that the top 1% of the observations with the highest predicted fraud probability from the

ensemble model captures 28.09% of the true frauds in the population. In addition, 18.59% of the top 1% of the observations with the highest predicted fraud probability from the ensemble model are true frauds. These results are impressive considering the fact that less than 1% of the firms in the population are true frauds. While still far from perfect, our empirical results from NDCG@k suggest that our ensemble method based on raw accounting data is significantly better in out-of-sample performance relative to the best fraud detection models available in the extant literature.

We contribute to the existing fraud detection literature in several important ways. First, we are the first to propose using raw accounting data rather than ratios to predict accounting frauds. The promising results from this study raise the possibility that we can further improve the performance of future fraud prediction models by using hundreds of additional readily available raw accounting data items from the financial statements. Second, we are the first study to demonstrate the usefulness of ensemble learning, a state-of-the-art machine learning method, in fraud prediction, a challenging task due to severe class imbalance. Third, we are the first to introduce a new performance evaluation metric for ranking problems, NDCG@k, to the setting of fraud prediction. Considering the significant costs of investigating a larger number of false positive fraud observations, NDCG@k provides a valuable alternative benchmark to rank fraud prediction models. Finally, our results also have important implications for the ongoing research that compares the usefulness of text data versus quantitative data to predict 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 Dechow et al.'s ratios-based logistic regression model significantly understates the value of financial data in fraud prediction.

The rest of the paper is organized as follows. Section two provides an introduction to the ensemble method. Section three discusses the out-of-sample performance evaluation metrics. Section four introduces the accounting fraud sample and explains the selection of the 24 raw accounting data items. Section five reports the out-of-sample performance of the basic and modified Dechow et al. models. Section six reports the out-of-sample performance of Cecchini et al.'s SVM-FK method. Section seven discusses the logistic regression results of fraud prediction using raw accounting data items. Section eight analyzes its out-of-sample performance of the ensemble method. Section nine concludes.

2. An Introduction to Ensemble Learning

Ensemble learning, one of the main paradigms in machine learning, has achieved great success in many real-word applications in recent years (see pages 17-19 in Zhou [2012] for a review of applications of ensemble methods). Different from conventional machine learning methods (e.g., SVM methods) which usually generate one single estimator, ensemble method combines the predictions of a set of base estimators (e.g., decision trees) in order to improve the generalizability or robustness over any single estimator. Previous studies (Zhou [2012]) show that ensembles can usually outperform any single base estimator. However, due to the class imbalance problem, conventional ensemble methods usually need to be combined with a sampling technique in order to balance the class distribution of training data by either adding examples to the minority class (over-sampling) or removing examples from the majority class (under-sampling) (Liu and Zhou [2013]). There have been various methods for combining ensemble methods and data sampling techniques for class imbalance learning (Liu and Zhou [2013]). In this study we employ one variation of ensemble learning called RUSBoost (Seiffert, Khoshgoftaar, Van Hulse, and Napolitano [2010]). RUSBoost seeks to take advantage of both the efficient under-sampling technique

(Liu, Wu, and Zhou [2009]) and the most influential ensemble algorithm called AdaBoost (Freund and Schapire [1997]). A review by Galar, Fernandez, Barrenechea, Bustince, and Herrera [2012] find RUSBoost (Seiffert et al. [2010]) to demonstrate the best performance and are also more computationally efficient due to its simplicity. The superiority of RUSBoost is further confirmed by Khoshgoftaar, Van Hulse, and Napolitano [2011].

In the following, we first introduce the AdaBoost algorithm, and then describe how it is combined with an under-sampling technique in our empirical implementation of RUSBoost. The AdaBoost algorithm is one of the most important ensemble methods due to its solid theoretical foundation, strong predictive power, and great 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, while the weights of correctly classified observations will be decreased. In this way, the weak classifier in each iteration will be forced to concentrate on the observations that are difficult to predict in the previous iterations. Finally, a strong classifier can be produced by taking the weighted average of all week classifiers, where the weight is based on a weak classifier's classification error rate on the training sample. The weak classifiers with lower classification error rates will receive higher weights.

RUSBoost is a variant of AdaBoost that combines the Random Under-Sampling (RUS) for class imbalance learning (Seiffert et al. [2010]). It works in the same manner as AdaBoost except that RUS is performed in each iteration to address the imbalance of fraud and nonfraud firms. Specifically, when training the weak classifier in each iteration, the RUS algorithm uses the full sample of fraud firms in the training period and a randomly generated

subsample of nonfraud firms in the same training period.[3] The estimation of RUSBoost requires the selection of the ratio between the number of under-sampled majority class observations (i.e., nonfraud) and the number of minority class observations (i.e., fraud). In this paper, we construct two versions of RUSBoost by setting this ratio equal to 1:1 and 2:1. That is, when the ratio is set to 1:1, we simply sample the same number of fraud observations and non-fraud observations. When the ratio is set to 2:1, we sample non-fraud observations twice as many as fraud observations.

3. Performance Evaluation

3.1. OUT-OF-SAMPLE PERFORMANCE EVALUATION

Most existing fraud detection papers in accounting don't evaluate the out-of-sample performance of their models. However, the in-sample prediction error is a very optimistic estimate of performance in an out-of-sample data set. Hence, it is necessary to assess the out-of-sample prediction error of a fraud detection model in 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 [2003]) as follows: (1) split the data into n roughly equal samples (folds), where n is typically set at 10; (2) estimate the fraud detection model using only n-1 folds, leaving one fold out for out-of-sample performance evaluation; and (3) the out-of-sample performance of the model is evaluated using the left-out fold. These steps are repeated n times, one for each of the above n samples.

Our fraud data are inter-temporal in nature and hence performing the standard n-fold cross validation as suggested above is less appropriate because such a cross-validation would

[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.

destroy the inter-temporal nature of our fraud data. Therefore, we follow Cecchini et al. [2010] and Dechow et al. [2011] by using the last three years of our sample period 2003-2005 as the out-of-sample test period and a window of at least 11 years to train and validate our fraud detection models.[4] Specifically, we use 1991-2001 as the training period for test year 2003, 1991-2002 as the training period for test year 2004, and 1991-2003 as the training period for test year 2005. To minimize the look-ahead bias, we leave a gap of two years between the training period end and a test year because on average it takes approximately two years for the discovery and disclosure of a fraud (Dyck et al. [2005]). To our best knowledge, all of the prior published fraud detection models don't impose the two-year gap requirement.

3.2. OUT-OF-SAMPLE PERFORMANCE EVALUATION METRIC ONE: AUC

Since our fraud prediction task can be cast as a binary classification problem (fraud versus nonfraud), we can measure the fraud detection performance using the evaluation metrics for classification problems. One standard classification performance metric is accuracy, defined as $\frac{TP+TN}{TP+FN+FP+TN}$, where TP (True Positive) is the number of fraud firm-years that are correctly classified as frauds; FN (False Negative) is the number of fraud firm-years that are misclassified as non-frauds; TN (True Negative) is the number of non-fraud firm-years that are correctly classified as non-frauds; and FP (False Positive) is the number of non-fraud firm-years that are misclassified as frauds. Unfortunately, this standard classification performance metric is not appropriate in our scenario due to the imbalanced nature of our fraud versus nonfraud data (recall the fraud percentage in our sample is less than 1% each year). For example, one naïve strategy of classifying all firm-years as non-frauds in our sample would lead to an extremely high accuracy of more than 99% based on the standard classification performance metric. However, fraud prediction models with such a high

[4] Inferences for our ensemble learning method are similar if the test period is extended to a longer five-year period 2001-2005 (untabulated).

performance are of little value in our scenario because our main concern is to accurately detect as many fraud firm-years as possible without misclassifying too many non-fraud firm-years. That is, we care about both the true negative rate (i.e., *specificity* defined below) and the true positive rate (i.e., *sensitivity* defined below).

