The prevalence and price distorting effects of undetected financial misrepresentation: Empirical evidence *

Abdullah Alawadhi
University of Washington
a90@uw.edu

Jonathan Karpoff
University of Washington
karpoff@uw.edu

Jennifer L. Koski
University of Washington
jkoski@uw.edu

Gerald D. Martin
American University
gmartin@american.edu

Under revision: April 23, 2023

Abstract. We use a comprehensive database of regulatory enforcement actions for financial misrepresentation to estimate prediction models using logistic, machine learning, and bivariate probit classifiers. Our parsimonious logistic model and three versions of a Support Vector Machine learning model perform well, each with an average area under the ROC curve (AUC) of 0.78 in out of sample tests. The base logistic model implies that 22.3% of Compustat-listed firms are engaged in financial misrepresentation that is potentially sanctionable by regulators in an average year. The average violation period is 3.1 years, implying that 22.3%/3.1 = 7.2% of firms initiate financial reporting practices each year that are potentially sanctionable. Of these firms, 3.5% eventually are sanctioned by regulators. We use these findings to infer the fraction of firms that misrepresent their financials and yet never face regulatory penalties, to estimate the size of the price distortions imposed by misrepresentation on the shares of both misrepresenting and non-misrepresenting firms, and to estimate the size of firms' ex ante expected costs of engaging in financial misrepresentation that incorporate both the probability of getting caught and the penalties if caught.

JEL classifications: G38, K22, L51, M48
Keywords: Fraud, prediction, enforcement, social cost

*We thank Hank Bessembinder, Bidisha Chakrabarty, Weili Ge, Saad Siddiqui, and participants at the PCAOB/TAR Conference on Auditing and Capital Markets, and the University of Washington finance brown bag seminar for helpful comments.

The prevalence and price distorting effects of undetected financial misrepresentation: Empirical evidence

1. Introduction

Financial misrepresentation imposes large costs on firms, investors, financial markets, and national economies.[1] But how common is it? Table 1 shows that, on average, 0.27% of Compustat-listed firms initiated programs of financial misrepresentation each year from 1976–2014 that subsequently triggered enforcement action by the Securities and Exchange Commission (SEC) and/or the Department of Justice (DOJ). These firms' misreporting periods average 3.1 years in duration, indicating that, in an average year, 0.84% of Compustat-listed firms are engaged in financial misrepresentation that subsequently prompts regulatory enforcement action. These firms that are caught, however, are just the tip of the iceberg. How many other firms engage in similar types of misconduct yet remain undetected, and what is the probability misrepresenting firms are caught and face sanctions? This paper seeks to answer these questions.

Our approach is to estimate multiple versions of prediction models based on logistic regression (e.g., Beneish 1999; Dechow, Ge, Larson, and Sloan 2011), machine learning approaches (e.g., Cecchini, Aytug, Koehler, and Pathak 2010; Bao, Ke, Li, Yu, and Zhang 2019), and bivariate probit models (Poirier 1980; Wang 2013). We use data from all SEC and DOJ enforcement actions from 1978-2017 for financial misrepresentation at publicly traded companies that occurred from 1976-2014 and incorporate predictor variables from prior attempts to predict unobserved misconduct. We then apply Receiver Operating Characteristics (ROC) methodology to measure each model's sensitivity (i.e., the ability to accurately identify firms that are caught by regulators for financial misconduct) and specificity (i.e., the ability to avoid false positives). To assess each model's performance, we calculate the area under the ROC curve (AUC) in both in-sample and out-of-sample tests.[2]

[1] For a survey of papers that measure the cost to firms when they are caught for financial misconduct, see Amiram, Bozanic, Cox, Dupont, Karpoff, and Sloan (2018), Section 4.2. For evidence on the effects of financial fraud on investors, markets, and economies, see Graham, Li, and Qiu (2008), Beatty, Liao, and Yu (2013), Giannetti and Wang (2017), Gurun, Stoffman, and Yonker (2018), and Dupont (2023).

[2] At the optimum threshold level that maximizes the model's overall predictive ability, the AUC is the sum of the model's sensitivity and specificity.

This approach offers four advantages over previous attempts to estimate the extent of unobserved financial misconduct. First, we examine a wide range of candidate models as opposed to only one or two models, including logistic models and machine learning approaches based on Nearest Neighbor, Decision Trees, Ensemble Methods, Neural Networks, Discriminant Analysis, and Support Vector Machines. Second, to identify financial misrepresentation we use a comprehensive sample of regulatory enforcement actions for financial misrepresentation. This is in contrast to the limited samples used in previous prediction models based on accounting restatements, class action lawsuits, and Accounting and Auditing Enforcement Releases (AAERs) to identify financial misrepresentation. As Karpoff, Koester, Lee, and Martin (2017) show, such samples are subject to data constraints that can severely impact empirical inferences. Third, we consider a large set of candidate predictor variables that includes but is not limited to the characteristics used in the most prominent prediction models in the literature. Fourth, we employ ROC methods to explicitly consider the tradeoffs between model sensitivity and specificity. Several recent papers also use the area-under-the-curve (AUC) metric to evaluate fraud prediction models (e.g. Cecchini et al. (2010), Larcker and Zakolyukina (2012), Perols, Bowen, Zimmerman, and Samba (2017), and Bao, Ke, Li, Yu, and Zhang (2019)). We show how the inferences derived from a prediction model are affected by adjusting the error cost ratio, which summarizes the relative costs of false positives and false negatives. Adjustments to the error cost ratio, in turn, reflect the researcher's underlying prior beliefs about regulators' objectives or the underlying prevalence of misconduct. We use this insight to incorporate a wide range of prior beliefs simply by adjusting the error cost ratio.

Among the models we estimate, one logistic model and three support vector machine (SVM) learning models perform the best as judged by their AUCs. Each of these four models has an out-of-sample AUC of 0.78, indicating an ability to discriminate reasonably well as judged by standards in the ROC literature (e.g., Swets 1998). The top four models yield similar estimates of the proportion of firms engaged in financial misconduct, including undetected firms. Of these models, however, the logistic model has two advantages. First, the estimated coefficients from the logistic model yield insight into the characteristics of firms that engage in financial misrepresentation and their average marginal effects. For example, our base logistic model implies that that a 10% increase in market capitalization is associated with a 4.9% increase in the probability of misconduct, and firms reporting a loss have a 0.52% higher probability of misconduct in that year. The SVM models, in contrast, generate predictions based on combinations of the underlying covariates that offer no meaningful intuition about the characteristics of firms that are engaged in financial misconduct. The second advantage of the logistic model is that it is computationally simple and requires less computer processing time compared to most machine learning approaches. We therefore emphasize the logistic model when making inferences about the likelihood, prevalence, and costs of financial misconduct.

Using the logistic model results, our base estimate is that 22.3% of Compustat-listed firms are engaged in reporting activities that potentially are sanctionable as financial misrepresentation in any given year. Misrepresentation periods average 3.1 years, implying that, in an average year, 22.3% ÷ 3.1 = 7.2% of Compustat-listed firms initiate new reporting activities that potentially are sanctionable as financial misrepresentation. Of these firms, 3.5% eventually are sanctioned in enforcement actions for financial misrepresentation. These estimates are similar to those from the other three top-performing SVM machine learning classifiers, which yield estimates of the probability of potential violation from 22% to 27% and estimates of the probability of attracting enforcement activity from 3.1% to 3.6%.

All estimates from prediction models such as these depend on an assumption about the duration of the unobserved violations. The average in-sample observed violation period is 3.1 years, so our base estimates assume that unobserved violations last three years. One contribution of our approach is to show that, by making this assumption explicit, we can examine the effects of changing it. For example, if we assume unobserved violations last only one year, the model flags 28.5% of firms as committing misconduct each year and the estimated conditional probability of getting caught falls to 3.2%.[3]

[3] We are aware of no other paper that considers the effect of violation duration on one's inferences from a prediction framework. The most common (implicit) assumption is that violations persist for one year (e.g., Dechow et al. (2011), Bao et al., 2019). As illustrated in Section 7.1, however, one's implicit assumption about the undetected violations' durations can have a meaningful effect on inferences.

The base estimates also reflect an assumption that Type I errors (falsely classifying an innocent firm as misrepresenting) and Type II errors (inaccurately classifying a misrepresenting firm as innocent) are equally costly, i.e., that the error cost ratio equals 1.0. For each prediction model we estimate, we show how the probability estimates change with alternate assumptions about the pre-specified relative costs of Type I and Type II errors. For example, U.S. jurisprudence typically assigns a higher cost to Type I errors (falsely charging a firm with misrepresentation) compared to Type II errors (letting actual misrepresentation go unpunished). If regulators act as if the cost of a Type I error is 50 percent higher than the cost of a Type II error, our base model implies that an average of 13.5% of Compustat-listed firms are engaged in misrepresentation in any given year (compared to the base estimate of 22.3%), of which 4.6% eventually face enforcement action (compared to the base estimate of 3.5%). These estimates illustrate how a modeler's prior beliefs about the error cost ratio affect one's inferences from a prediction model.

We report on out-of-sample tests to assess the validity of our estimates, including k-fold cross validation and rolling regressions with different model estimation and holdout periods to assess the model's validity. Throughout these tests, our base logistic model yields stable estimates and predictions across time. In k-fold cross validation tests, the average out-of-sample AUC equals 78%, and in year-by-year rolling regressions the out-of-sample AUC varies within a narrow range of 73% to 77%.

Next, we combine measures of the prevalence of misconduct with prior research findings to gain insight into the nature of the price distortions caused by financial misrepresentation. Our base estimates imply that, in any given year, 22.3% of all firms' share prices are artificially inflated by 10.3% over their full information values. Even though most of these firms never face an enforcement action for their misconduct, investors likely will consider the fact that some firms are engaged in misrepresentation that is undetected. To avoid paying for overpriced shares on average, investors will rationally discount the values of non-misrepresenting firms by 3.0%. Both types of share price distortions – artificial inflation of 10.3% by misrepresenting firms and discounts of 3.0% for other firms – impose social costs, as they lead to suboptimal investment and risk-bearing compared to the hypothetical counterfactual in which no financial misrepresentation occurs (e.g., see Bond, Edmans, and Goldstein, 2012). To our knowledge, these are the first empirical measures of the pervasiveness and size of the price distortions caused by financial misrepresentation.

Finally, we combine our results with prior findings to examine the implications for managers and regulators. Our base estimates imply that a firm misrepresenting its financial statements faces a 3.5% chance of being subject to penalties that average 31.7% of market capitalization. A firm's unconditional expected penalty for engaging in financial misrepresentation therefore equals 1.1% of the firm's market capitalization – an amount that, for most firms, justifies a substantial investment in regulatory compliance and internal controls. For regulators, these estimates imply that shareholders in an average firm can benefit financially from misrepresentation it if generates benefits that are expected to exceed 1.1% of the firm's market capitalization. That is, shareholders can expect to benefit from a firm's financial misrepresentation only in limited situations in which the expected benefits to the firm are very large. One such situation could be shortly before the firm issues new equity to outside investors (e.g., see Dechow, Sloan, and Sweeney 1996).

2. Alternative approaches to estimate the prevalence of financial misrepresentation

The first task of this paper – to measure the rate of undetected financial misconduct – is related to the fraud prediction literature. We consider three types of models from this literature. The first type uses binary dependent variable regressions, typically logistic models, to identify the characteristics of firms that are known to be violators. Other firms with similar characteristics are flagged as also being likely violators. This approach has wide applications, including predicting financial distress (Altman 1968) and consumer credit card fraud (Wiginton 1980). As notable examples, Beasley (1996) and Beneish (1997, 1999) construct logistic models that characterize the features of companies that are identified in AAERs, and Beneish (1999) uses the fitted values from his model to distinguish between "manipulators" and "non-manipulators" among all Compustat-listed firms. Dechow et al. (2011) extend this approach to a sample of 435 firms identified in AAERs and use fitted values from their model to generate an "F-score" for each firm-year. The F-score is the model-generated probability of misconduct for a given firm-year divided by the unconditional expectation of misconduct in the sample. Firm-years with F-scores above 1.0 are classified as having fraud. In Section 4, we extend this approach to generate a prediction model that generates inferences about the prevalence of financial misrepresentation using the modeler's prior beliefs about prevalence and Type I and Type II classification errors.

The second type of model uses machine learning classifiers to identify undetected financial misconduct.[4] Cecchini et al. (2010) use a support vector machine (SVM) classifier to identify a nonlinear combination of 23 firm characteristics that helps to identify firms engaged in misconduct. Perols (2011) finds that a support vector machine and logistic model approach both perform well as prediction models compared to four other machine learning models they consider. Bao et al. (2019) construct a model based on ensemble learning, which in their sample outperforms the Dechow et al. (2011) and Cecchini et al. (2010) models. In Section 5, we construct prediction models using 14 different machine learning classifiers, including the classifiers considered by Cecchini et al. (2010), Perols (2011), Perols et al. (2017), and Bao et al. (2019), and use the top three performing models to generate measures of the prevalence of financial misrepresentation.

A third approach is the bivariate probit model initially proposed by Poirier (1980). Wang (2013) applies this approach to financial misconduct to identify firm characteristics that are associated with committing misconduct and characteristics that are associated with facing a security class action lawsuit for its misconduct. In Section 6, we extend Wang's (2013) model to generate specific estimates of the probability of financial misconduct and the probability a firm is caught for its misconduct. We do not emphasize these estimates in making inferences about the prevalence of financial misrepresentation, however, for two reasons. First, using the AUC metric, the bivariate probit model does not perform as well as the leading logistic and machine learning models. Second, in our data the bivariate probit model is unstable and yields inferences that differ widely depending on small changes in specification, sometimes not converging at all.

[4] See Cecchini et al. (2010), Perols (2011), Perols et al. (2017), and Bao et al. (2019). Of these, only Bao et al. (2019) use their model to classify firm-years into misconduct versus non-misconduct firm-years.