To properly gauge the performance of a fraud detection model, one could use balanced accuracy (BAC) as an alternative performance evaluation metric (He and Ma [2013]), which is defined as the average of the fraud detection accuracy within fraud observations and the nonfraud detection accuracy within non-fraud observations. Specifically, $BAC = \frac{1}{2} * (sensitivity + specificity)$, where $Sensitivity = \frac{TP}{TP+FN}$ and $Specificity = \frac{TN}{TN+FP}$. 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 classifier, usually automatically determined by the classifier to maximize the BAC. In other words, by setting a different threshold of the classifier, one would obtain a different BAC value. In the absence of any knowledge on the costs of misclassifying false positives versus the costs of misclassifying false negatives, one couldn't determine the optimal predicted fraud probability threshold for the purposes of classifying frauds and nonfrauds. Second, measures such as S*ensitivity* 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 area under the Receiver Operating Characteristics (ROC) curve as our out-of-sample performance evaluation metric. A ROC curve is a two-dimensional depiction of a classifier's performance by combining the True Positive Rate (i.e., *Sensitivity*) and the False Positive Rate (i.e., 1-*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 detection model to a single scalar by computing the area under the ROC curve (AUC). Since 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 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).

3.3. OUT-OF-SAMPLE PERFORMANCE EVALUATION METRIC TWO: NDCG@K

Our fraud detection task can also be thought of as a ranking problem by limiting our out-of-sample performance evaluation to only a small number (i.e., k as defined below) of firm years with the highest probability of fraud predicted by a fraud detection model. In this scenario, the performance of a fraud detection model can be measured by the following performance evaluation metric for ranking problems: Normalized Discounted Cumulative Gain at the position k (NDCG@k). NDCDG@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 proved to be effective theoretically (Wang, Wang, Li, He, Chen, and Liu [2013]).

Formally, the Discounted Cumulative Gain at the position k (DCG@k) is defined as follows: $DCG@k = \sum_{i=1}^{k}(2^{rel_i} - 1) \,/\, log_2\,(i+1)$, where rel_i equals 1 if the i-th 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 probability of fraud predicted by a fraud detection model (referred to as the ranking list). In our subsequent empirical analyses, we select a 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 frauds sanctioned by the SEC's AAERs is typically less than 1% each year.

The definition of DCG@k has two 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, i.e., the denominator $log_2\ (i + 1)$.

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

Relative to the first performance evaluation metric AUC, NDCG@k avoids the significant direct and indirect costs of investigating a large number of predicted frauds, the majority of which are likely false positives due to the severe data imbalance of true frauds. Because the average frequency of accounting frauds sanctioned by the SEC's AAERs is typically less than 1%, even the best performing fraud detection model from prior research (e.g., Cecchini et al. [2010]) would result in 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 non-fraud observations in the out-of-sample test period, the best among the competing models considered in their study. However, applying Cecchini et al.'s model to our test period 2003-2005 would result in too many false positives. Because only 118 (0.67%) of the 17,716 firm years are true frauds in our test period 2003-2005, Cecchini et al.'s method would mislabel 1,654 ((1-90.6%)*(17,716-118)) non-fraud observations as frauds, significantly bigger than the number of true frauds in the test period. Clearly, it is impractical to investigate all the predicted frauds

due to limited resources. Even if one wishes to investigate all the predicted fraud observations, the direct and indirect costs of investigating so many predicted frauds would be prohibitively large while the benefit would be small because the majority of the predicted fraud observations are false positives. NDCG@k avoids this problem by limiting its investigation to no more than k number of firm years in the test period (i.e., top 1% or 177 predicted frauds in our setting).

4. The Data

4.1. THE FRAUD SAMPLE

Our initial accounting fraud sample comes from the SEC's *Accounting and Auditing Enforcement Releases* (AAERs) over the period between May 17th, 1982 and September 1st, 2010.[5] Following Dechow et al. (2011), we exclude firms in the insurance and banking industry (i.e. SIC 6000-6999). The AAERs cover accounting frauds that occurred during the fiscal years 1971-2008. However, there are only 13 fraud firm-years before 1979. Table 1 shows the distribution of the frauds by year over the fiscal years 1979-2008. There are a total of 785 fraud firm years, but the frequency of frauds is very low, typically less than 1% per year. The fraud percentage is abnormally low in the last three years 2006-2008, largely due to the fact that our sample of AAERs ends in 2010 but it typically takes five years from the fraud occurrence to the AAER publication date. Hence, we exclude the last three years 2006-2008 from our sample.

We also exclude the years before 1991 because there is a significant shift in U.S. firms' fraud behavior. First, the SEC changed its enforcement around 1990. In their review of the history and evolution of the SEC enforcement program, Atkins and Bondi [2008] suggest

[5] We thank Dechow et al. [2011] for allowing us access to their hand collected AAER fraud data.

that the SEC's enforcement program shifted from a remedial purpose to a punitive purpose in nature in the 1990s. Prior to 1990, the SEC's statutory purpose was to provide remedial relief for aggrieved investors and to deter future violations. Since mid-1980s through late 1980s, the Congress passed a series of laws that expanded the SEC's powers and provided the SEC with new penalty authority. These laws include the Insider Trading Sanctions Act of 1984, the Insider Trading and Securities Fraud Enforcement Act of 1988, and the Securities and Enforcement Remedies and Penny Stock Reform Act of 1990. As a result of these laws, the SEC gained three significant new sets of powers: (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]). The SEC started to seek or impose more punitive actions, such as civil monetary penalties to a broad range of conduct in 1990.

Second, the use of stock options as a form of executive compensation rose dramatically during the 1990s, as did other forms of 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 of accounting frauds in the AAERs published in the 1990s than in the 1980s (Beasley et al. [1999] and [2010]).[6]

Third, an analysis of the accounting frauds included in the AAERs in the 1980s and 1990s by Beasley, Carcello, and Hermanson [1999] and Beasley, Carcello, Hermanson, and Neal [2010] respectively reveals subtle changes in the nature of the accounting frauds over

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

time. The two most common techniques used to fraudulently misstate financial statement information involved overstating revenues and assets in both time periods. However, misstatements through the understatement of expenses/liabilities became a more frequently used fraud technique in the 1990s (an increase from 18% to 31%).

4.2. THE RAW ACCOUNTING DATA

A key feature of our fraud detection models is to use raw accounting data directly from the financial statements. While there are a lot of raw accounting data items from a company's financial statements, we limit our empirical analyses to only 24 raw data items so that we can compare the performance between our fraud detection models and the traditional fraud detection methods whose fraud determinants are also derived from the same or a similar set of raw accounting data. Specifically, we follow Cecchini et al. [2010], one of the most recent and comprehensive studies on fraud detection, in the selection of the raw accounting data items. After reviewing a comprehensive list of existing academic papers, including Beneish [1999], Dechow et al. [2011], Summers and Sweeney [1998], and Green and Choi [1997], Cecchini et al. [2010, Table 3] identified an initial list of 40 raw accounting data items used by prior fraud detection research to construct the regression variables. Cecchini et al. retained a final list of 23 raw accounting data items after imposing the requirement that no raw variable has more than 25% of its values missing. Following the same sample selection procedures, we obtain a final list of 24 raw accounting data items during our sample period 1991-2005. Table 2 shows the list of the 40 raw accounting data items from Cecchini et al. [2010] in column (1) and our final list of 24 raw accounting data items in column (2).

As noted in the Introduction, we use Dechow et al.'s [2011] ratios-based logistic regression model as a benchmark for our subsequent out-of-sample performance analysis. Because the primary focus of our study is to use data from the readily available financial

statements to predict accounting frauds, we don't exactly replicate Dechow et al.'s models in Table 7 that use both financial and non-financial data. Specifically, we begin with the initial larger list of candidate regression variables identified by Dechow et al.'s [2011] Table 3. Dechow et al. [2011, Table 3] suggest three types of fraud determinants: (a) nine accruals quality related variables which can be calculated using the accounting numbers in the annual financial statements such as balance sheets and income statements; (b) two nonfinancial variables and four off-balance-sheet variables, which can be calculated using annual report disclosures; and (c) eight market-related variables which can be computed using either the annual report disclosures or stock price data or both. We exclude all the nonfinancial variables and off-balance-sheet variables. 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 the financial statements.[7] We include all five variables under the category of "Performance variables" except for *Deferred tax expense* because variable *Income Taxes, Deferred* required to calculate this ratio has more than 25% of its values missing in our sample period. Dechow et al. didn't include this variable either in their subsequent regression analyses (see their Tables 7 and 9). We keep only *Actual issuance* and *Book-to-market* under the category "Market-related incentives" because the raw accounting data for both variables are readily available in Compustat. In addition, Dechow et al. [2011] include *Actual issuance* in their basic model 1 of Table 7 and B*ook-to-market* in their out-of-sample prediction model in Table 9. Accordingly, our partial replication of Dechow et al.'s fraud detection model, referred to as the basic Dechow et al. model, contains 11 financial ratios from Dechow et al.'s Table 3.

[7] It is important to note that Dechow et al. [2011] didn't include these four discretionary accrual measures in their subsequent regression models either.

The last column of Table 2 displays the 23 raw accounting data items required to compute the 11 financial ratios used by the basic Dechow et al. model. While there is a significant overlap in the raw accounting data items between our replication of Cecchini et al.'s [2010] model in column (2) and the basic Dechow et al. model in column (3), there are a few key differences.[8] First, Dechow et al.'s raw variable list in column (3) contains three raw accounting data items (i.e., *Accounts Payable, Trade*, *Sale of Common and Preferred Stock*, and *Long-Term Debt Issuance*) excluded from Cecchini et al.'s raw data list in column (2). These three raw accounting data items are used to construct two financial ratios (*Change in cash margin* and *Actual issuance*) used in the the basic Dechow et al. model. Second, Dechow et al.'s raw variable list in column (3) excludes five raw accounting data items included in Cecchini et al.'s list in column (2): *Net Income (Loss)*, *Interest and Related Expense, Total*, *Income Taxes, Total*, *Retained Earnings*, *Depreciation and Amortization*. These raw accounting data items are used by Cecchini et al. to construct three financial ratios, *Depreciation index* in Beneish [1999], *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 list of raw accounting data between the basic Dechow et al. model in column (3) and Cecchini et al.'s model in column (2), we also augment the basic Dechow et al. model with three additional financial ratios noted above (i.e., *Depreciation Index*, *Retained earnings over total assets* and *EBIT*). As a result, this modified Dechow et al. model uses 28 raw accounting data, including the list of the entire 24 raw accounting data items from Cecchini et al.'s model in column (2).

[8] Dechow et al. [2011] use *Property, Plant, and Equipment, Net* while Cecchini et al. [2010] use *Property, Plant, and Equipment, Gross*. We don't believe this is a significant difference and therefore treat these two items as equivalent.

5. The Out-of-Sample Performance of the Basic and Modified Dechow et al. Models

We start with the estimation of the basic Dechow et al. model (11 regression variables) in Panel A of Table 3, using the sample firm years over the training period 1991-2001 for test year 2003 (column 1), 1991-2002 for test year 2004 (column 2), and 1991-2003 for test year 2005 (column 3). Five of the 11 regression coefficients are significant and in the predicted directions in at least one column. The other regression variables are insignificant.[9]

Panel B of Table 3 reports the regression results of the modified Dechow et al. model by adding three additional regression variables to the basic Dechow et al. model: *Depreciation index*, *Retained earnings over total assets*, and *EBIT* (14 regression variables). The inferences for the coefficients on the first eleven regression variables are qualitatively similar to those in Panel A of Table 3 except that the coefficient on *Change in inventory* loses significance. However, the coefficient on the newly added regression variable *Depreciation index* is significant and as predicted.

Table 4 reports the out-of-sample performance evaluation of the basic Dechow et al. model and the modified Dechow et al. model for the test period 2003-2005 using the evaluation metrics of both AUC and NDCG@k. Consistent with Dechow et al. (2011), the average AUC for the three test years is 0.638 for the basic Dechow et al. model and 0.654 for the modified Dechow et al. model, much higher than 0.50, the AUC of random guesses. The difference in the average AUCs for the two Dechow et al. models is also marginally significant (two-tailed p value=0.080). However, the average NDCG@ks are extremely low for both models (0.018 and 0.014 respectively). For example, the average value of *sensitivity*

[9] We include two accrual proxies (*WC accruals* and *RSST accruals*) in the same logit regression. Hence, the coefficients on these two variables could be hard to interpret due to potential multicollinearity. However, our primary objective is to predict frauds out of sample and hence multicollinearity is not a concern (Makridakis, Wheelwright and Hyndman [1998]). The out-of-sample performance is similar if we exclude one of the two proxies in the logit model.

(defined in Table 4) for the basic Dechow et al. model is 2.72%, meaning that only 2.72% of all the true frauds in the population is captured in the top 1% observations with the highest predicted fraud probabilities. Similarly, the average value of *precision* (defined in Table 4) for the basic Dechow et al. model is 1.68%, meaning that only 1.68% of the top 1% observations with the highest predicted fraud probabilities are true frauds.

6. Replication of Cecchini et al. [2010]

Cecchini et al. [2010] developed an innovative SVM method based on a financial kernel that maps raw accounting data into a list of predefined ratios that is broader than the ratio list used by prior fraud detection literature in accounting, hereafter referred to as SVM-FK.[10] Cecchini et al. [2010, Table 7] show that their SVM-FK significantly outperforms several representative fraud detection models in accounting, including Dechow et al. [2011].

In this section, we replicate Cecchini et al.'s 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 frauds and non-frauds). When searching for the optimal parameter $C^{+1}: C^{-1}$ that maximizes the value of AUC, Cecchini et al. [2010] directly perform the search using the test sample rather than a holdout validation sample and therefore Cecchini et al.'s implementation procedures are subject to a look-ahead bias, which is directly acknowledged on page 1156 of their paper. We avoid this limitation by performing the grid search in a holdout validation sample.

[10] Specifically, the SVM-FK method maps a firm's raw accounting data in two consecutive years (referred to as year 1 and year 2 respectively) into the following six types of predefined ratios, representing both intra-year ratios and year-over-year changes in ratios: $\phi(u) = \left(\frac{u_{1,i}}{u_{1,j}}, \frac{u_{1,j}}{u_{1,i}}, \frac{u_{2,i}}{u_{2,j}}, \frac{u_{2,j}}{u_{2,i}}, \frac{u_{1,i}u_{2,j}}{u_{1,j}u_{2,i}}, \frac{u_{1,j}u_{2,i}}{u_{1,i}u_{2,j}}\right), i, j = 1, \dots, n, i < j$, where $u_{1,i}$ is raw accounting data item i in year 1 and n represents the total number of raw accounting items. Since we have a total of 24 raw accounting data items (i.e., n=24), the above transformation function would result in a total of 1,656 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].