Several papers use Benford's law to detect financial misrepresentation.[5] Benford's law states that the leading digits of a set of randomly-generated numbers should follow a known distribution. For example, the number one (1) should appear as a leading digit in 30.1% of the numbers, while the number nine (9) should appear as a leading digit in 4.6% of the numbers. Financial statement numbers that deviate from the Benford distribution can indicate that the numbers were generated in a non-arbitary manner, consistent with financial manipulation. Amiram et al. (2015) and Chakrabarty, Moulton, Pugachev, and Wang (2019) find that a measure of financial statement numbers' deviation from the Benford distribution is related to the incidence of a misstatement, so we incorporate a Benford's law measure as a candidate predictor variable in our tests.

Several additional papers use innovative strategies to gain insights that are related to fraud prediction. For example, Dyck, Morse, and Zingales (2023) use the demise of Arthur Andersen following Enron's 2001 financial fraud to identify otherwise-undetected fraud by former Arthur Andersen clients that were forced to switch to other auditors. In our sample, however, this approach yields improbable results (i.e., detection probabilities exceeding 90% or even 100%) because the fraction of former Arthur Anderson clients that eventually were detected for misconduct is similar to that for clients of other accounting firms. Zakolyukina (2018) constructs a structural model of a CEO's decision to manipulate earnings, which, when calibrated, implies that 60% of CEOs misstate earnings at least once in their careers. Soltes (2019) uses internal records from several large corporations and reports an average of more than two instances of some type of financial misconduct per firm each week. Bedi, Schrand, and Soltes (2019) survey managers' responses to structured scenarios, and find that nearly 60% of surveyed managers say they would likely misreport their financials in at least one of the experimental scenarios presented to them. These approaches yield novel insights, but they do not constitute prediction models and do not identify specific firms or firm-years as subject to financial misrepresentation.

[5] See Carslaw (1988), Thomas (1989), Durtschi, Hillison, and Pacini (2004), and Nigrini (2012), Amiram et al. (2015), and Chakrabarty, Moulton, Pugachev, and Wang (2019).

In another innovative approach, Larcker and Zakolyukina (2013) construct a prediction model using CEOs' and CFOs' language in quarterly earnings conference calls. This approach has an unusual advantage in that the prediction model's classifications can be compared against observed indicators of detected misconduct. Whereas Larcker and Zakolyukina (2013) model predictions of misconduct that eventually is detected, our focus is to measure the extent of both detected and undetected misrepresentation.

3. Data description

3.1. Financial misrepresentation sample

To develop prediction models, we focus on financial misrepresentation that triggers enforcement action by the SEC and/or DOJ for violation of one or more of the 13(b) provisions of the Securities Exchange Act of 1934 as amended by the Foreign Corrupt Practices Act (FCPA) of 1977 or the related Code of Regulations (CFR) rules. The section 13(b) provisions are cited whenever regulators bring enforcement action for inaccurate books and records (13(b)(2)(A)), inadequate internal controls (13(b)(2)(B)), knowingly falsifying books and records, or circumventing or knowingly failing to implement a system of internal controls (13(b)(5)). The Code of Regulations (CFR) rules promulgated under these sections include 13b2-1 (that no person shall directly or indirectly, falsify or cause to be falsified, any book, record or account) and 13b2-2 (that no person shall make or cause to be made a materially false or misleading statement to an accountant in connection with an audit, preparation of financial statements, or any required filing). As documented by Karpoff et al. (2017), 13(b) charges frequently are accompanied by other charges, including fraud under Section 10(b)-5 of the 1934 Exchange Act. However, we know of no instances of a regulatory action for financial misrepresentation at a publicly traded company that does not invoke section 13(b) or rules thereunder.

Using this screen, we identify 1,069 enforcement actions initiated by the SEC and DOJ from 1978 through 2017 for violations that occurred during the 1976–2014 period. Our investigation focuses on the period in which the firm's books were misrepresented. Information on the specific violation period for each case in our sample is obtained from details in the documents pertaining to the enforcement action, including

SEC enforcement releases. The median time from the end of the violation period until the SEC or DOJ's first enforcement proceeding – an event that is required for the firm to appear in our sample – is 28.9 months. Our data on SEC enforcement activity is through 2017. To avoid truncation bias due to the lag between a firm's violation period and when it is targeted for enforcement action, we calibrate and evaluate the prediction models using data on misrepresentation that occurs only through 2014.

Table 1 reports the annual number of newly initiated violations and ongoing violations in the sample. From 1978-2017 there are a total of 1,069 enforcement actions involving 1,039 distinct firms (including 30 recidivist firms) for financial misrepresentation covering a total of 3,311 violation-years during the 1976-2014 period. The average duration of a firm's violation, i.e., the period during which the firm's financials are misrepresented, is 3.1 years. On average, 0.84% of Compustat-listed firms are engaged in financial misrepresentation that subsequently attracts SEC enforcement action in any given year. This fraction reaches a maximum of 2.16% in 2001 during the dot-com bubble, decreasing to 0.40% in 2013 and 0.17% in 2014. The low fraction for 2014 could be due to truncation bias, as some violations may not have yet prompted SEC enforcement action by the end of our sample of enforcements in 2017. We find, however, that none of the qualitative results from our tests are affected if we decrease the possibility of truncation further by dropping 2014 violations (or violations from successively earlier years) from our sample.[6]

Table 3 reports on the distribution of misrepresentation firm-years by size. Consistent with prior findings (e.g., Dechow et al., 2011), the frequency of violation-years is positively correlated with firm size. For example, only 2.5% of firms are in the smallest size decile during violation firm-years, while 33.5% are in the largest decile. Table 4 shows that there also is some industry clustering in the sample. For example, firms in the *Business Equipment* industry constitute 23.9% of the violation firm-years in the sample even though only 14.9% of the population of firm-years in Compustat are in that industry. The

[6] As reported in Section 5.3, we obtain similar estimates in rolling regressions that subsequently add one year of new observations to the sample, building to the complete sample of violations through 2014. These results show that the base case model and resulting estimates are stable over time.

Healthcare industry also is overrepresented in the violation sample, whereas other industries, such as the *Utilities* and *Telecom* industries, are underrepresented.

3.2. Potential predictor variables

To identify potential predictor variables, we begin by including all the variables used by Beneish (1999), Dechow et al.'s (2011) Model 2, the *MAD score* based on Benford's Law as developed by Amiram et al. (2015), and the top five predictor variables emphasized by Cecchini et al. (2010). These variables – 23 in all – are summarized in the Appendix. To this list we add 32 additional candidate variables that are suggested by prior research or regulatory enforcement releases. These 32 additional variables fall into the following six groups:

(i) Common financial ratios and firm-level controls (9 additional variables)

Analysts emphasize the importance of several basic firm characteristics for financial analysis, including firm size (e.g., market capitalization), the market-to-book ratio, leverage, profit margin, return on assets (ROA), and basic earnings power (e.g., see Buffett and Clark, 2008). Several additional financial ratios can reflect the nature and opacity of the firm's assets and serve as an indicator of fraud. We include the ratio of research and development (R&D) expenses to firm sales, the ratio of intangible assets to total assets, and inventory turnover.[7]

The literature cited in footnote 7 suggests that the likelihood of misrepresentation is positively related to market capitalization, leverage, the ratio of R&D expenses to firm sales, the ratio of intangible assets to total assets, and inventory turnover; and negatively related to profit margin, ROA, and basic earnings power. This is because previous research shows that firms' internal control problems tend to

[7] For arguments supporting the potential predictive ability of these variables, see Association of Certified Fraud Examiners (2018), Bell and Carcello (2000), Kedia and Philippon (2009), Lou and Wang (2009), Ozcan (2016), Persons (2011), and the Public Company Accounting Oversight Board (2020).

increase with firm size, the opacity of firm assets, and poor performance. To the extent the market-to-book ratio is correlated with opacity, it too should be negatively related to the likelihood of misrepresentation.

(ii) Financial distress (3 additional variables)

Prior research finds that a disproportionate number of financial frauds concentrate among financially troubled firms, including several papers that conclude that Altman's z-score is negatively related to fraud.[8] We therefore consider three measures related specifically to financial distress, including *Altman's z-score*, a *Distress flag* that equals 1 if the Altman z-score is less than 1.75, and a *Loss flag* that equals 1 if the firm's net income is negative.

(iii) Alternate accruals measure (1 additional variable)

Accruals are widely considered to create opportunities to manage and/or misrepresent earnings. Dechow et al. (2011) include the accruals measure developed by Richardson et al. (2006), *RSST accruals*. We consider several additional accruals measures (e.g., see Larson, Sloan, and Giedt, 2018) include the accrual measure developed by Allen, Larsen, and Sloan (2013, *ALS accruals*) because, among the accruals measures, it is most highly correlated with financial misrepresentation in our sample. We expect a positive relation between accruals – as measured using either *RSST accruals* or *ALS accruals*, or both – and financial misrepresentation.

(iv) Non-financial firm characteristics (5 additional variables)

We include five measures of firm-specific characteristics that are not directly reported on firms' financial statements, including the number of employees, number of business segments, segment concentration, the number of geographic segments, and the average distance to the firm's markets. The

[8] See, for examples: Amoa-Gyarteng (2014), Bhavani and Amponsah (2017), Chan and Landry (2019), Drabkova (2015), MacCarthy (2017), Mahama (2015), Ofori (2016), Simpson (2016), Mehta and Bhavani (2017), and Taherinia and Talebi (2019).

Association of Certified Fraud Examiners (2018) argues that the likelihood of financial misreporting is positively related to the number of firm employees because both the likelihood of, and opportunity for, rogue actors increases with the number of employees. This argument is reflected in Dechow et al.'s (2011) F-Score, which includes the abnormal change in employees as a predictor variable.

Ex ante, we expect the likelihood of misrepresentation to increase with the firm's operational complexity and geographic diversity. Operational complexity, which we measure with the number of business segments and a measure of segment concentration, increases the cost of internal monitoring. It may also increase the net benefits of financial misrepresentation to divisional heads, who influence the flow of financial information to corporate headquarters. Likewise, an increase in geographic spread tends to increase the cost of internal monitoring. We therefore expect a positive relation between fraud likelihood and both the number of geographic segments and the average distance from corporate headquarters to the firm's markets.

(v) Oversight (3 additional variables)

Previous work (e.g., Kedia and Rajgopal, 2011) indicates that external oversight from auditors and regulators decreases the likelihood of financial misrepresentation. We measure auditor oversight with two measures. The first is a dummy variable, *Auditor opinion flag*, that equals one if the auditor issues an adverse or qualified opinion about the firm's financial statements. The second is a dummy variable, *Big N auditor flag*, that equals one if the financial statements were audited by a Big N auditor. An adverse auditor opinion indicates a disagreement between the firm's independent auditor and its managers about the representation of the firm's financial statement. A Big N auditor is widely considered to offer reputational capital and expertise that decrease the likelihood of misrepresentation. We also include the *Distance to regulator* because Kedia and Rajgopal (2011) find that the physical distance from a firm's headquarters to the nearest SEC office is positively related to the likelihood of misconduct.

(vi) Market and industry characteristics (11 additional variables)

We include a measure of the firm's industry concentration, the *Herfindahl index*, because several previous papers infer that the likelihood of financial misconduct is related to the competitiveness of the firm's industry (e.g., Wang and Winton, 2012; Boone, Grieser, Li, and Venkat, 2019). Finally, we include fixed effects for 10 industry groups because Choi, Lou, Karpoff, and Martin (2023) find that financial misrepresentation occurs in industry-specific waves. We define industries using the Fama-French 10-industry portfolio definitions.

In total, our tests consider 55 candidate predictor variables. Table 5 reports summary measures for these variables for violation firm-years (columns 1 and 2) compared to non-violation years for the same violating firms (columns 3 and 4), and also compared to all firm-years for non-violating firms (columns 6 and 7). The univariate comparisons show a number of patterns that are consistent with the prior models' findings. Among the variables used by Beneish (1999), for example, violating firms' gross margins, asset quality, sales growth, and leverage are all significantly higher during their violation years compared to their non-violation years. Gross margins, sales growth, and total accruals also are significantly higher during the violation years than for non-violators, on average, and leverage is significantly lower.

Similarly, most of the variables used by Dechow et al. (2011) are significantly different and in the expected direction for violation firm-years compared to non-violation firm-years. In their violation years, firms have relatively large RSST accruals, large changes in receivables, inventory and cash sales, a higher fraction of soft assets, a lower change in the number of employees, and are more likely to have an existing operating lease and a security issuance. Unlike Dechow et al. (2011), however, we do not find that violation firm-years are associated with a high change in return on assets.

Consistent with the results in Amiram et al. (2015) and Chakrabarty et al. (2019), we find that the *MAD score* is lower in violation years than non-violation years. This means that violation years have a significantly lower deviation from the Benford's law distribution than do non-violation years. The variables from Cecchini et al.'s (2010) prediction model also are generally different between the violation and non-violation firm-years, as are most of the additional firm characteristics in Panel E of Table 5. Overall, the univariate comparisons in Table 5 suggest a large number of characteristics that may be useful in constructing a prediction model for financial misrepresentation.