Specifically, we train the SVM-FK model using 1991-1999 and validate the model using 2000-2001 for the test years 2003-2005. We use two years instead of one year for validation because of the low frequency of frauds in a typical year.[11] After determining the optimal parameter C^{+1}: C^{-1} that maximizes the value of AUC in the validation period, we use the combined training and validation period 1991-2001 to retrain the model before performing the out-of-sample fraud prediction.

Second, we differ from Cecchini et al. [2010] by using 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 of frauds and nonfrauds in the training period for the purposes of model training and validation. However, it is problematic to use only the matched fraud and nonfraud firm years in the test period to evaluate the out-of-sample performance of the SVM-FK model because such a procedure is subject to the look-ahead bias and therefore couldn't be implemented in real time. Specifically, because it takes two years on average for the detection and disclosure of an accounting fraud (Dyck et al. [2005]), a relevant decision maker (e.g., a regulator or investor) typically doesn't know whether the financial statements of a company in an industry year are fraudulent or not in the year of the company's financial statement release and therefore could not match a fraud firm year to the nonfraud firm years in the same industry year. Instead, the decision maker requires a fraud prediction model to identify the predicted frauds from the entire *population* of firms in a test period rather than a hypothetical matched sample of fraud firms and nonfraud firms in the test period. Hence, a more appropriate out-of-sample performance

[11] The logistic regression and the ensemble method don't require a similar grid search because there are no parameters to tune.

evaluation approach is to evaluate SVM-FK's 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] follows Cecchini et al. by using a matched sample of frauds and nonfrauds for training and validation but uses the entire population of firm years in the test period 2003-2005 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 fraud and nonfraud firms constitutes only 16.50% (2,924/17,716) of the population of fraud and nonfraud firms in the test period 2003-2005. 118 (0.67%) of the 17,716 observations in the test period 2003-2005 are true frauds. In contrast, for the matched sample of frauds and nonfrauds based on industry and year in the test period 2003-2005, 98 (3.35%) of the 2,924 observations are true frauds. Untabulated results show that the average AUC for the SVM-FK model is 0.80 using a matched sample of fraud and nonfraud observations in the test period 2003-2005 but drops significantly using the full population of firm years in 2003-2005. Specifically, as shown in Table 4, the average AUC after correcting the two look-ahead biases is only 0.740. Nevertheless, the performance of the SVM-FK model is still significantly better than the performance of both the basic and modified Dechow et al. methods. Using the NDCG@k as an alternative evaluation criterion, we find that the average value of NDCG@k for our replication of Cecchini et al.'s SVM-FK method is only 0.017, comparable to those from the two Dechow et al. models. For the top 1% predicted frauds by the SVM-FK model in the test period 2003-2005, the average values of *sensitivity* and *precision* are always below 2%. Overall, we conclude that Cecchini et al.'s SVM-FK model performs better than the basic and modified Dechow et al. models in terms of AUC but performs similarly to the basic and modified Dechow et al. models in terms of NDCG@k.

7. Predicting Frauds Using the Logit Model Based on Raw Accounting Data

We next examine whether it is possible to improve the out-of-sample performance of the logistic regression model using raw accounting data directly as fraud determinants. Table 5 shows the logistic regression results of fraud prediction using Cecchini et al.'s [2010] 24 raw accounting data items. To minimize the effect of scale differences across different raw data items, it is common in the data mining literature to normalize the raw data items (Han, Kamber, and Pei [2006]).[12] In this study 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, i.e., $x' = \frac{x}{||x||}$, where the division are element-wise. For example, a vector of $(1, 2)$ is normalized to $(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$. Six of the 24 raw accounting data items are significant in at least one of the three regressions, involving three asset accounts, one liability account, and two expense accounts. These six raw data items are used in the computation of the following six financial ratios in Table 3: *WC accruals*, *RSST accruals*, *%Soft assets*, *Change in cash margin*, *Actual issuance*, and *Depreciation Index*. It is interesting to note that the coefficients on *WC accruals*, *RSST accruals*, and *Change in cash margin* are insignificant in Table 3, suggesting direct evidence that converting the raw accounting data into financial ratios would result in loss of useful information for the purposes of fraud prediction.

As shown in Table 4, the average AUC for the logistic regression model based on the 24 raw accounting data items is 0.747, significantly higher than the average AUC for both the basic and modified Dechow et al. models (the p values for the differences are 0.057 and 0.074, respectively), but no different from the average AUC for the more computationally intensive SVM-FK model. However, the mean value of NDCG@k for the logistic regression based on the 24 raw data items is similar to the mean values of NDCG@k for the two

[12] 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).

Dechow et al. methods and the SVM-FK model. Given that the 24 raw data items are a subset of the 28 raw data items used in the modified Dechow et al. model, the better out-of-sample AUC performance of the logistic model based on the 24 raw data items relative to the modified Dechow et al. model suggests that the financial ratios identified by accounting experts have not fully extracted the useful information from the raw accounting data for the purposes of fraud prediction.

8. Predicting Frauds Using Ensemble Learning

8.1. THE OUT-OF-SAMPLE PERFORMANCE RESULTS

Limiting to the same 24 raw accounting data items, we next examine whether it is possible to further improve the out-of-sample fraud prediction performance by using the more advanced data mining method, ensemble learning, rather than the logistic regression method.[13] Table 4 reports the out-of-sample performance of the two variations of the RUSBoost ensemble learning (denoted as RUSBoost (2:1) and RUSBoost (1:1)) over the test period 2003-2005 based on the same 24 raw accounting data items. The average AUCs for the two ensemble method variations are almost identical (approximately 0.82), consistent with prior research that one advantage of ensemble methods is that they are not sensitive to model parameters as other methods (e.g., SVM). More importantly, the average AUCs for both ensemble method variations are significantly larger than the average AUC for the logistic regression based on the 24 raw accounting data items (0.747). Using NDCG@k as an alternative evaluation criterion, we find that the values of NDCG@k for the two ensemble method variations are significantly larger than the values of NDCG@k for the logistic regression based on the 24 raw accounting data items. For the top 1% predicted frauds in the test period 2003-2005, the average values of *sensitivity* and *precision* for the best ensemble

[13] In untabulated analyses we also tried the SVM method based on raw accounting data items and found no evidence that it outperforms our ensemble method discussed below.

model RUSBoost(2:1) are 28.09% and 18.59%, respectively. In contrast, the corresponding values for the logit model based on the 24 raw accounting data items are only 2.14% and 1.69%, respectively. Considering the fact that both the ensemble method and the logistic regression start with the same 24 raw accounting data items, our results suggest that the ensemble method is more powerful in predicting frauds out of sample than the logistic regression typically used in accounting.

The results in Table 4 also show that the performance of the ensemble method is significantly better than the performance of Cecchini et al.'s [2010] SVM-FK model using both performance evaluation metrics. The average AUC for the ensemble models (around 0.82) is significantly higher than the average AUC of the SVM-FK model (0.740). Likewise, the average value of NDCG@k for the best ensemble model (0.329) is significantly higher than the average value of NDCG@k for the SVM-FK model (0.017).

8.2. UNCOVERING THE BLACK BOX BEHIND THE PERFORMANCE OF THE ENSEMBLE METHOD

One 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. While individual decision trees can be interpreted easily by simply visualizing the tree structure, ensemble 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 features (i.e., raw data items in our setting) in fraud prediction (Tuv, Borisov, Runger, and Torkkola [2009]). In this paper, we use the "predictorImportance" function implemented in MATLAB to estimate the feature importance of the 24 raw accounting items

used in our best ensemble model RUSBoost (2:1).[14] 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, Friedman, Stone, and Olshen [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 6 reports the descriptive statistics on the importance of 24 raw data items for our RUSBoost (2:1) model for each of the three test years as well as the average of the three years combined. The 24 raw data items are sorted from high to low based on the average feature importance over the three test years in Table 6. Figure 1 also visualizes the feature importance for the three test years. The top 10 most important features (or raw data items) that contribute to RUSBoost (2:1) model's superior performance are listed below starting with the most important: (1) *Cash and Short-Term Investments*; (2) *Assets, Total*; (3) *Common Shares Outstanding*; (4) *Price Close, Annual, Fiscal*; (5) *Property, Plant and Equipment, Total*; (6) *Depreciation and Amortization*; (7) *Inventories, Total*; (8) *Cost of Goods Sold*; (9) *Receivables, Total*; and (10) *Debt in Current Liabilities, Total*.