3.3. Evaluation of model performance

To evaluate a model's performance, we initially focus on its in-sample AUC measure. For the top-performing models we extend the analysis to conduct out-of-sample tests, k-fold validation tests, and other sensitivity tests. Each model generates fitted values for each firm-year observation that can be interpreted as the probability of misrepresentation in that firm-year. Firm-years with fitted values that exceed a certain probability threshold are classified as violation firm-years (positives), and values below the threshold are classified as non-violation firm years (negatives). We then compare the model's predictions to the observed violation firm-years in our sample, creating the following 2x2 confusion matrix:

Observed:

		Violation firm-years	Non-violation firm years	Total
Model predicted:	Positives (violation firm-years)	True positives (TP)	False positives (FP)	Predicted positives (i.e., violations)
	Negatives (not violation firm-years)	False negatives (FN)	True negatives (TN)	Predicted negatives (not violation firm-years)
	Total	Observed positives (violation firm-years)	Observed negatives (not violation firm-years)	Total observations (N)

The model's sensitivity is the rate at which it accurately identifies violation firm-years (i.e., avoids Type II errors), and is defined as the ratio of the number of true positives (TP) to the number of observed positives. The model's specificity is the rate at which its classifications avoid false positives (i.e., avoids Type I errors), and is defined as the ratio of the number of true negatives (TN) to the number of observed negatives. We use these statistics to generate each model's ROC curve. The ROC curve plots the sensitivity (true positive rate) as a function of the false positive rate (1-specificity) for all possible probability threshold levels. As noted previously, the AUC is the integral of the ROC curve and yields a single measure of the model's performance. At the optimum threshold level that maximizes the model's overall predictive ability, the AUC is the sum of the model's sensitivity and specificity.[9]

4. Logistic prediction models

4.1. Previous logistic prediction models

We begin by estimating the prediction models based on the subsets of predictor variables used by Beneish (1999) and Dechow et al.'s (2011) Model 2. Panels A and B of Table 6 report our replications of each of these two models using our violation sample from 1976-2014. Our replications closely approximate the results reported in the original papers, indicating that most of the empirical correlations between firm characteristics and misconduct that each of these models identifies are persistent in our sample as well. Like Dechow et al. (2011), for example, we find that the likelihood of a violation firm-year is positively related to accruals, change in receivables, change in inventory, the fraction of soft assets, change in cash sales, the existence of an operating lease, and a contemporaneous new security issue. The likelihood is negatively related to the change in ROA and a measure of the abnormal change in the number of firm employees.

Figure 1 displays the ROC curves for these two models based on within-sample tests using all firm-year data from 1976-2014. For the Beneish model, the AUC is 0.54, and for the Dechow et al. (2011) model the AUC is 0.67.[10]

[9] Models with an AUC of 0.5 show no discrimination (i.e. the classification of violation vs. non-violation years is as good as a coin flip). As a general rule, an AUC below 0.70 indicates poor model ability to discriminate, and an AUC between 0.70 and 0.80 shows acceptable discrimination. See Green and Swets (1966), Swets (1988), Hosmer and Lemeshow (2000, p. 162), and Chakrabarty et al. (2019). It is important to emphasize that in our application – identifying undetected misrepresentation – we do not have a separate observation into the truth that allows us to unmistakably identify false negatives and false positives, a situation that some researchers call the "gold standard" in ROC procedures, e.g., Joseph, Gyorkow, and Coupal (1995) and Zhou, Castelluccio, and Zhou (2014). Instead, we rely on in-sample observations of misconduct to calibrate the model, and then assume that the model's fitted values accurately identify firms that are misrepresenting. In Section 7.5 we show how to incorporate the modeler's prior beliefs about the model's sensitivity and specificity into the estimation process, and how changes in such beliefs affect our specific estimates.

[10] Perols et al. (2017, Table 3) and Bao et al. (2019, Table 3) also report AUCs for their replications of the Dechow et al. and the Cecchini et al. models. Although there is some variation, their values are similar, with AUCs ranging from

4.2. Logistic model based on Benford's Law

As noted, several papers use Benford's Law to identify undiscovered financial misconduct. Amiram et al. (2015) and Chakrabarty et al. (2019) focus on the Benford *MAD score*, which measures the mean absolute deviation of the distribution of leading digits of the numbers reported in a firm's financial reports from the distribution implied by Benford's Law. Panel C of Table 6 reports on our estimation of a prediction model based on the *MAD score*. Consistent with Amiram et al. (2015, Figure 4) and Chakrabarty et al. (2019), we find that the *MAD score* is significantly and negatively related to the likelihood of a violation firm-year in a logistic regression. Figure 1 displays the ROC curve for the prediction model based on the model in Panel C of Table 6. The AUC is 0.63, higher than the AUC for the Beneish model but lower than the AUC for the Dechow et al. model.

4.3. A broader logistic model

In this section we build upon prior models to construct a logistic model that has better properties of sensitivity and specificity in identifying firms that commit financial misconduct. Our goal is to capture the salient characteristics of firms that are found to engage in financial misconduct over the sample period. We consider all of the variables listed in Table 5, which include the firm characteristics used in the prior prediction models.

We estimate the following logistic regression:

$$\Pr(Financial\ Misconduct_{it}) = \alpha_{it} + \beta\ '\ Beneish_{it} + \gamma\ '\ DGLS_{it} + \delta\ Benford_{it} + \eta\ '\ Additional + \epsilon_{it}$$

The dependent variable is an indicator variable that takes the value of one for each firm-year of observed financial misconduct and zero otherwise. The subscript *i* denotes a firm, and the subscript *t* denotes a fiscal year. *Beneish* is a vector of variables used in the Beneish (1999) model. *DGLS* is a vector of variables used

0.58 to 0.67. Their AUCs are out-of-sample estimates, while those in Figure 1 are in-sample. We discuss out-of-sample performance in Section 7.3.

in Model 2 of Dechow et al. (2011). *Benford* includes the *MAD score*, which is the mean absolute deviation of the leading digit of firm financials from the theoretical Benford's Law distribution. There are a total of 23 Beneish, DGLS, and Benford variables. *Additional* is a vector of 27 of the 32 additional variables listed in Table 5 and discussed in Section 3. (In the tests reported we omit the five variables from Cecchini et al. (2010) because they are statistically insignificant, economically unintuititive, and require data that reduce the sample size. Excluding these five variables has no meaningful effect on the other estimates.) All variables are defined in the Appendix. To control for within-firm correlation, we cluster robust standard errors at the firm level. We begin by estimating the full model with all 50 predictor variables (omitting the Cecchini et al. (2010) variables) and then derive a more parsimonious model based on a subset of the variables.

Table 7 reports the results from estimating the full model. To accommodate all the predictor variables we report the results across vertical panels in the table. Among the variables used in Beneish's (1999) model, violations are positively related to gross margins, sales growth, and SG&A expenses – similar to the results for the individual Beneish model reported in Table 6, Panel, A. Among the Dechow et al. (2011) variables, violations are positively related to the change in receivables, change in inventory, the soft assets ratio, the operating lease flag, and the security issue flag. The *MAD score* is negatively related to violations, but its coefficient is not statistically significant.

The incidence of misrepresentation is significantly related to several of the *Additional* variables that are not included in the prior prediction models. For example, it is positively related to market capitalization, the dummy variable indicating recent earnings losses, and the indicator variable for a qualified audit opinion, and it is negatively related to the Allen, Larson and Sloan (2013) measure of accruals, the concentration ratio of the firm's sales across business segments, and a *Big N* auditor dummy variable.

The full model's AUC is 0.7857, which is substantially higher than any of the prior logistic models and offers an acceptable level of discrimination according to Hosmer and Lemeshow (2000). The AUC of a model that excludes the *Additional* variables is 0.6761. The *Additional* variables, which are not included in any of the previous logistic models, therefore contribute substantially to the explanatory power of our model.

A drawback with the full model is that it has a large number of predictor variables. To develop a more parsimonious model, we follow Dechow et al. (2011) by running a stepwise regression procedure with backward elimination. At each step, we re-estimate the model after excluding the predictor variable with the largest p-value, continuing until all remaining predictor variables are significant at the 5% level. The results of this procedure are reported as Model 2 in Table 7.

This more parsimonious model consists of a total of 17 predictor variables. This set includes three variables from the Beneish (1999) model (*Gross margin index*, *Sales growth index*, and *SG&A index*) and five variables from Dechow et al. (2011) (*Change in receivables*, *Change in inventory*, *% Soft assets*, *Operating lease flag*, and *Security issue flag*), while the *MAD Score* drops out. Nine of the *Additional* variables also remain in the final model, including *Market cap*, the Allen et al. (2013) abnormal accruals measure (*ALS accruals*), *Loss flag*, the *Segment concentration index*, the *Number of geographic segments*, *Auditor opinion flag*, *Big N auditor flag*, and two industry indicators *(Business equipment* and *Telecom*). The area under the ROC curve for the parsimonious model remains high, at 0.7813.[11]

5. Machine learning prediction models

5.1. Machine learning classifiers

The logistic models in Section 4 use theory to identify potentially important fraud predictors, check for statistical significance, and run diagnostic tests. Machine learning classifiers, in contrast, use pattern matching between the outcome variable and potential predictors without concern for economic intuition or statistical issues such as multicollinearity. In this section we compare the logistic model from Section 4.3

[11] The operating lease flag does not apply to fiscal years beginning after January 1, 2019, as firms must now report operating lease-related assets and liabilities on their balance sheets (IAS 16 of the International Accounting Standards Board and the Financial Accounting Standards Board's Accounting Standards Update 842). Replicating our tests without the *Operating lease flag* yields similar results, with an AUC of 0.7800 for the parsimonious model.

to 14 machine learning models used previously in the fraud prediction literature. These include linear models (stochastic gradient descent and Gaussian naïve Bayes), k-nearest neighbors, decision trees, ensemble methods (random forest, extra trees, and random under-sampling boosting), artificial neural network (multi-layer perceptron), discriminant analysis (quadratic), and support vector machines (SVMs) using five different kernels (two linear, sigmoid, polynomial, and radial bias function).[12] The two linear kernels differ by their implementation using different library functions, as described in Internet Appendix Table IA.1 (the LinearSVC kernel is less flexible but faster on large datasets). We consider these five different SVMs because prior research suggests these algorithms perform well in fraud prediction models.[13]

To implement and compare the machine learning models, we make several decisions regarding model dimensionality, corrections for overfitting, and hyperparameters. Regarding dimensionality, we begin by using the same 17 variables derived from the backward elimination process of the logistic model discussed in section 4.3, and by standardizing each variable. This allows us to maintain a common sample and to compare the models directly.[14]

Because they are explicitly atheoretic data-mining algorithms, machine learning classifiers generally perform well in classifying the variable of interest within the data on which the models are trained, even if they fail to predict anything useful on yet unseen data. To avoid this problem, we follow common practice by holding out part of the available data as a test set. We use a random 60/40 split, which results in a training dataset containing 107,202 random cases from the original dataset and a testing dataset containing 71,468 cases.

[12] The SVM kernel is the mathematical function used to map the feature set (i.e., the predictor variables) to the variable of interest (fraud). It is a weighing factor between two sequences of data that can assign more weight to one data point at one time point than another data point.

[13] See for example Singh et al. (2012) A Machine Learning Approach for Detection of Fraud based on SVM. *International Journal of Scientific Engineering and Technology*, Vol. No.1, Issue No.3, pp. 194-198 and Sangeetha et al. (2017) Benefits of SVM and Deep Learning in Credit Card Fraud Detection – A Survey, *Int. Conf. on Signal, Image Processing Communication & Automation,* Grenze Scientific Society.

[14] The Internet Appendix reports on tests in which we do not constrain the number of variables to this group of 17. Starting with larger sets of variables available in the Compustat database, we considered feature reduction techniques including variance threshold, best scoring, false positive and false discovery rate, penalized linear models, and lasso regression. These approaches yield reduced sets of variables that overlap significantly with the 17 variables used in the models reported here and do not result in significant improvements in model performance. We also considered other variable transformations, including min-max scaling, but standardizing yields the best model fits.

We also use k-fold cross-validation, in which the training data are partitioned into k smaller sets (we set k = 5 for the tabulated results). The model is trained using k-1 folds of the training dataset and is validated on the remaining part of the data (i.e., it is used to compute a performance measure such as accuracy). The process is repeated k times subsequently withholding each fold once as the validation set. The performance measure reported in the training step by k-fold cross-validation is then the average of the values computed in a loop through all k-folds of the training dataset. While computationally expensive in larger datasets ($n \approx 100{,}000$), this process does not waste data and avoids an extreme result obtained by a chance selection of the validation set, which is a major advantage in problems of inverse inference such as fraud where the number of samples with the variable of interest infrequently occur. This process additionally aids in the prevention of overfitting.

To perform well, most machine classifiers require additional parameters not learned within the estimator algorithm – called hyperparameters – that are determined before applying the estimation process. Most machine learning libraries establish pre-set default values for these hyperparameters, but pre-set values generally lead to suboptimal predictions. The most popular approach to adjust hyperparameter values, and the one used here, is to determine the optimal hyperparameter values for each classifier using an exhaustive grid search over a range of possible values on the training dataset. The combination of hyperparameter values that produces the best predictions in a withheld sample are then used in the training and testing process. Table IA.2 in the Internet Appendix presents the grid search values we use and the resulting optimal hyperparameters for each of the machine learning classifiers.

5.2. Machine learning model results

Table 8 summarizes the results for the 14 machine learning models plus the logistic regression model from Section 4.3. For each model, we report the average of the following performance metrics from a 100-trial bootstrap procedure:

Balanced accuracy (Bal Acc) – the average of how well the classifier predicts each class;

Sensitivity (Sens) – the percentage of misrepresentation firm-years in the test sample that the model accurately flags as misrepresentation firm-years (i.e., true positives);

Specificity (Spec) – the percentage of non-misrepresentation firm-years in the test sample that the model accurately flags as non-misrepresentation firm-years (i.e., true negatives);

Geometric mean (GMean) – the geometric mean of sensitivity and specificity;

Type I (false-positive) error rate – the percentage of non-misrepresentation firm-years in the test sample that the model flags as misrepresentation firm-years;

Type II (false-negative) error rate – the percentage of misrepresentation firm-years in the test sample that the model flags as non-misrepresentation firm-years;

AUC (area under receiver operating characteristic curve) – a measure of how well the model distinguishes between the misrepresentation and non-misrepresentation firm-years, where 1.0 represents perfect classification and 0.5 is no better than a random coin toss for classification;

Rank – the rank order of the model based upon the AUC; and

Standard deviation (SD) – the standard deviation of the AUC over the 100 trials.