To benchmark the performance of the RUSBoost (2:1) model, Panel B of Table 6 also lists the specific financial statement accounts affected by the misstatements reported by the AAERs over the same period 2003-2005. We directly obtained the data used in Panel B, including the ten individual account categories, from Dechow et al. [2011], who in turn collected this data from the AAERs. We can draw the following conclusions from the top ten lists of both panels. First, there is a significant overlap of the two top ten lists as shown in the last column of Panel B. This evidence suggests that our ensemble method has been relatively effective in identifying the specific accounts affected by the misstatements. Second, while the

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

account "*rev*" in Panel B is the most frequently affected account by misstatements, the same account "*Sales/Turnover (Net)*" in Panel A is only ranked #17 in terms of feature importance. However, "*Sales/Turnover (Net)*" is also closely related to "*Receivables, Total*" and we wish to note that the RUSBoost (2:1) model ranked "*Receivables, Total*" #9 in terms of feature importance in Panel A. Third, "*Common Shares Outstanding*" and "*Price Close, Annual, Fiscal*" are not part of the ten individual account categories in Panel B but they are ranked #3 and #4, respectively, in terms of feature importance in Panel A. We suspect that "*Common Shares Outstanding*" is useful in predicting accounting frauds simply because misstatement firms often issue common equity during the misstatement period. Similarly, "*Price Close, Annual, Fiscal*" provides valuable information about the likelihood of frauds because of informed trading during the misstatement period by both company employees and other related stakeholders. We leave to future research to confirm the validity of these two conjectures.

9. Conclusion

Accounting frauds are rare and difficult to detect. Hence, an important research question in accounting research is to develop effective methods to detect corporate accounting frauds on a timely basis so that the extent of damages from such frauds can be minimized. The objective of this study is to develop a new out-of-sample fraud detection model based on a sample of publicly traded U.S. firms over the period 1991-2005. To preserve the inter-temporal nature of fraud prediction, we follow Cecchini et al. [2010] and Dechow et al. [2011] by using the last three years of our sample period 2003-2005 as the out-of-sample test period and the years prior as the training period. To avoid potential look-ahead bias we also require a minimum gap of two years between the training period end and a test

year because on average it takes approximately two years for the discovery and disclosure of a fraud (Dyck et al. [2005]).

Following prior research, we use only readily available financial data as inputs in fraud prediction, but we depart from prior fraud research in accounting in several important ways. First, we directly use raw accounting data from the financial statements to predict frauds. In contrast, the extant accounting fraud detection research typically uses financial ratios identified by accounting experts to predict frauds. Second, we use ensemble learning, one of the state-of-the-art paradigms in machine learning, in fraud prediction rather than the commonly used logit regression. Third, we predict frauds out of sample rather than explain fraud determinants within sample.

While there are a lot of raw accounting data items from the financial statements, we limit our empirical analyses to only 24 raw data items from Cecchini et al. [2010] in order to compare the performance between our fraud detection models and the traditional fraud detection methods whose fraud determinants are also derived from the same or a similar set of raw accounting data.

We adopt two types of benchmark fraud prediction models. First, we use a logistic fraud prediction model based on 11 financial ratios from Dechow et al. [2011], referred to as the basic Dechow et al. model. While the raw accounting data items required for the 11 financial ratios substantially overlap with the 24 raw data items from Cecchini et al. [2010], the basic Dechow et al. model requires three raw accounting data items excluded from the 24 raw data items. On the other hand, five of the 24 raw data items are not used by the basic Dechow et al. model. Accordingly, we also estimate a modified Dechow et al. model based on 11 financial ratios from the basic model and three additional ratios derived from the 24 raw data items excluded from the basic model. As a result, the modified Dechow et al. model contains 14 financial ratios derived from 28 raw accounting data items, including the 24 raw

data items from Cecchini et al. [2010]. Our second type of benchmark model is the fraud detection model developed by Cecchini et al. based on support vector machines with a financial kernel (SVM-FK) that maps the 24 raw accounting data into a broader set of financial ratios and changes in financial ratios.

We find that the out-of-sample performance of the basic and modified Dechow et al. models is similar and significantly better than the performance of random guesses. Consistent with Cecchini et al. [2010], we find that the SVM-FK method outperforms both the basic and modified Dechow et al. models. However, a logistic regression based on the 24 raw accounting data items outperform both the basic and modified Dechow et al. models by a significant margin and perform similarly to Cecchini et al.'s more advanced SVM-FK model. This evidence suggests that the financial ratios identified by accounting experts and used in the basic and modified Dechow et al. methods have not fully utilized the useful information from the raw accounting data. In addition, we show that the ensemble method based on the same 24 raw accounting data further outperforms the logistic regression model based on the same 24 raw accounting data by a significant margin. This evidence suggests that conditioning on using the same set of raw accounting data items, our ensemble learning method is more powerful than the logistic regression method in extracting valuable information from the raw data items for the purposes of fraud prediction out of sample.

Overall, our results suggest that the existing fraud prediction models haven't fully utilized the information from the publicly available raw financial statement data. In addition, we show that it is possible to extract such useful information by adopting better fraud prediction models based on raw accounting data.

To benchmark with prior fraud detection research, we limit our analyses to 24 raw accounting data items, a small fraction of the raw accounting items from financial statements readily available from Compustat. Hence, there exists an exciting possibility that more

powerful fraud detection models can be developed by tapping this large volume of unused and low cost raw financial data items.

Our findings are also relevant to a growing accounting literature that attempts to harvest the text information from corporate filings for the purposes of predicting frauds or firm performance (e.g., Larcker and Zakolyukina [2012], Li [2010], Lo, Ramos, and Rogo [2015]). To demonstrate the usefulness of textual data, a common employed 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

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. 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. "The Detection of Earnings Manipulation." *Financial Analysts Journal* 55 (1999): 24-36.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. "Classification and regression trees." CRC press, 1984.

Cecchini, M., H. Aytug, G. J. Koehler, and P. Pathak. "Detecting Management Fraud in Public Companies." *Management Science* 56 (2010): 1146-1160.

Dechow, P. M., W., Ge, C. R. Larson, and R. G. Sloan. "Predicting Material Accounting Misstatements." *Contemporary Accounting Research* 28 (2011): 17-82.

Dyck, A., A. Morse, and L. Zingales. "Who Blows the Whistle on Corporate Fraud?" Working paper, *University of Michigan*, 2005.

Ernst & Young. "Driving ethical growth—New markets, new challenges." 11th Global Fraud Survey, 2010. Available online at http://www.ey.com/Publication/vwLUAssets/Driving_ethical_growth_new_markets,_new_challenges:_11th_Global_Fraud_Survey/$FILE/EY_11th_Global_Fraud_Survey.

Efron, B., and R. J. Tibshirani. "An introduction to the bootstrap." *CRC press*, 1994.

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): 113-143.

Fawcett, T. "An Introduction to Roc Analysis." *Pattern Recognition Letters* 27 (2006): 861-874.

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 - 139.

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): 463-484.

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): 915-945.

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." Morgan Kaufmann, 2006.

Hastie, T., R. Tibshirani, and J.H. Friedman. "The Elements of Statistical Learning." *New York: Springer*, 2003.

He, H., and Y. Ma. "Imbalanced Learning: Foundations, Algorithms, and Applications." *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*, forthcoming.

Järvelin K. and J. Kekäläinen. "Cumulated Gain-Based Evaluation of IR Techniques." *ACM Transactions on Information Systems*, 20 (2002): 422-446.

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* 41 (2011): 552-568.

Larcker, D. and A. A. Zakolyukina. "Detecting Deceptive Discussion in Conference Calls." *Journal of Accounting Research* 50 (2012): 495-540.

Li, F. "The informantion content of forward-looking statements in corporate filings - a naïve bayesian machine learning approach." *Journal of Accounting Research*, 48 (2010): 1049-1102.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. "Exploratory Undersampling for Class-Imbalance Learning." *IEEE Trans Syst Man Cybern B Cybern* 39 (2009): 539-550.

Liu, X.-Y., and Z.-H. Zhou. "Ensemble Methods for Class Imbalance Learning." *Imbalanced Learning: Foundations, Algorithms, and Applications*, (eds H. He and Y. Ma), John Wiley & Sons, Inc, 2013.

Lo, K., F. Ramos, R. Rogo. "Earnings management and annual report readability." Working paper, the University of British Columbia, 2015.

Makridakis, S. G., Wheelwright, S. C. and Hyndman, R. J. "Forecasting: Methods and Applications, 3rd ed." Wiley, New York, 1998.

Murphy, K. J. "Executive compensation." *Handbook of Labor Economics* 3 (1999): 2485-2563.

Schiesel, S. "Trying to Catch WorldCom's Mirage." *New York Times*, June 30, 2002.

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): 185-197.

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): 131-146.

Tuv, E., Borisov, A., Runger, G., and Torkkola, K. "Feature selection with ensembles, artificial variables, and redundancy elimination." *The Journal of Machine Learning Research* 10 (2009): 1341-1366.

Wang Y., Wang L., Li Y., He D., Chen W., Liu T.-Y. "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." *Morgan Kaufmann*, 2005.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. "Top 10 algorithms in data mining." *Knowledge and Information Systems*, (2008): 1-37.

Zhou, Z.-H. "Ensemble Methods: Foundations and Algorithms." *CRC Press*, 2012.

APPENDIX:

Variable Definitions

Financial Ratio Variable	Predicted sign	Calculation from raw accounting data items
Accruals quality related variables		
WC accruals	+	[(Δ*Current Assets, Total* – Δ*Cash and Short-Term Investments*) – (Δ*Current Liabilities, Total* – Δ*Debt in Current Liabilities, Total* – *ΔIncome Taxes Payable*)]/Average *Assets, Total*
RSST accruals	+	(ΔWC + ΔNCO + ΔFIN) / Average A*ssets, Total*, where WC = (*Current Assets, Total – Cash and Short-Term Investments*) – (*Current Liabilities, Total – Debt in Current Liabilities, Total*); NCO = (*Assets, Total – Current Assets, Total – Investment and Advances, Other*) – (*Liabilities, Total – Current Liabilities, Total – Long-term Debt, Total*); FIN=(*Short-Term Investments + Investment and Advances, Other*) – (*Long-term Debt, Total + Debt in Current Liabilities, Total + Preferred/Preference Stock (Capital), Total*)
Change in receivables	+	Δ*Receivables, Total* / Average *Assets, Total*
Change in inventory	+	Δ*Inventories, Total* /Average *Assets, Total*
% Soft assets	+	(*Assets, Total – Property, Plant and Equipment, Total – Cash and Short-Term Investments*) / *Assets, Total*
Performance variables		
Change in cash sales	–	Percentage change in cash sales, where cash sales is measured as *Sales/Turnover (Net) –* Δ*Receivables, Total*
Change in cash margin	–	Percentage change in cash margin, where cash margin is measured as 1 – [(*Cost of Goods Sold* – Δ*Inventories, Total* + Δ*Account Payable, Trade*) / (*Sales/Turnover (Net) – ΔReceivables, Total*)]
Change in return on assets	+	(*Income Before Extraordinary Items* $_{t}$ / Average *Assets, Total* $_{t}$) – (*Income Before Extraordinary Items* $_{t-1}$ /Average *Assets, Total* $_{t-1}$)
Change in free cash flows	–	Δ [*Income Before Extraordinary Items* – RSST Accruals] / Average *Assets, Total*
Market-related incentives		
Actual issuance	+	An indicator variable coded 1 if the firm issued securities during year t (i.e., an indicator variable coded 1 if *Sale of Common and Preferred Stock* > 0 or *Long-Term Debt Issuance* > 0)
Book-to-market	–	*Common/Ordinary Equity, Total* / (*Common Shares Outstanding* × *Price Close, Annual, Fiscal*)

APPENDIX *(Continued):*

Financial Ratio Variable	**Predicted sign**	**Calculation**
Additional financial ratios		
Depreciation Index	+	[*Depreciation and Amortization* $_{t-1}$ / *(Depreciation and Amortization* $_{t-1}$ + *Property, Plant and Equipment, Total* $_{t-1}$)]/ [*Depreciation and Amortization* $_{t}$ / *(Depreciation and Amortization* $_{t}$ + *Property, Plant and Equipment, Total* $_{t}$)]
Retain earnings over total assets	+	*Retained Earnings/ Assets, Total*
EBIT	+	(*Net Income (Loss) + Interest and Related Expense, Total + Income Taxes, Total*) / *Assets, Total*

TABLE 1

Distribution of Fraud Firms by Year over 1979-2008

Table 1 shows the yearly fraud percentages after merging AAERs announced in or prior to 2010 with COMPUSTAT fundamental data over the years 1979-2008. The bold numbers highlight that the fraud firm-years in the years 2006–2008 are abnormally low due to the fact that frauds committed in 2006-2008 have not been fully reported by AAERs as of 2010, the ending date of our AAER sample.

Year	Number of fraud firms	Total number of firms	Percentage
1979	4	3,232	0.12%
1980	9	3,212	0.28%
1981	9	3,428	0.26%
1982	16	3,297	0.49%
1983	8	3,250	0.25%
1984	10	3,109	0.32%
1985	6	2,990	0.20%
1986	12	2,976	0.40%
1987	11	3,246	0.34%
1988	18	4,246	0.42%
1989	19	4,372	0.43%
1990	15	4,388	0.34%
1991	26	4,554	0.57%
1992	24	4,826	0.50%
1993	22	5,238	0.42%
1994	20	5,542	0.36%
1995	18	6,153	0.29%
1996	29	6,662	0.44%
1997	38	6,684	0.57%
1998	52	6,613	0.79%
1999	72	6,813	1.06%
2000	80	6,774	1.18%
2001	68	6,442	1.06%
2002	65	6,155	1.06%
2003	52	5,982	0.87%
2004	39	5,928	0.66%
2005	27	5,806	0.47%
2006	**11**	**5,838**	**0.19%**
2007	**4**	**5,749**	**0.07%**
2008	**1**	**5,515**	**0.02%**
Total	**785**	**149,202**	**0.53%**

TABLE 2

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

Table 2 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].

	(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	-	-

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
Market Value Items			
Price Close, Annual, Fiscal	Yes	Yes	Yes
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	-	-

TABLE 3

The Regression Results of the Basic Dechow et al. Model and the Modified Dechow et al. Model in the Training Periods.

Panel A The basic Dechow et al. using 11 financial ratios.