Ranked by AUC, the support vector machine classifier using the radial basis function kernel estimator provides the best prediction ability of all classifiers with an average AUC of 78.32%. The logistic model and support vector machine classifier using the linear function kernel are close seconds, each with an average AUC of 77.97%. The k-nearest neighbor and the multi-layer perceptron artificial neural network models have poor predictive qualities, each with average AUCs under 70%.[15] All other classifiers have average AUCs in the 0.7 to 0.8 range. Some classifiers, however, exhibit poor consistency

[15] Again, we use the following rule of thumb as suggested in Hosmer & Lemeshow (2013), Applied Logistic Regression. p.177:

0.5 = No discrimination, model is no better than random guessing.
0.5-0.7 = Poor discrimination, not much better than random guessing.
0.7-0.8 = Acceptable discrimination
0.8-0.9 = Excellent discrimination
> 0.9 = Outstanding discrimination

as indicated by relatively high standard deviations. For example, the ensemble method using the random under-sampling boosting estimator as proposed in Bao et al. (2019) is the least consistent among the machine learning classifiers. The standard deviation of its AUC measures is 0.0374, which is nearly seven times the standard deviation of the logistic model AUCs. This indicates that the ensemble method model is highly sensitive to the data on which it is trained.

The results in Table 8 limit the machine learning algorithms to the 17 variables used in our reduced logistic model. It turns out that this limitation does not have a large effect on the model outcomes. As reported in the Internet Appendix Table IA.3, the performance of each model as measured by the AUC improves only slightly if we use the candidate predictor variables listed in Table 5.[16] For example, the AUC of the SVC(radial basis function) model improves slightly from 78.32% to 79.26%. The AUC of the k-nearest neighbor model increases the most, from 57.44% to 66.18%.[17]

In total, the results in Table 8 indicate that the logistic model of Section 4.3 and three of the SVM models perform the best as measured by their average AUCs. In subsequent analyses we focus on these four models. Of these models, only the logistic model is motivated by economic intuition and yields economically meaningful estimates of the marginal impacts of various firm characteristics on the likelihood of fraud. We therefore refer to the logistic model of Section 4.3 as our base model for estimating the prevalence of misrepresentation.

6. Bivariate probit models

Poirier's (1980) bivariate probit model provides an alternate approach to logistic regression and machine learning models to address the problem of partial observability. The bivariate probit approach models the outcome of misrepresenting and being caught, P(Detected), as the product of two latent processes that are estimated simultaneously: the probability of misrepresenting, P(Misrepresentation), and

[16] Again, of the 55 variables in Table 5, we exclude the five variables from Cecchini et al. (2010).

[17] In predictive modeling, modelers sometimes tweak the inputs ex-post in an attempt to improve the model's predictive ability. We have not made any ex post tweaks because such adjustments are difficult to replicate and could reflect researcher bias.

the probability of getting caught conditional on misrepresenting, P(Detected|Misrepresentation). Wang (2013) applies this model to financial misconduct and identifies several firm characteristics that are associated with each of these probabilities. For example, a firm's R&D expense is positively related to the probability of committing fraud but negatively related to the probability of getting caught.

The Appendix Table IA.4 reports on our efforts to estimate prediction models using bivariate probit models. We replicate Wang's (2013) model and estimate an extension of this model using a larger number of covariates. A problem with using Wang's (2013) specific covariates is that some of them are measured after the firm's misrepresentation begins, making this specific model inappropriate for prediction. We relegate the details of our bivariate probit models to the Appendix, however, for two reasons. First, using the AUC metric, the models we estimate do not perform as well as the top-performing logistic and machine learning models. Second, in our application the bivariate probit models are highly sensitive to specification changes, and in many instances do not converge at all. We therefore focus on the logistic and top machine learning models.

7. The prevalence of financial misrepresentation

In this section we draw from the top-performing prediction models in Sections 4 and 5 to generate measures of the prevalence of financial misrepresentation, including misrepresentation that is not detected. We focus on the logistic model's results from Section 4, but the estimates are similar using the three top-performing machine learning models from Section 5. Again, in a prediction context, these four models have the highest AUCs (0.78) among the various models we estimated.

For each model, we use the model's estimates to classify each firm-year into violation and non-violation buckets by comparing the firm-year's fitted value from the model to a threshold level. Using standard ROC methods, at each candidate threshold level we calculate the sensitivity and specificity of the classifications assuming uncaught firm-years are true non-violators.[18] The optimal threshold level is determined by weighing the tradeoff between the Type I error rate (falsely labeling a non-misrepresenting firm-year as a misrepresention firm-year) and Type II error rate (falsely labeling a misrepresenting firm-year as non-misrepresenting). Our initial estimates assume the Type I and Type II errors are equally costly (i.e., the error cost ratio is 1/1), in which case the optimal threshold level maximizes the simple average of the sensitivity and specificity.

In our sample, violations have an average duration of 3.1 years. This is consistent with the information provided in regulatory proceedings and indicates that regulators frequently prosecute longer-running programs of financial misrepresentation rather than instances in which a single financial statement's numbers are in error. To reflect a tendency for enforcement actions to target longer-running violations, we use each model to classify a firm as engaging in misconduct only if the model classifies misrepresentation for at least three consecutive years.

7.1. Logistic model estimates of the prevalence of misrepresentation

Table 9, Panel A, reports the confusion matrix for our base case estimates using the parsimonious logistic model. The cutoff level that maximizes the model's combined sensitivity and specificity, i.e., the optimal cutoff, is 0.0137. Firm-years with fitted values of 0.0137 are therefore indications of potential misrepresentation. To reflect our observation that the average violation period is slightly longer than three years, and we classify a firm-year to be a misrepresentation year it is part of a sequence of three or more consecutive years in which the fitted values from the logistic model are at least 0.0137.

Panel B of Table 9 reports that, using this procedure, the model sensitivity (true positive rate) is 60.1% because it correctly predicts 1,405 of the 2,339 (60.1%) misrepresentation firm-years in the sample.

[18] See Joseph, Gyorkow, and Coupal (1995) and Zhou, Castelluccio, and Zhou (2014). In Section 7.3 below, we show how our estimates change if we instead impose alternative assumptions about the rate at which observed non-violating firm-years are, in fact, non-violating firm-years.

Similarly, model specificity (true negative rate) is 78.2%, as the model correctly predicts 137,847 of the 176,331 (78.2%) of non-misrepresentation firm-years.

This prediction model can be used to identify individual misrepresenting firms and the years of their model-implied violations, as illustrated in Table 2. Our objective, however, is to estimate the fraction of firms engaged in financial misrepresentation (P(violate)) and conditional on misconduct, the probability that a firm is caught (P(caught|violate)). These estimates are reported in Table 9. P(violate) equals (TP + FP)/N, or 22.3%, indicating that 22.3% of firm-years in the sample are characterized by financial misrepresentation. P(caught|violate) = TP/(TP+FP), or 3.5%, indicating that, of all violation firm-years, 3.5% eventually become subject to enforcement action.

It is important to note that the average violation in our sample of enforcement actions persists for 3.1 years. This number can be used to provide insight into the probability that a given firm (as opposed to a firm-year) engages in misrepresentation. In particular, the estimate that 22.3% of firm-years are violation firm-years implies that an average of 7.2% (= 22.3% ÷ 3.1 years) of firms begin new programs of financial misrepresentation each year.

Our base estimates reflect assumptions about the duration of predicted violations and the relative costs of Type I and Type II errors. Table 10 reports on how the estimates are affected by changes in these assumptions. The highlighted row reports our base case estimates in which we assume equal costs of Type I and Type II errors and require that the model generates three consecutive violation years to classify an instance of misconduct. If we maintain the 1/1 error cost ratio for Type I and Type II errors but require violation periods of only one year to be classified as an instance of misconduct, the model estimates that 28.5% of all firm-years involve financial misrepresentation, of which 3.2% become subject to regulatory enforcement action. In general, imposing a shorter time period to identify misconduct causes the model to classify more firm-years as violation firm-years, with a corresponding decrease in the probability of detection and enforcement action.

The estimates also are sensitive to changes in the error cost ratio. The top rows with an error cost ratio of 1/1.5 report the model estimates if we assign a 50% higher cost to Type II errors relative to Type I errors. The bottom rows with an error cost ratio of 1.5/1 assume the opposite – that Type I errors are 50% more costly than Type II errors. In criminal cases, U.S. jurisprudence emphasizes the principle that a defendant is innocent until proven guilty, and a long literature argues that U.S. and common law generally treat Type I errors as very costly. We conjecture that SEC and DOJ enforcement of financial misconduct follows similar principles, and that, if we deviate from an assumed error cost ratio of 1/1, it should be toward higher error cost ratios in which the model is trained to treat Type I errors as more costly.

The implied fraction of misconduct firm-years decreases with the error cost ratio, since a lower error cost ratio assigns a higher cost to falsely classifying a firm-year as in violation. If we maintain the three-year duration requirement and assume an error cost ratio of 1/1.5, the model identifies 13.5% of Compustat-listed firms as engaging in financial misrepresentation during an average year. Of these violation firm-years, 4.6% are eventually identified as parts of enforcement actions in which firms are caught and penalized.

Figure 2 provides a visual representation of the range of model estimates reported in Table 10. This figure plots the probability of violation and (conditional on a violation) the probability of getting caught as a function of the assumed error cost ratio and the violation duration. The figure illustrates how changes in the assumed error cost ratio generate relatively large changes in the implied prevalence of misrepresentation. Using a three-year violation duration, for example, changing the error cost ratio from 1.5/1 to 1/1.5 changes the implied probability of misrepresenation from 37.1% to 13.5%.

7.2. Machine learning model estimates of the prevalence of financial misrepresentation

Figure 3 reports parallel implications for the prevalence of misrepresentation using the results from the top three performing machine learning models. Using a 1/1 error cost ratio and a three-year violation duration, the estimates for the probability of violation range between 22% and 26% (Panel A), and the estimates for the probability of being caught range between 3.1% and 3.6% (Panel B). For each model, however, the implied probability of violation is negatively related to the assumed duration of the undetected violation and positively related to the assumed Type I/Type II error cost ratio. Similarly, the implied probability that a violating firm faces enforcement action is positively related to the assumed violation duration and negatively related to the error cost ratio. Assuming undetected violations last one year and that regulators assess false positives to be 50% more costly than false negatives, for example, the Support Vector Machine classifier using a radial basis function implies that the probability of a firm-year being in violation is 18.3%, with the probability of facing enforcement action equal to 4.1%.

Overall, the top-performing machine learning models generate results that are similar to each other and similar to those of the logistic model. These implied probabilities of financial misrepresentation range from 12.8% to 45.8%, and the implied probabilities of getting caught range from 2.4% to 4.7%. As previously observed, the logistic model yields results that are easily interpreted and are not computationally demanding. The rest of our analysis therefore focuses on the logistic model results.

7.3. Logistic model performance: Out-of-sample tests

The basic observation that motivates our investigation – the fact that some financial misrepresentation is undetected – also prohibits us from independently verifying each model's accuracy. In the ROC literature, this is known as an application that does not meet the gold standard of performance assessment. One substitute method to probe the model's reliability is to examine the model's sensitivity and performance in out-of-sample tests.[19] In this section, we focus on the logistic model's out-of-sample tests and extensions.

As noted above, the logistic model performs well in-sample, with an AUC of 0.78. To conduct out-of-sample tests, we partition the data into different training and holdout subsamples, each time estimating the model on the training sample and evaluating its performance in the holdout sample. The first type of partition is via date. We begin by estimating the model using firm-year observations through 1990 and test the model performance using post-1990 data. In successive iterations, we extend the training data period

[19] The literature refers to prediction models that can be tested against independent indication of true positives and true negatives as satisfying the "gold standard." See, e.g., Joseph, Gyorkow, and Coupal (1995) and Zhou, Castelluccio, and Zhou (2014). Related papers that also emphasize out-of-sample tests include Larcker and Zakolyukina (2012), Perols et al. (2017), and Bao et al. (2019).

by one year, stopping when the training data include all observations through 2014 – which, since it is the last year of our sample period, yields results that are the same as our base case.

Figure 4 reports the results of this procedure. In this figure, we present the out-of-sample AUC from the holdout subsample and, for comparison, the in-sample AUC for the training subsample. Model fit as measured by the AUC is quite stable over time, with in-sample AUCs in the 0.75-0.79 range. The out-of-sample AUC also is high and extremely stable, ranging from 0.73-0.77. The average out-of-sample AUC is 0.7523.

Bao et al. (2019) note that a bias can arise when an individual misrepresentation spans the training and the test periods. Our tests may be subject to such bias because our misrepresentation periods average 3.1 years. To examine this issue, we repeat the analysis in Figure 4 with Bao et al.'s correction, leaving a two-year gap between the training and test periods. For example, when the training period extends through 2004, our test period begins in 2006. The average out-of-sample AUC from this procedure is 0.7482, which is only slightly lower than the average out-of-sample average AUC of 0.7523 without the two-year gap. We infer that our results are not severely affected by a potential overlap between training and test periods.

We next evaluate the model's out-of-sample performance using a stratified k-fold cross-validation test.[20] The key difference in this test is that the training and holdout samples are determined randomly instead of by consecutive years. The results, which are graphed in Figure 5, are extremely stable. Not only is the mean out-of-sample AUC equal to 0.78, but the range in AUC scores is narrow, from 0.76 to 0.81, and in none of the estimation samples does the AUC fall below 0.70.

7.4. Model flexibility and the error cost ratio

These tests indicate that the logistic model's estimates are stable over time and have good prediction capability. Our base logistic model also yields estimates of the prevalence of financial misrepresentation

[20] To conduct this test, as a first step, the data are randomly split into six folds, or subgroups. The number of violations in each bucket is equal. In the second step, we use data from five of the folds to fit the model and use the sixth fold to test the model. This test results in an ROC curve. We repeat this step for each of the folds, resulting in six different ROCs. Figure 5 plots the average ROC along with information on the distribution across the six iterations.

that are similar to those generated by the best-fit machine learning models. Still, we cannot rule out measurement error in these estimates. It turns out, however, that model classification error can be viewed as a problem of choosing the best error cost ratio.

To illustrate this point, suppose the modeler has a prior belief that the true unconditional probability that a firm-year is in violation of 13(b) reporting rules is lower than the 22.3% implied by our base case. Such a belief is isomorphic to assuming that the base case identifies too many firm-years as misrepresentation firm-years because it underweights the cost of false positives. If we believe the model generates too many false positives, we can use this belief to assign a higher cost to false positives, i.e., by choosing a higher error cost ratio. Adjusting the error cost ratio allows the modeler to combine the data with her prior beliefs to estimate the prevalence of misrepresentation.

Alternatively, we can use the error cost ratio to explicitly incorporate beliefs about the enforcement regime. Suppose, for example, we believe that regulators assign a 50% higher cost to false positives than to false negatives in their allocation of enforcement resources. We can apply an error cost ratio of 1/1.5, thereby changing the specific firm-years that are flagged as misrepresentation firm-years. Using Table 10, the implied probability of misrepresentation with a 1/1.5 error cost ratio is 13.5% (assuming a violation period of three or more years).

As another illustration, we can adjust for biases that cause our model to identify too few firm-years as misrepresentation years. For example, the fact that our known true misrepresentation firm-years rely on the SEC taking enforcement action suggests that there are additional (undetected) misrepresentation periods that we do not use to calibrate our prediction model. We can offset a concern about identifying too few misrepresenting firm-years by *decreasing* the threshold at which the prediction model classifies a firm-year as in violation, i.e., by increasing the assumed error cost ratio.

In summary, out-of-sample tests show that the logistic prediction model yields estimates that are stable across time and that perform well. Additional concerns about model biases can be recast as concerns that the model should weight Type I or Type II errors at something other than a 1/1 ratio. Table 10 shows how to translate alternate assumptions about the error cost ratio into data-driven estimates of the prevalence of financial misrepresentation and the likelihood that misrepresenting firms face regulatory sanctions.

8. The social cost of financial misrepresentations and policy implications

8.1. Share price distortions implied by our estimates

Researchers, practitioners, and regulators maintain that financial misrepresentation is socially costly because it distorts the information available to investors in pricing financial securities, leading to suboptimal investment and risk-bearing decisions.[21] Empirical estimates of the size of this social cost, however, are elusive. In this section we use our estimates of the prevalence of misconduct to provide new insight into the nature and extent of the price distortions generated by financial misrepresentation.

We begin by noting that the logistic model's base estimate is that 22.3% of Compustat-listed firms are misrepresenting their financial statements in any given year. Karpoff, Lee, and Martin (2008a) report data indicating that, before misconduct is revealed, the share prices of misrepresenting firms are inflated by an average 10.3% compared to their non-inflated values.[22] Together, these estimates imply that, in any given year, an average of 22.3% of all firms' share values are artificially inflated by 10.3%.

Previous research emphasizes the potential for artificial price distortions in the shares of firms that are directly affected by financial misconduct.[23] But the impacts are not limited to misrepresenting firms. If non-misrepresenting firms' share prices were unaffected, investors would systematically overpay for shares. That is, misrepresenting firms' share prices would be inflated by an average of 10.3% and non-misrepresenting firms' share prices would not be inflated. Aggregating across all firms, the average firm's share price would be artificially inflated by 22.3% * 10.3% = 2.7%.

[21] E.g., see Rajan and Zingales (1998), Duffie and Lando (2001), Graham, Harvey, and Rajgopal (2005), Sadka (2006), and Giannetti and Wang (2016).

[22] This estimate is implied by Karpoff et al.'s (2008a) finding that firms caught misrepresenting their financials experience an average decrease in share values of 38.06% and that 9.34 percentage points (or 24.5%) of this loss is the average amount by which the share values were artificially inflated. The price inflation compared to the non-inflated (i.e., pre-misconduct) share price is [1/(1-9.34%]-1) = 10.3%.

[23] E.g., see Fishman and Hagerty (1992), Ferrell and Saha (2011), and White (2020).

An outcome in which investors systematically overpay for shares, however, is not an equilibrium. We should expect investors to consider the fact that some firms' share prices are artificially inflated by fraudulent reporting, and to discount all share prices accordingly. If investors fully price the expected price inflation, they will discount shares of non-misrepresenting firms by an amount ∂ to offset the expectation of loss from investing in artificially inflated shares. The discount ∂ can be calculated by setting the expected value of an invested dollar to \$1, i.e., $1 = .223(1+10.3\%) + (1-.223)\ (1-\partial)$, implying that $\partial = 3.0\%$. This implies that even non-fraud firms experience a cost of misrepresentation, as investors discount the values of non-misrepresenting firms by 3.0%. The 3.0% discount is ex ante protection against the likelihood that some shareholdings will turn out to be in firms with artificially inflated prices.

Thus, financial misrepresentation causes two types of price distortions. First, misconduct firms' share prices are inflated. Our base-case estimate indicates that, in any given year, 22.3% of firms artificially inflate their share values by an average of 10.3%, compared to the full-information scenario in which these firms did not misrepresent their financials. Second, investors rationally adjust to the possibility of investing in inflated shares. As a result, the share values of the 77.8% of firms that are not misrepresenting their financials are discounted by an average we estimate at 3.0%.

These are, of course, rough estimates. If we assume that share price inflation tends to be higher among firms that eventually are caught than among cheating firms that are not caught – perhaps because regulators target the most egregious cases of misrepresentation – our estimate of an average price inflation of 10.3% is too high. On the other hand, if only large instances of misrepresentation attract regulatory attention, there likely exists a fair amount of lesser price inflation among many firms that our estimates do not incorporate. This implies that some of the 77.8% of firms that we classify as not misrepresenting do, in fact, engage in some manipulations that inflate their share prices. If so, our estimate of a 3.0% discount for non-cheating firms is too low. Taking into account such offsetting considerations, we conclude that our specific estimates – that share prices are inflated by 10.3% for a sizeable minority of firms, but discounted by 3.0% for the majority of non-cheating firms – are subject to bias, but the direction of bias is unclear.

8.2. Other costs of financial misrepresentation

Society devotes resources to monitor firm behavior and discipline fraudulent activities. Costs of enforcement include (at least) the enforcement budgets of agencies with enforcement powers, including the SEC and the securities division of the DOJ as well as state enforcement agencies, and a substantial portion of the private legal enforcement process. These are real costs of misconduct, but they are not directly attributable to the price distortions for which our results provide estimates.

An additional cost of misconduct is the loss in firms' reputational capital when they are caught. As Karpoff et al. (2008a) and others show, lost reputation frequently is much larger than any costs imposed on firms through legal and other direct penalties for financial misconduct. The reputation cost tends to be internalized by the misconduct firm and its investors, so it does not represent an external social cost. Since it is a cost to the firm, however, its magnitude is relevant for business and social policy, as discussed in the next section.

8.3. Implications for business policy and enforcement policy

Our estimate of the probability that misrepresenting firms are caught can be used to gain insight into optimal firm and enforcement policies. Our base case estimate is that the probability of getting caught for financial misrepresentation is 3.5%. Karpoff et al.'s (2008) results indicate that firms that are caught misrepresenting their financials suffer ex post penalties that sum to 31.7% of the firm's pre-misconduct value. These penalties include both direct (legal and regulatory) and indirect (reputational) losses.[24] So, ex ante, the expected cost to a firm of engaging in financial misrepresentation averages .035 * 31.7% = 1.1% of pre-misconduct share value. This estimate does not capture the costs that are internalized by the firm's managers, which can include the loss of job, debarment from serving as a corporate executive, or even jail time (Karpoff, Lee, and Martin, 2008b). Nevertheless, it demonstrates that, from the firms' shareholders'

[24] This is indicated by Karpoff et al.'s (2008a) finding that ex post penalties average 28.7% of the pre-revelation share price, and the pre-revelation share price is artificially inflated by 10.3% because of the misrepresentation: (1 + .103) (1 – .3806) – 1 = 31.7%.

perspective, the expected cost of engaging in financial misrepresentation is roughly 1.1% of the firm's pre-misconduct market capitalization.

Becker's (1968) theory of crime implies that penalties for misconduct are optimized when, at the margin, the expected penalties equal the expected benefit from the underlying activity. Using this framework, our estimate provides a starting point for considering optimal fraud penalties. For example, Amiram et al. (2018) summarize evidence about managers' motives to engage in fraud. One motive is that, by inflating the share value, managers can extract higher prices when issuing new shares. Our results indicate that such price inflation can be optimal from existing shareholders' views if the expected benefits exceed the expected cost of 1.1% of share value.

From a business policy perspective, the ex ante cost of financial misrepresentation at the firm is on the order of 1.1%. This provides a starting point for considering the firm's optimal investment in internal controls and fraud prevention. For example, Burns and Kedia (2006) show that managers are incentivized to engage in artificial share price inflation to meet compensation performance thresholds. Share-price based incentives, combined with poor internal monitoring, therefore can expose the firm to managerial-initiated misconduct that imposes an expected cost of 1.1% of firm value. More broadly, any managerial incentives to artificially inflate share prices expose the firm to the risk of legal and reputational penalties. Our estimates indicate that the ex ante cost of such risk, considering the probability of getting caught, is 1.1% of equity value.

9. Conclusions

A large empirical literature examines the motives for and consequences of financial misrepresentation and fraud (for a review, see Amiram et al. 2018). Inferences from most of this research, however, come with a caveat: we can examine the characteristics and effects only of misrepresentation that is detected, usually because it triggers a restatement, lawsuit, or regulatory enforcement action. Most commentators agree that a large amount of misrepresentation remains undetected, at least in the samples from which researchers and policymakers draw inferences. This limits our ability to understand the drivers and effects of misrepresentation in general as opposed to the observations of detected misrepresentation.

This paper seeks to facilitate a broader understanding of otherwise undetected misrepresentation by developing a model that identifies firms as likely violating or not violating financial reporting rules in a given year. We consider a large number of firm characteristics that can serve as predictor variables and develop a prediction model using data from a uniquely comprehensive sample of firms that were sanctioned by the SEC and DOJ for financial misrepresentation that occurred from 1976-2014. To guide the application of the model, we employ Receiver Operating Characteristics procedures, which allow us to explicitly consider and trade off Type I and Type II classification errors.

Our base case results indicate that, in an average year, 22.3% of Compustat-listed firms engage in misrepresentation practices that are potential targets for regulatory sanctions. A total of 3.5% of these violation firm-years are targeted by SEC enforcement action for financial misrepresentation. The average misrepresentation period lasts 3.1 years, so our estimates imply that 22.3% ÷ 3.1 years = 7.2% of firms initiate new programs of potentially prosecutable financial misrepresentation in an average year.

The model and its corresponding estimates are stable over time and perform well in out-of-sample tests. The estimates, however, are sensitive to assumptions about the relative costs of Type I and Type II classification errors and about the minimum duration of model-indicated misconduct required to classify a firm-year as in violation of reporting rules. Plausible variations in these two assumptions from our base case tend to work in opposite directions, the effect of which is to yield fairly stable estimates of the prevalence of misrepresentation and the likelihood that it is detected. Assuming both a 50% higher cost of Type I errors (false positives) compared to Type II errors (false negatives), and classifying firm-years as violations even if the violation persists for only one year, the estimated probability of misrepresentation changes to 18.0% and a corresponding probability of detection to 4.1% – estimates that are close to our base case estimates.

As illustrated in Table 2, our model can be used to identify specific firms and years in which financial misrepresentation is likely. However, we use our estimates to provide new evidence on the prevalence and costs of financial misrepresentation. Our base case estimates indicate that, in an average year, 22.3% of firms' share values are artificially inflated by 10.3%. The share price inflation affects non-misrepresenting firm values also, as investors with knowledge that some firms are inflating their share values will discount all other firms' values by an amount that we estimate at 3.0%. These estimates provide evidence on the size of the price distortions induced by financial misrepresentation, which in turn can impose social costs through their effects on firms' investment decisions and investors' portfolio choices.

Our estimates also provide new insight into the size of firms' potential costs from financial misrepresentation and the optimal public policy to deter misrepresentation. Our base-case estimates imply that the unconditional expected cost to a firm's shareholders of engaging in financial misrepresentation averages 1.1% of the firm's market capitalization. This estimate provides a starting point from which to guide firm and public policy. For firms, the expected cost of financial misrepresentation – incorporating both the costs if detected and the probability of detection – is in the neighborhood of 1.1% of market capitalization, providing a guide to the optimal amount of investment in internal controls to deter misrepresentation. For regulators, current enforcement and regulatory penalties impose an ex ante expected cost of 1.1% of a firm's market capitalization. Financial misrepresentation that yields expected benefits in excess of 1.1% of market capitalization will continue to be optimal for some firms unless penalties or enforcement increase. To the extent that the managers who engage in the misconduct do not internalize a large share of the cost, some managers will engage in misrepresentation even when the expected benefits to the firm are less than 1.1% of market capitalization.

Appendix: Potential predictor variable definitions

This table presents definitions and transformations of the variables used in our analyses for predicting financial misrepresentation. We obtain firm financial information from the Compustat Annual North American data file for the years from 1976 to 2014. Segment data are obtained from the SEG_GEO and SEG_ANNFUND data files of Compustat from Wharton Research Data Services (WRDS).