Panel A shows the logit regression results of the training fraud detection model based on 11 financial ratios from Dechow et al. [2011]. Training period in column (1) is year 1991- 2001 and the corresponding test period is year 2003; in column (2), training and testing periods are year 1991- 2002 and year 2004, respectively; in column (3), training and test periods are 1991-2003 and year 2005, respectively.

Training Period Test Period	Predicted Sign	(1) 1991-2001 2003	(2) 1991-2002 2004	(3) 1991-2003 2005
Variables		coefficient (p-value)	coefficient (p-value)	coefficient (p-value)
Constant		-7.306*** (0.000)	-7.214*** (0.000)	-7.266*** (0.000)
WC accruals	+	-0.074 (0.812)	-0.113 (0.678)	-0.187 (0.429)
RSST accruals	+	0.237 (0.307)	0.235 (0.248)	0.273 (0.133)
Change in receivables	+	1.767*** (0.000)	1.764*** (0.000)	1.771*** (0.000)
Change in inventory	+	1.292** (0.012)	1.076** (0.025)	1.054** (0.019)
% Soft assets	+	1.906*** (0.000)	1.900*** (0.000)	1.913*** (0.000)
Change in cash sales	–	0.011 (0.571)	0.003 (0.901)	-0.001 (0.964)
Change in cash margin	–	-0.011 (0.480)	-0.016 (0.260)	-0.016 (0.263)
Change in return on assets	+	-0.227 (0.185)	-0.230 (0.152)	-0.196 (0.161)
Change in free cash flows	–	-0.395** (0.011)	-0.330** (0.015)	-0.249** (0.032)
Actual issuance	+	1.328*** (0.000)	1.294*** (0.000)	1.370*** (0.000)
Book-to-market	–	-0.049 (0.118)	-0.021 (0.479)	-0.011 (0.716)
N		65082	71469	77586
Pseudo R-sq		4.2%	3.9%	3.8%

Robust p-value clustered by firm in parentheses. P-values are based on two-tailed tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

TABLE 3 *(Continued)*

Panel B The modified Dechow et al. model using 14 financial ratios.

Panel B shows logit regression results of the modified Dechow et al. model after adding three financial ratios to the basic Dechow et al. model.

		(1)	(2)	(3)
Training Period	Predicted	1991-2001	1991-2002	1991-2003
Test Period	Sign	2003	2004	2005
Variables		coefficient (p-value)	coefficient (p-value)	coefficient (p-value)
Constant		-7.503***	-7.340***	-7.355***
		(0.000)	(0.000)	(0.000)
WC accruals	+	-0.144	-0.202	-0.285
		(0.769)	(0.625)	(0.438)
RSST accruals	+	-0.017	0.340	0.499
		(0.966)	(0.400)	(0.176)
Change in receivables	+	1.736***	1.795***	1.796***
		(0.002)	(0.000)	(0.000)
Change in inventory	+	0.614	0.382	0.391
		(0.322)	(0.497)	(0.455)
% Soft assets	+	2.183***	2.138***	2.114***
		(0.000)	(0.000)	(0.000)
Change in cash sales	–	0.025	0.012	0.007
		(0.259)	(0.599)	(0.738)
Change in cash margin	–	-0.013	-0.023	-0.023
		(0.470)	(0.188)	(0.192)
Change in return on assets	+	-0.234	-0.285	-0.262
		(0.383)	(0.258)	(0.251)
Change in free cash flows	–	-0.526*	-0.161	0.040
		(0.073)	(0.574)	(0.876)
Actual issuance	+	1.190***	1.139***	1.212***
		(0.000)	(0.000)	(0.000)
Book-to-market	–	-0.107**	-0.075*	-0.069*
		(0.023)	(0.070)	(0.071)
Depreciation Index	+	0.167**	-7.340***	-7.355***
		(0.050)	(0.000)	(0.000)
Retain earnings over total assets	+	0.176	-0.202	-0.285
		(0.293)	(0.625)	(0.438)
EBIT	+	-0.076	0.340	0.499
		(0.847)	(0.400)	(0.176)
N		59492	65104	70542
Pseudo R-sq		4.6%	4.3%	4.3%

Robust p-value clustered by firm in parentheses. See the appendix for variable definitions. P-values are based on two-tailed tests. *** p < 0.01, ** p < 0.05, * p < 0.1

TABLE 4

The Out-of-Sample Performance Evaluation Metrics

Panel A. The results of the out-of-sample performance evaluation over the test period 2003-2005

Panel A shows fraud detection models' performance comparison using the following performance metrics averaged over the test period 2003–2005:

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).* The Discounted Cumulative Gain at the position k (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, i.e., $\text{NDCG@k} = \frac{\text{DCG@k}}{\text{ideal DCG@k}}$, 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, $Sensitivity = \frac{\text{TP}}{\text{TP+FN}}$, 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 non-frauds 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, $Precision = \frac{\text{TP}}{\text{TP+FP}}$, 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.

			Performance Metrics averaged over the test period 2003-2005			
			Metric one	**Metric two**		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
11 Financial Ratios	1)	Logit	0.638	0.018	2.72%	1.68%
14 Financial Ratios	2)	Logit	0.654	0.014	2.42%	1.30%
24 Raw Accounting Data Items	3)	Logit	0.747	0.015	2.14%	1.69%
	4)	SVM-FK	0.740	0.017	1.96%	1.85%
	5)	RUSBoost(1:1)	0.822	0.271	22.15%	14.64%
	6)	RUSBoost(2:1)	0.816	0.329	28.09%	18.59%

TABLE 4 *(Continued)*

Panel B Paired t-test of AUC across the fraud detection models over the test period 2003-2005

Panel B shows the p-values of the paired t-tests of AUC values among fraud detection models listed in Table 4 Panel A.

Input Variables		Method	1)	2)	3)	4)	5)
11 Financial Ratios	1)	Logit					
14 Financial Ratios	2)	Logit	**0.080***				
24 Raw Accounting Data Items	3)	Logit	**0.057***	**0.074***			
	4)	SVM-FK	**0.015****	**0.023****	0.671		
	5)	RUSBoost(1:1)	**0.022****	**0.019****	**0.074***	**0.069***	
	6)	RUSBoost(2:1)	**0.033****	**0.032****	**0.078***	0.101	0.513

P-values are based on two-tailed tests. *** p < 0.01, ** p < 0.05, * p < 0.1

TABLE 4 *(Continued)*

Panel C Paired t-tests of NDCG@k across fraud detection models over the test period 2003-2005

Panel C shows the p-Values of paired t-tests of NDCG@k values among fraud detection models listed in Table 4 Panel A.

Input Variables		Method	1)	2)	3)	4)	5)
11 Financial Ratios	1)	Logit					
14 Financial Ratios	2)	Logit	0.412				
24 Raw Accounting Data Items	3)	Logit	0.808	0.947			
	4)	SVM-FK	0.974	0.898	0.855		
	5)	RUSBoost(1:1)	**0.002*** **	**0.004*** **	**0.000*** **	**0.001*** **	
	6)	RUSBoost(2:1)	**0.001*** **	**0.001*** **	**0.001*** **	**0.005*** **	**0.055***

P-values are based on two-tailed tests. *** p < 0.01, ** p < 0.05, * p < 0.1

TABLE 5

Results of the Logit Regression on 24 Raw Accounting Variables from Cecchini et al. [2010]

Table 5 shows the logistic regression results of fraud prediction using Cecchini et al.'s [2010] 24 raw accounting data items. To minimize the effects of scale differences across different raw data items, we normalize an observation's input vector such that the normalized vector has a unit length, i.e., $x' = \frac{x}{||x||}$, where the division are element-wise. For example, a vector of (1,2) is normalized to $(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$.