Variable	Definition	Transformation
Beneish model variables		
Days' sales in receivables index	The ratio of days' sales in receivables in year t to year t-1.	
Gross margin index	The ratio of gross margin in year t-1 to year t.	
Asset quality index	The ratio of PP&E to total assets in year t to year t-1.	
Depreciation index	The ratio of the depreciation rate in year t-1 to year t.	
Sales growth index	The ratio of sales in year t to year t-1.	
Leverage index	The ratio of total debt to total assets in year t to t-1.	
Total accruals/total assets	The ratio of total accruals (change in working capital accounts other than cash less depreciation) to total assets.	
SG&A index	The ratio of SG&A in year t to year t-1.	
Dechow et al. model variables		
RSST accruals	Richardson, Sloan, Soliman, and Tuna (2006) accruals.	
Change in receivables	Change in receivables divided by average total assets.	
Change in inventory	Change in inventory divided by average total assets.	
% Soft assets	Total assets less PP&E less cash and cash equivalent divided by total assets.	
Change in cash sales	Sales less the change in accounts receivable.	
Change in ROA	Change in return on assets.	
Abnormal change in employees	Percentage change in the number of employees less the percentage change in assets.	
Operating leases flag	Dummy equals 1 if future operating lease obligations is greater than zero.	
Security issue flag	Dummy equals 1 if the firm issued securities during the year.	
Benford's Law variable		
MAD score	The summation of the absolute difference between the empirical distribution of leading digits in the financial statement and the theoretical Benford's law distribution, scaled by the number of leading digits.	
Cecchini et al. model variables		
Lag sales / lag preferred stock	Ratio of sales in year t-1 to total preferred stock in year t-1.	Scaled by 1,000,000
SG&A / investment and advances	Ratio of SG&A in year t to investment and advances in year t.	Scaled by 1,000,000
Lag total assets / lag investment and advances	Ratio of total assets in year t-1 to investment and advances in year t-1.	Scaled by 1,000,000
Lag sales / lag investment and advances	Ratio of sales in year t-1 to investment and advances in year t-1.	Scaled by 1,000,000
Total assets / short-term investments	Ratio of total assets in year t to total short-term investments in year t.	Scaled by 1,000,000

Other candidate predictor variables		
Market cap	The product of price and total shares outstanding.	Log(0.000001 + *Market Cap*)
Market to book	The summation of market capitalization and total assets minus total liabilities, divided by total assets.	Winsorized at the 1% level
Leverage index	The ratio of total liabilities to total assets.	Winsorized at the 1% level
Employees	The number of employees (in thousands). If missing, we use step extrapolation of EMP by firm and year.	Log(0.001 + *Employees*)
Profit margin	The ratio of net income to sales.	Winsorized at the 1% level
Basic earning power	The ratio of operating income after depreciation to total assets.	Winsorized at the 1% level
Inventory turnover	The ratio of cost of goods sold to inventory.	Winsorized at the 1% level
Intangibles to total assets	The ratio of total intangible assets to total assets.	Winsorized at the 1% level
ALS accruals	Allen, Larsen, and Sloan (2013) abnormal accruals.	Winsorized at -1 and 1
R&D to Sales	The ratio of research and development expenses to sales.	Winsorized at the 1% level
ROA	The ratio of net income to total assets.	Winsorized at the 1% level
Loss flag	Dummy equals 1 if net income is negative.	

Altman Z score	A generalized version of the Altman Z score, calculated as follows:
	Z’ = 3.25 + 6.56X1 + 3.26X2 + 6.72X3 + 1.05X4, Where:
	X1 = (ACT* - LCT*) / AT
	X2 = RE* / AT
	X3 = EBIT* / AT
	X4 = (AT - LT) / LT
		* 0 (zero) substituted for missing values

Altman Z distress flag	Dummy equals 1 if Altman Z score is less than 1.75.	
Number of business segments	The number of business segments as reported in SEG_ANNFUND in Compustat with STYPE = "BUSSEG" and SID ≠ 99. It is set to 1 if no number is reported.	Log(# *Business Segments*)
Segment concentration index	The Herfindahl-Hirschmann Index (HHI) applied to the percentage sales of the firm’s reported business segments. If no segments are reported for a given firm-year, we assume 1 segment.	Log(*Segment concentration index*)
Average distance to markets	The average distance from firm headquarters to its geographic segments' centroids, weighted by segment sales.	Log(*Average distance to markets*)
Herfindahl index	The HHI industry concentration index calculated using sales within 4-digit SIC codes.	
Number of geographic segments	The number of geographic segments reported.	Log(# *Geographic segments*)
Distance to regulator	The distance (in miles) between firm headquarters to the closest regional SEC office or U.S. Attorney office.	Log(*Distance to regulator*)
Auditor opinion flag	Dummy equals 1 if Auditor's Opinion variable (AUOP) is one of the following: ("0", "2", "3", "5").	
Big N auditor flag	Dummy equals 1 if Auditor variable (AU) is one of the following: ("1", "2", "3", "4", "5", "6", "7", "8").	
Industry flag	Dummy equals 1 if the firm's 4-digit SIC code falls in the Fama French 10-industry portfolio definitions.	

References

Agrawal, Anup, and Sahiba Chadha, 2005, Corporate governance and accounting scandals, *The Journal of Law & Economics* 48, 371–406.

Allen, Eric J., Chad R. Larson, and Richard G. Sloan, 2013, Accrual reversals, earnings and stock returns, *Journal of Accounting and Economics* 56, 113-129.

Altman, Eduard I., 1968, Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy, *Journal of Finance* 23, 589-609.

Amiram, Dan, Zahn Bozanic, James D. Cox, Quentin Dupont, Jonathan M. Karpoff, and Richard Sloan, 2018, Financial reporting fraud and other forms of misconduct: A multidisciplinary review of the literature, *Review of Accounting Studies 23*, 732-783.

Amiram, Dan, Zahn Bozanic, and Ethan Rouen, 2015, Financial statement errors: evidence from the distributional properties of financial statement numbers, *Review of Accounting Studies* 20, 1540–1593.

Amoa-Gyarteng K., 2014, Analyzing a listed firm in Ghana for early warning signs of bankruptcy and financial statement fraud: An empirical investigation of AngloGold Ashanti, *European Journal of Business and Management* 6(5), 10–17.

Association of Certified Fraud Examiners, 2018, Report to the nations: 2018 Global study on occupational fraud and abuse, available at https://s3-us-west-2.amazonaws.com/acfepublic/2018-report-to-the-nations.pdf.

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang, 2019, Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach, forthcoming, *Journal of Accounting Research*.

Bhavani, Ganga and Christian Amponsah, 2017, M-Score and Z-Score for detection of accounting fraud, *Journal of the Association for Accountancy & Business Affairs* 16, 68-86.

Beasley, Mark S., 1996, An empirical analysis of the relation between the board of director composition and financial statement fraud, *The Accounting Review* 7, 443–465.

Beatty, Anne, Scott Liao, and Jeff Jiewei Yu, 2013, The spillover effect of fraudulent financial reporting on peer firms' investments, *Journal of Accounting and Economics* 55(2-3),183-205.

Becker, Gary S., 1968, Crime and punishment: An economic approach. *Journal of Political Economy* 76, 169–217.

Bedi, Suneal, Catherine Schrand, and Eugene Soltes, 2019, Managerial proclivities to financially misreport, working paper.

Bell, Timothy B. and Joseph V. Carcello, 2000, A decision aid for assessing the likelihood of fraudulent financial reporting, *Auditing: A Journal of Practice & Theory* 19(1), 169-184.

Beneish, Messod D., 1997, Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance*, Journal of Accounting and Public Policy* 16, 271-309.

Beneish, Messod D., 1999, The detection of earnings manipulation, *Financial Analysts Journal* 55, 24–36.

Beneish, Messod and Patrick Vorst, 2022, The cost of fraud prediction errors, *The Accounting Review* (2022) 97(6), 91–121.

Boone, Audra, William Grieser, Qingqiu Li, and Parth Venkat, 2019, Product differentiation, benchmarking, and corporate fraud, available at SSRN: https://ssrn.com/abstract=3070375.

Buffett, Mary and David Clark, 2008, Warren Buffett and the Interpretation of Financial Statements: The Search for the Company with a Durable Competitive Advantage, New York, NY: Scribner.

Burns, Natasha, and Simi Kedia, 2006, The impact of performance-based compensation on misreporting, *Journal of Financial Economics* 79, 35–67.

Carslaw, Charles., 1988, Anomalies in income numbers: Evidence of goal oriented behavior, *The Accounting Review*, 63, 321–327.

Cecchini, Mark, Haldun Aytug, Gary J. Koehler, and Praveen Pathak, 2010, Detecting management fraud in public companies, *Management Science* 56, 1146-1160.

Chakrabarty, Bidisha, Pamela C. Moulton, Leo Pugachev, and Frank Wang, 2019, Catch me if you can: Improving the scope and accuracy of fraud prediction, working paper.

Chan, Canri and Steven P. Landry, 2019, Financial statements too good to be true? An instructional case assessing that question using analytical procedures and Beneish's M-Score, *Journal of Forensic and Investigative Accounting*, 11(2), 380-394.

Choi, Hae Mi, Xiaoxia Lou, Jonathan M. Karpoff, and Gerald S. Martin, 2023, Enforcement waves and spillovers, *Management Science*, forthcoming.

Dechow, Patricia M., Weili Ge, Chad R. Larson, and Richard G. Sloan, 2011, Predicting material accounting misstatements, *Contemporary Accounting Research* 28, 17–82.

Dechow, Patricia M., Richard G. Sloan, and Amy P. Sweeney, 1996, Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC, *Contemporary Accounting Research* 13, l-36.

Drabkova, Zita, 2015, Analysis of possibilities of detecting the manipulation of financial statements in terms of the IFRS and Czech accounting standards, Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 63(6), 1859–1866.

Duffie, Darrell and David Lando, 2001, Term structures of credit spreads with incomplete accounting information, *Econometrica* 69, 633-664.

Dupont, Quentin, 2023, Households' Financial Fallout From Institutional Scandals: Evidence From The U.S. Catholic Clergy Abuse Crisis. Georgetown McDonough School of Business Research Paper No. 3875648, Available at https://ssrn.com/abstract=3875648.

Durtschi, Cindy, William Hillison, and Carl Pacini, 2004, The effective use of Benford's law to assist in the detecting of fraud in accounting data, *Journal of Forensic Accounting* 5, 17–34.

Dyck, Alexander, Adair Morse, and Luigi Zingales, 2023, How pervasive is corporate fraud? Review of Accounting Studies https://doi.org/10.1007/s11142-022-09738-5.

Ferrell, Allen and Atanu Saha, 2011, Forward-casting 10b-5 damages: A comparison to other methods, Harvard Law and Economics Discussion Paper, available at https://ssrn.com/abstract=1811068.

Files, Rebecca, Gerald S. Martin and Stephanie Rasmussen, 2019, Regulator-cited cooperation credit and firm value: Evidence from enforcement actions, *The Accounting Review* 94(4), 275-302.

Fishman, Michael J. and Kathleen M. Hagerty, 1992, Insider trading and the efficiency of stock prices, *The RAND Journal of Economics* 23, 106 – 122.

Giannetti, Mariassunta and Tracey Yue Wang, 2016, Corporate scandals and household stock market participation, *Journal of Finance* 71, 2591-2636.

Graham, John R., Campbell R. Harvey, and Shiva Rajgopal, 2005, The economic implications of corporate financial reporting, *Journal of Accounting and Economics* 40, 3-73.

Graham, John R., Si Li, and Jiaping Qiu, 2008, Corporate misreporting and bank loan contracting, *Journal of Financial Economics* 89(1), 44-61.

Green, David and John Swets, 1966, *Signal detection theory and psychophysics*. New York, NY: John Wiley and Sons Inc., ISBN 0-471-32420-5.

Gurun, Umit G ., Noah Stoffman, and Scott E Yonker, 2018, Trust Busting: The Effect of Fraud on Investor Behavior, *The Review of Financial Studies* 31(4), 1341–1376.

Joseph, Lawrence, Theresa W. Gyorkos, and Louis Coupal, 1995, Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard, *American Journal of Epidemiology* 141, 263–72.

Hosmer, David W. and Stanley Lemeshow, 2000, *Applied Logistic Regression*, second edition, Wiley Series in Probability and Statistics (New York).

Karpoff, Jonathan M., Allison Koester, D. Scott Lee, and Gerald S. Martin, 2017, Proxies and databases in financial misconduct research, *The Accounting Review* 92, 129-163.

Karpoff, Jonathan M., D. Scott Lee, and Gerald S. Martin, 2008a, The cost to firms of cooking the books. *Journal of Financial and Quantitative Analysis* 43, 581-612.

Karpoff, Jonathan M., D. Scott Lee, and Gerald S. Martin, 2008b, The consequences to managers for financial misrepresentation, *Journal of Financial Economics* 88, 193-215.

Kedia, Simi and Thomas Philippon, 2009, The economics of fraudulent accounting, *Review of Financial Studies*, 22(6), 2169-2199.

Kedia, Simi and Shiva Rajgopal, 2011, Do the SEC's enforcement preferences affect corporate misconduct? *Journal of Accounting and Economics*, 51(3), 259-278.

Larcker, David and Anastasia Zakolyukina, 2012, Detecting deceptive discussions in conference calls, *Journal of Accounting Research* 50, 495-540.

Larson, Chad, Richard Sloan, and Jenny Zha Giedt, 2018, Defining, measuring, and modeling accruals: A guide for researchers, *Review of Accounting Studies* 23, 827-781.

Lawson, Bradley P., Gerald S. Martin, Leah Muriel, and Michael S. Wilkins, 2019, How do auditors respond to FCPA risk?, *Auditing: A Journal of Practice & Theory* 38(4), 177-200.

Lou, Yung-I. and Ming-Long Wang, 2009, Fraud risk factor of the fraud triangle assessing the likelihood of fraudulent financial reporting, *Journal of Business & Economics Research*, 7(2).

MacCarthy, John, 2017, Using Altman Z-score and Beneish M-score models to detect financial fraud and corporate failure: A case study of Enron Corporation, *International Journal of Finance and Accounting* 6(6), 159-166.

Mahama, Muntari, 2015, Detecting corporate fraud and financial distress using the Altman and Beneish models, *International Journal of Economics, Commerce and Management* 3(1):1–18.

Mehta Anupam and Ganga Bhavani, 2017, Application of forensic tools to detect fraud: The case of Toshiba, *Journal of Forensic and Investigative Accounting* 9(1), 692–710.

Nigrini, Mark, 2012, *Benford's law: Applications for forensic accounting, auditing, and fraud detection*. Hoboken, N.J.: Wiley.

Ofori, Edmond, 2016, Detecting corporate financial fraud using modified Altman Z-Score and Beneish M-Score. The case of Enron Corp, *Research Journal of Finance and Accounting* 7(4), 59–65.

Ozcan, Ahmet, 2016, Firm characteristics and accounting fraud: A multivariate approach, *Journal of Accounting, Finance and Auditing Studies* 2(2), 128-144.