	(1)	(2)	(3)
Training Sample	1991-2001	1991-2002	1991-2003
Test Sample	2003	2004	2005
Variables	coefficient (p-value)	coefficient (p-value)	coefficient (p-value)
Constant	-7.350***	-7.359***	-7.422***
	(0.000)	(0.000)	(0.000)
Balance Sheet Items			
Cash and Short-Term Investments	-1.985*	-1.996*	-1.904*
	(0.073)	(0.071)	(0.093)
Receivables, Total	-1.230	-1.633	-1.703
	(0.390)	(0.250)	(0.230)
Inventories, Total	0.827	0.361	0.182
	(0.507)	(0.767)	(0.881)
Short-Term Investments, Total	-1.629	-1.616	-1.534
	(0.200)	(0.183)	(0.227)
Property, Plant and Equipment, Total	-4.025***	-3.795***	-3.635***
	(0.000)	(0.000)	(0.000)
Investment and Advances, Other	0.061	0.252	0.522
	(0.950)	(0.777)	(0.559)
Assets, Total	5.630**	5.525**	4.663*
	(0.045)	(0.039)	(0.076)
Debt in Current Liabilities, Total	-2.667	-2.689	-3.526*
	(0.187)	(0.174)	(0.077)
Income Taxes Payable	5.284	5.824	4.175
	(0.393)	(0.387)	(0.570)
Current Liabilities, Total	2.011	2.050	2.175
	(0.254)	(0.238)	(0.215)
Long-Term Debt, Total	1.053	1.051	0.951
	(0.410)	(0.394)	(0.452)
Liabilities, Total	1.150	1.453	2.478
	(0.679)	(0.585)	(0.346)
Common/Ordinary Equity, Total	-0.678	-0.347	0.616
	(0.812)	(0.899)	(0.819)
Preferred/Preference Stock (Capital), Total	-1.941	-1.865	-1.016
	(0.650)	(0.642)	(0.793)
Retained Earnings	0.519	0.268	0.092
	(0.222)	(0.500)	(0.812)

TABLE 5 *(Continued)*

	(1)	(2)	(3)
Training Sample	1991-2001	1991-2002	1991-2003
Test Sample	2003	2004	2005
Variables	coefficient (p-value)	coefficient (p-value)	coefficient (p-value)
Income Statement Items			
Sales/Turnover (Net)	-0.017	0.316	0.309
	(0.980)	(0.609)	(0.611)
Cost of Goods Sold	1.518*	1.304*	1.460*
	(0.070)	(0.092)	(0.055)
Depreciation and Amortization	1.664	-0.426	-1.366
	(0.707)	(0.923)	(0.761)
Interest and Related Expense, Total	-21.195**	-23.771***	-22.604**
	(0.021)	(0.009)	(0.011)
Income Taxes, Total	4.373	4.544	4.465
	(0.185)	(0.150)	(0.139)
Income Before Extraordinary Items	-1.935	-1.467	-0.889
	(0.125)	(0.302)	(0.545)
Net Income (Loss)	-0.045	-0.946	-1.387
	(0.963)	(0.431)	(0.260)
Market Value Items			
Common Shares Outstanding	0.803	0.445	0.409
	(0.188)	(0.469)	(0.494)
Price Close, Annual, Fiscal	0.680	0.518	0.375
	(0.398)	(0.529)	(0.653)
N	66301	72456	78438
Pseudo R-sq	6.1%	6.3%	6.4%

Robust p-value clustered by firm in parentheses. P-values are based on two-tailed tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 6. The importance of the 24 raw accounting data items used in the RUSBoost (2:1) model in fraud prediction

Panel A shows the feature importance of the 24 raw data items used in the RUSBoost (2:1) model. The reported values are the estimates of feature importance computed using the "predictorImportance" function in MATLAB (multiplied by 100). Panel B shows the ten specific financial statement accounts affected by the misstatements identified by the AAERs over the period 2003-2005. There are 118 misstatement firm years in 2003-2005 involving 141 specific accounts. Note a single misstatement firm year could affect more than one specific account. The data used in Panel A is hand collected from the AAERs by Dechow et al. (2011).

Panel A. The importance of the 24 raw data items in the RUSBoost (2:1) model

Rank	24 Raw data items	Test year 2003	Test year 2004	Test year 2005	Average of 2003-2005	Related account categories from Panel B
1	*Cash and Short-Term Investments*	0.010	0.009	0.008	0.009	*asset*
2	*Assets, Total*	0.010	0.009	0.008	0.009	*asset*
3	*Common Shares Outstanding*	0.010	0.009	0.008	0.009	
4	*Price Close, Annual, Fiscal*	0.009	0.008	0.008	0.008	
5	*Property, Plant and Equipment, Total*	0.009	0.008	0.007	0.008	*asset*
6	*Depreciation and Amortization*	0.009	0.008	0.007	0.008	
7	*Inventories, Total*	0.009	0.008	0.007	0.008	*Inv*
8	*Cost of Goods Sold*	0.009	0.008	0.007	0.008	*cogs*
9	*Receivables, Total*	0.009	0.008	0.007	0.008	*rec*
10	*Debt in Current Liabilities, Total*	0.009	0.008	0.007	0.008	*liab*
11	*Common/Ordinary Equity, Total*	0.008	0.008	0.007	0.008	*res*
12	*Retained Earnings*	0.008	0.007	0.007	0.008	*res*

13	*Current Liabilities, Total*	0.008	0.007	0.007	0.007	*liab*
14	*Interest and Related Expense, Total*	0.007	0.006	0.006	0.006	
15	*Long-Term Debt, Total*	0.007	0.006	0.006	0.006	*liab*
16	*Income Before Extraordinary Items*	0.007	0.006	0.005	0.006	
17	*Sales/Turnover (Net)*	0.007	0.006	0.005	0.006	*rev*
18	*Income Taxes, Total*	0.006	0.006	0.005	0.006	
19	*Liabilities, Total*	0.006	0.005	0.005	0.006	*liab*
20	*Investment and Advances, Other*	0.005	0.004	0.004	0.004	*asset*
21	*Income Taxes Payable*	0.005	0.004	0.004	0.004	*liab*
22	*Net Income (Loss)*	0.003	0.003	0.003	0.003	
23	*Short-Term Investments, Total*	0.003	0.003	0.002	0.003	*mkt_sec*
24	*Preferred/Preference Stock (Capital), Total*	0.002	0.002	0.001	0.002	

Panel B. The specific financial statement accounts affected by the detected accounting frauds over the test period 2003-2005

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	*rev*	Equals 1 if misstatement affected revenues, 0 otherwise	48	
2	*rec*	Equals 1 if misstatement affected accounts receivable, 0 otherwise	25	*Receivables, Total*
3	*res*	Equals 1 if misstatement affected reserves accounts, 0 otherwise. Dechow et al. (2011)	21	
4	*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	15	*Cash and Short-Term Investments; Assets, Total; Property, Plant and Equipment, Total*
5	*inv*	Equals 1 if misstatement affected inventory, 0 otherwise	13	*Inventories, Total*
6	*liab*	Equals 1 if misstatement affected liabilities, 0 otherwise	7	*Debt in Current Liabilities, Total*
7	*pay*	Equals 1 if misstatement affected accounts payable, 0 otherwise	6	
8	*cogs*	Equals 1 if misstatement affected cost of goods sold, 0 otherwise	6	*Cost of Goods Sold*
9	*mkt_sec*	Equals 1 if misstatement affected marketable securities, 0 otherwise	0	
10	*debt*	Equals 1 if misstatement affected bad debts, 0 otherwise	0	
Total			141	

FIGURE 1

A Graphical Representation of the Information in Panel A of Table 6