PCAOB, AS 2401: Consideration of fraud in a financial statement audit, https://pcaobus.org/Standards/Auditing/Pages/AS2401.aspx.

Perols, Johan, 2011, Financial statement fraud detection; An analysis of statistical and machine learning algorithms, *Auditing: A Journal of Practice & Theory* 30, 19-50.

Perols, Johan, Robert Bowen, Carsten Zimmerman, and Basamba Samba, 2017, Finding needles in a haystack: Using data analytics to improve fraud prediction, *The Accounting Review* 92, 221-245.

Persons, Obeua S., 2011, Using financial statement data to identify factors associated with fraudulent financial reporting, *Journal of Applied Business Research* 11(3), 38-46.

Rajan, Raghuram G. and Luigi Zingales, 1998, Financial dependence and growth, *American Economic Review* 88, 559-586.

Richardson, Scott A., Richard G. Sloan, Mark T. Soliman, and Irem Tuna, 2006, The implications of accounting distortions and growth for accruals and profitability, *The Accounting Review* 81, 713-743.

Sadka, Gil, 2006, The economic consequences of accounting fraud in product markets: Theory and a case from the US telecommunication industry (WorldCom), *American Law and Economics Review* 8, 439–475.

Simpson, Caesar, 2016, An examination of Enron Corporation: Could the fraud have been detected sooner? *International Journal of Scientific and Research Publications,* 6(7), 64-78.

Soltes, Eugene F., 2019, The frequency of corporate misconduct: Public enforcement versus private reality, *Journal of Financial Crime* 26(4), 923-937.

Swets, John, 1988, Measuring the accuracy of diagnostic systems, *Science* 240, 1285-1293.

Taherinia, Masoud and Reza Talebi, 2019, Ability of Fraud Triangle, Fraud Diamond, Beneish M Score, and Altman Z Score to predict financial statements fraud, *Journal of Economic and Social Research,* 18(2), 213-226.

Thomas, Jacob, 1989, Unusual patterns in reported earnings, *The Accounting Review* 5, 773–787.

Wang, Tracy Yue, 2013, Corporate securities fraud: Insights from a new empirical framework, *Journal of Law, Economics, and Organization* 29, 535-568.

Wang, Tracy Yue and Winton, Andrew, 2012, Competition and corporate fraud waves, available at https://ssrn.com/abstract=1783752.

Wiginton, J., 1980, A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15(3), 757-770.

White, Roger M., 2020, Insider trading: What really protects U.S. investors? *Journal of Financial and Quantitative Analysis* 55(4) 1305-1322.

Yang, Liu and Zhu, Min, 2022, Restatement Prediction with Detection Lag (Feb 28, 2022). Available at https://ssrn.com/abstract=4045172.

Yin, C., Cheng, X., Yang, Y. , and Palmon, D., 2021, Do Corporate Frauds Distort Suppliers' Investment Decisions?. *Journal of Business Ethics* 172, 115–132.

Zakolyukina, Anastasia A., 2018, How common are intentional GAAP violations? Estimates from a dynamic model, *Journal of Accounting Research* 56, 5–44.

Zhou, Xiao-Hua, Pete Castelluccio, and Chuan Zhou, 2004, Non-parametric estimation of ROC curves in the absence of a gold standard (July 2004). UW Biostatistics working paper, http://biostats.bepress.com/uwbiostat/paper231.

Zou, Kelly, A. James O'Malley, and Laura Mauri, 2007, Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models, *Circulation* 115, 654-657. http://circ.ahajournals.org/content/115/5/654.full.

Figure 1. ROC curves for previous prediction models

The ROC curve plots the sensitivity (true positive rate) as a function of the false positive rate (1-specificity) for all possible probability threshold levels. In this figure we plot the ROC curves corresponding to the models estimated in Table 6: Beneish (1999), Dechow et al. (DGLS, 2011), Benford Law's *MAD score*, and Cecchini et al.'s (2010) key predictive measures. "Area" reports the area under the curve (AUC) for each model.

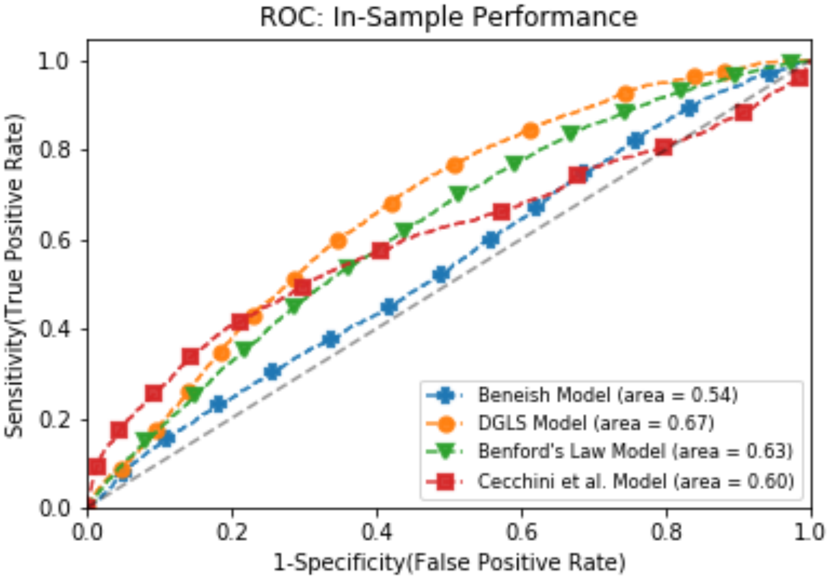
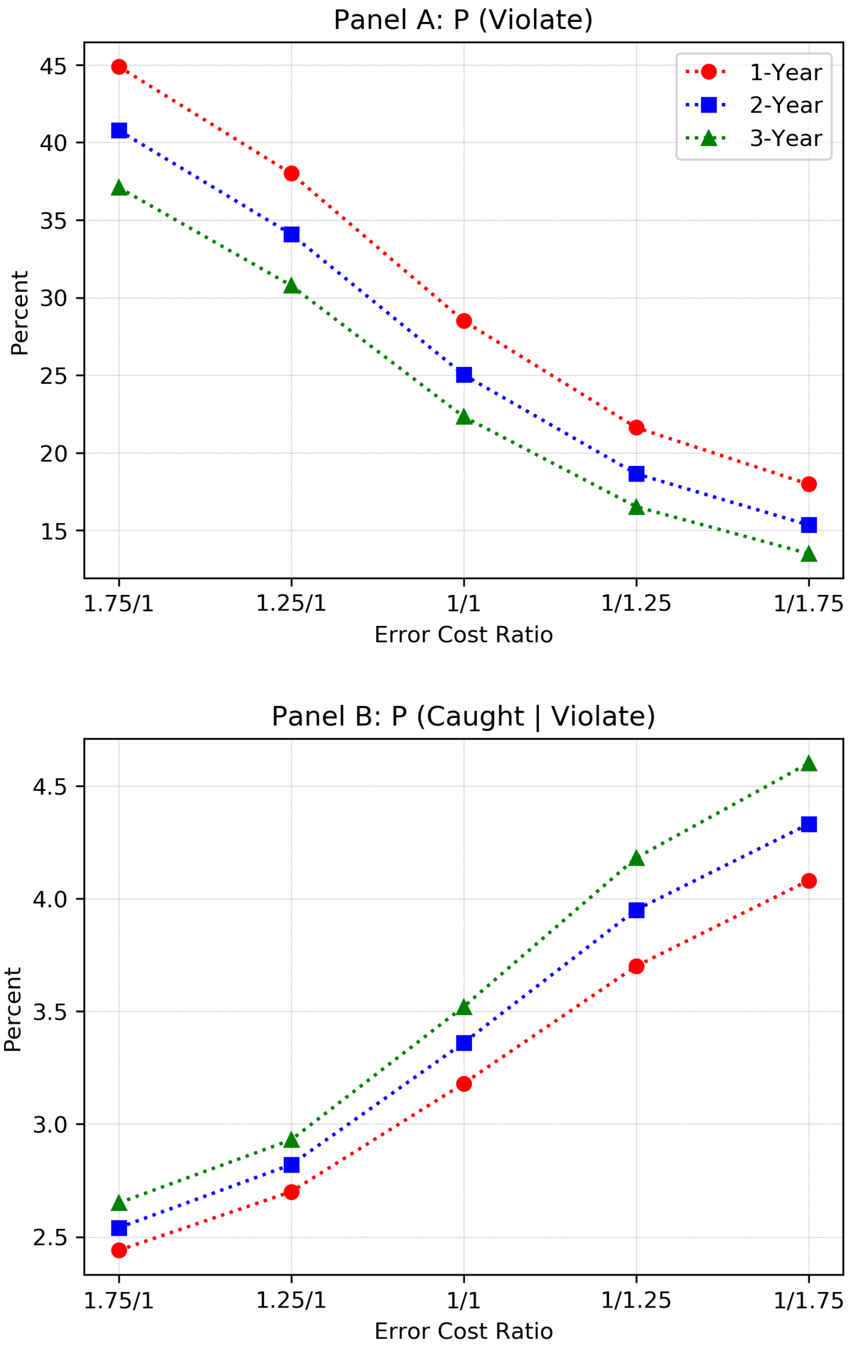


Figure 2. Sensitivity of model predictions to error cost ratios and violation durations

Panel A graphs the unconditional probability of a violation (P(Violate)) in a given firm-year implied by the prediction model summarized in column 2 of Table 7, for different error cost ratios and assumed violation durations. Panel B graphs the unconditional probability of a firm-year violation resulting in regulatory sanctions (i.e., "getting caught") conditional on a violation (P(Caught|Violate).


Panel A: P (Violate)

Panel B: P (Caught | Violate)

Figure 3. Sensitivity of probability estimates to changes in the error cost ratio and violation duration

Panel A shows how the unconditional probability of a violation (P(Violate)) in a given firm-year changes with different assumptions about the error cost ratio and violation duration, as implied by the top five performing prediction models summarized in Table 8 (our base logistic model and four machine learning models). Panel B shows how P(Caught|Violate) changes with the assumed error cost ratio and violation duration. P(Caught|Violate) is the unconditional probability of a firm-year violation resulting in regulatory sanctions (i.e., “getting caught”) conditional on a violation.

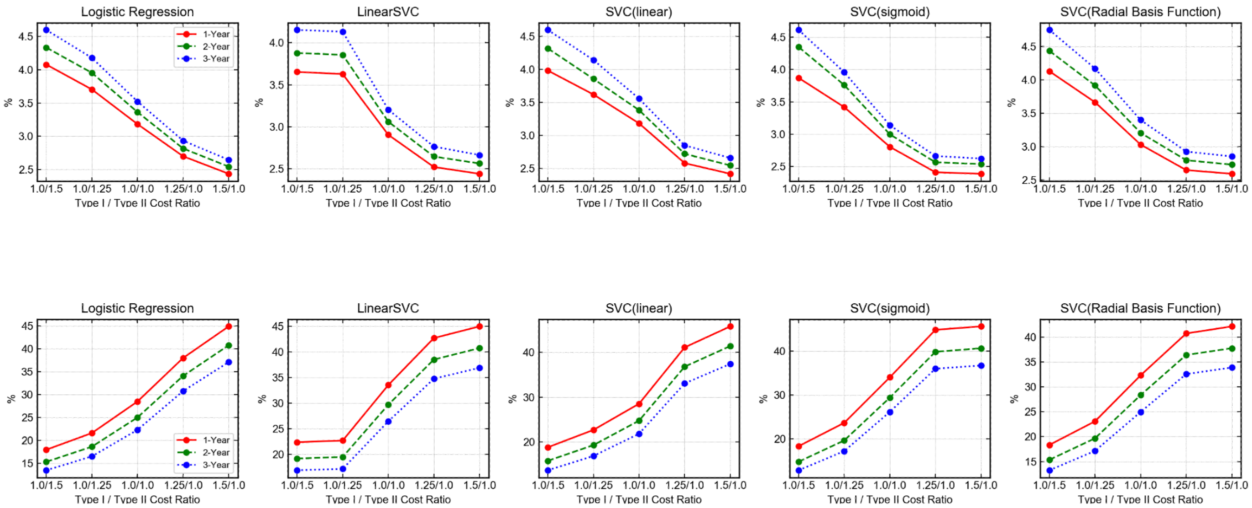


Figure 4. In-sample and out-of-sample model results, annual rolling model

This figure plots in-sample and out-of-sample model results as measured by the area under the ROC curve (AUC) for tests in which we truncate the data by successive years. Results are based on the final model in column 2 of Table 7. To conduct out-of-sample tests, we partition the data into different training and holdout subsamples, each time estimating the model on the training sample and evaluating its performance in the holdout sample. For the 1990 results, we estimate the model using firm-year observations through 1990 and test the model performance using post-1990 data. In successive iterations, we extend the training data period by one year, stopping when the training data include all observations through 2014.

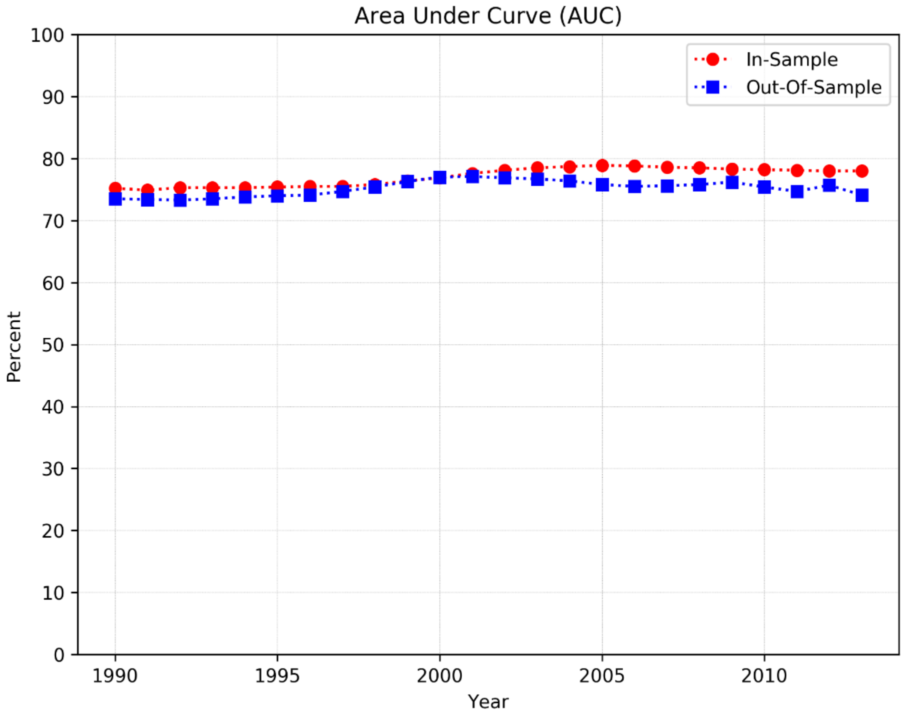


Figure 5. K-fold out-of-sample estimators for our prediction model

This figure reports results of k-fold estimators from our final prediction model in column 2 of Table 7 using a three-year violation duration. We randomly partition our data into different training and holdout subsamples, each time estimating the model on the training sample and evaluating its performance in the holdout sample. We repeat this process six times. The figure plots results from each trial, the mean across trials, and +/- 1 standard deviation bounds.

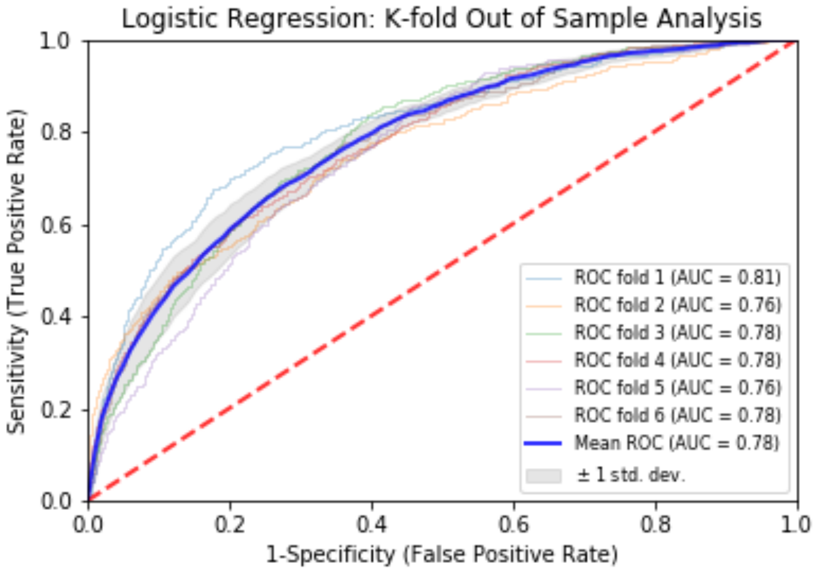


Table 1. Distribution of the sample of initiating firms and violation firm-years, by year

This table reports summary statistics for the sample of 1,069 firms targeted for enforcement action for financial misrepresentation by the SEC and/or DOJ under provisions of Section 13(b) of the Securities and Exchange Act of 1934 from 1978 through 2017, for misconduct occurring from 1976 through 2014. The sample includes 1,039 unique firms, 30 of which have two events each in the sample. The average violation lasts 3.1 years and the number of ongoing violations includes newly initiated violations and ongoing violations that began in a prior year.

Misconduct Year	# of firms initiating violations	# Compustat firms	Violation initiators (% of Compustat	Number of ongoing violations	Ongoing violations (% of Compustat)	Number of ongoing violations in our final	Number of Compustat firms in our final model
1976	23	6,751	0.34%	23	0.34%	18	3,196
1977	4	6,788	0.06%	26	0.38%	21	3,150
1978	3	6,761	0.04%	22	0.33%	16	3,216
1979	5	6,900	0.07%	19	0.28%	15	3,273
1980	15	6,889	0.22%	28	0.41%	17	3,302
1981	18	7,058	0.26%	37	0.52%	24	3,516
1982	18	7,685	0.23%	44	0.57%	28	3,583
1983	18	7,996	0.23%	44	0.55%	20	3,629
1984	18	8,243	0.22%	39	0.47%	18	3,760
1985	21	8,599	0.24%	38	0.44%	13	3,766
1986	19	9,017	0.21%	43	0.48%	21	3,744
1987	15	9,225	0.16%	37	0.40%	22	3,913
1988	23	9,333	0.25%	45	0.48%	22	3,962
1989	37	9,343	0.40%	66	0.71%	32	3,960
1990	14	9,571	0.15%	62	0.65%	29	3,911
1991	32	9,966	0.32%	65	0.65%	40	3,953
1992	33	10,705	0.31%	67	0.63%	39	4,061
1993	35	11,482	0.30%	68	0.59%	32	4,361
1994	30	11,898	0.25%	68	0.57%	44	4,692
1995	28	12,491	0.22%	70	0.56%	47	5,212
1996	32	12,623	0.25%	82	0.65%	56	5,441
1997	47	12,437	0.38%	97	0.78%	65	5,849
1998	60	12,554	0.48%	128	1.02%	86	5,855
1999	72	12,530	0.57%	184	1.47%	122	5,594
2000	91	12,092	0.75%	231	1.91%	157	6,029
2001	64	11,583	0.55%	250	2.16%	186	5,963
2002	36	11,251	0.32%	211	1.88%	159	5,783
2003	35	11,080	0.32%	198	1.79%	162	5,655
2004	21	10,925	0.19%	156	1.43%	129	5,592
2005	25	11,051	0.23%	139	1.26%	113	5,559
2006	26	11,125	0.23%	122	1.10%	101	5,522
2007	19	11,152	0.17%	102	0.91%	85	5,384
2008	32	10,966	0.29%	98	0.89%	79	5,212
2009	35	10,922	0.32%	106	0.97%	84	5,050
2010	21	11,122	0.19%	92	0.83%	78	4,898
2011	17	11,746	0.14%	73	0.62%	60	4,843
2012	19	11,755	0.16%	66	0.56%	53	4,774
2013	7	11,523	0.06%	46	0.40%	32	4,727
2014	1	10,919	0.01%	19	0.17%	14	4,780
Total	1,069	396,057	0.27%	3,311	0.84%	2,339	178,670

Table 2. Top 50 model-implied violation firm-years

This table illustrates the output from the prediction modeling process by listing the 50 firm-years within our sample with the highest probability of misrepresentation as implied by our base-case prediction model. A total of 15 of these firm-years were identified in subsequent regulatory enforcement actions as having misrepresented financial statements.

Rank	Company name	Year	Probability	Sanctioned by SEC/DOJ?
1	VIAVI SOLUTIONS INC	2000	0.516	No
2	TIME WARNER INC	2001	0.435	Yes
3	AT HOME CORP	1999	0.373	No
4	ARIBA INC	2000	0.363	No
5	VERISIGN INC	2000	0.346	No
6	SANOFI	2004	0.309	No
7	CISCO SYSTEMS INC	2000	0.302	No
8	NORTEL NETWORKS CORP	1999	0.300	No
9	NORTEL NETWORKS CORP	2000	0.297	Yes
10	APPLE INC	2011	0.294	No
11	GENERAL ELECTRIC CO	2007	0.290	No
12	AVIS BUDGET GROUP INC	1997	0.288	Yes
13	GENERAL ELECTRIC CO	2006	0.287	No
14	AUTONATION INC	1996	0.287	No
15	CISCO SYSTEMS INC	1999	0.273	No
16	COMPAQ COMPUTER CORP	1998	0.269	No
17	CISCO SYSTEMS INC	2001	0.267	No
18	GENERAL ELECTRIC CO	2005	0.265	Yes
19	ALCATEL-LUCENT	2000	0.260	Yes
20	FREESCALE SEMICONDUCTOR INC	2006	0.256	No
21	DELHAIZE AMERICA INC	2001	0.256	No
22	VERITAS SOFTWARE CORP	1999	0.249	No
23	SORRENTO NETWORKS CORP	1996	0.249	No
24	LUCENT TECHNOLOGIES INC	1999	0.243	No
25	TELEFONAKTIEBOLAGET LM ERICS	1999	0.243	No
26	LVMH MOET HENNESSY LOUIS V	2010	0.241	No
27	APPLE INC	2010	0.241	No
28	INTL BUSINESS MACHINES CORP	2011	0.240	No
29	INTL BUSINESS MACHINES CORP	2001	0.237	Yes
30	NOKIA CORP	2000	0.237	No
31	NOKIA CORP	2007	0.228	No
32	NEC CORP	1998	0.226	No
33	HP INC	2002	0.223	Yes
34	TELEFONAKTIEBOLAGET LM ERICS	2000	0.221	No
35	MONSTER WORLDWIDE INC	1999	0.221	Yes
36	INTL BUSINESS MACHINES CORP	2000	0.220	Yes
37	GENERAL ELECTRIC CO	2008	0.220	No
38	NOKIA CORP	1999	0.219	No
39	INTL BUSINESS MACHINES CORP	1999	0.219	Yes
40	INTL BUSINESS MACHINES CORP	2003	0.218	Yes
41	VODAFONE GROUP PLC	1999	0.217	No
42	SIEMENS AG	2007	0.217	Yes
43	INTL BUSINESS MACHINES CORP	2004	0.217	Yes
44	GENERAL ELECTRIC CO	2010	0.216	No
45	AVENTIS SA	2000	0.216	No
46	TYCO INTERNATIONAL PLC	2001	0.216	Yes
47	GENERAL ELECTRIC CO	2009	0.215	No
48	CANON INC	2006	0.213	No
49	EMC CORP/MA	2000	0.213	No
50	I2 TECHNOLOGIES INC	2000	0.212	Yes

Table 3. Size distribution of violating firms

This table reports the distribution of firm size, by deciles of firms listed on Compustat, for the 3,117 firm-years targeted for enforcement action for financial misrepresentation by the SEC and/or DOJ under provisions of Section 13(b) of the Securities and Exchange Act of 1934 from 1978 through 2017, and for which sufficient Compustat data are available to compute firm size. (The sample of misrepresentation firm-years includes an additional 194 firm-years for which market capitalization information is not available on Compustat.) Firm size is measured as market capitalization of the firms' outstanding common shares as of the end of the year before the violation firm-year.

Decile	**Freq.**	**Percentage**
Small	78	2.5%
2	140	4.5%
3	162	5.2%
4	172	5.5%
5	201	6.4%
6	248	8.0%
7	297	9.5%
8	333	10.7%
9	442	14.2%
Large	1044	33.5%
Overall	3117	100.0%

* 194 firm-years have missing market cap information

Table 4. Distribution of violation firm-years by industry

This table reports the distribution by industry of the sample of 3,311 violation firm-years from 1976-2014 for firms targeted for enforcement action for financial misrepresentation by the SEC and/or DOJ under provisions of Section 13(b) of the Securities and Exchange Act of 1934 from 1978 through 2017. Industry classifications are based on the Fama French 10-industry portfolio definitions.

FF Industries	Compustat Frequency	Percent of Population	Violation Frequency	Percent of Violations	% Violating in Industry
Nondurables	20,285	5.1%	189	5.7%	0.93%
Durables	9,170	2.3%	84	2.5%	0.92%
Manufacturing	46,544	11.8%	448	13.5%	0.96%
Energy	20,999	5.3%	135	4.1%	0.64%
Business Equipment	58,997	14.9%	791	23.9%	1.34%
Telecom	11,911	3.0%	74	2.2%	0.62%
Shops	35,275	8.9%	352	10.6%	1.00%
Healthcare	30,063	7.6%	348	10.5%	1.16%
Utilities	13,666	3.5%	57	1.7%	0.42%
Other	149,147	37.7%	833	25.2%	0.56%
Total	396,057	100.0%	3,311	100.0%	0.84%

Table 5. Summary statistics for firm characteristics associated with violations

This table reports summary statistics for the variables used to construct our prediction model for financial misrepresentation. Panels A through D report variables used in four prior models that can be used to predict financial misconduct, and Panel E reports additional variables that we consider in our tests. Variable definitions are in the Appendix. Column 1 reports on the mean value of the firm characteristic during a violation firm-year and column 2 reports the number of firm-years with data available to calculate the mean. Columns 3 and 4 report values for non-violation firm-years for the violation firms. Column 5 reports a t-statistic comparing means in columns 1 and 3. Columns 6 and 7 report values for all firm-years for firms that are not targeted for enforcement action for misrepresentation. Column 7 reports a t-statistic comparing means in columns 1 and 6.

	Violating firms							
	Violation years		Non-violation years			Non-violating firms		
	Mean	N	Mean	N	t-stat	Mean	N	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Beneish (1999) model variables								
Days sales in receivables index	1.08	3,009	1.09	13,034	-0.72	1.11	268,420	-2.50
Gross margin index	1.02	3,056	0.98	13,442	2.41	0.98	280,406	2.10
Asset quality index	2.12	3,130	1.81	13,781	2.61	2.08	298,677	0.31
Depreciation index	1.06	3,038	1.04	13,263	1.71	1.05	282,367	0.63
Sales growth index	1.34	3,067	1.20	13,497	9.51	1.24	282,686	6.63
Leverage index	1.13	3,120	1.10	13,747	1.93	1.16	297,617	-2.45
Total accruals/total assets	-0.08	3,240	-0.08	14,530	0.95	-0.11	324,324	4.36
SG&A index	1.04	2,588	1.03	11,395	1.52	1.04	221,175	0.45
Panel B: Dechow et al. (2011) variables								
RSST accruals	0.05	3,037	0.02	13,308	4.94	0.02	284,450	5.38
Change in receivables	0.03	3,134	0.02	13,803	7.00	0.02	299,460	9.32
Change in inventory	0.02	3,134	0.01	13,803	6.09	0.01	299,460	10.69
% Soft assets	0.64	3,240	0.61	14,530	6.58	0.54	324,324	18.67
Change in cash sales	0.33	3,001	0.15	13,147	7.33	0.14	278,300	7.05
Change in ROA	-0.01	2,965	-0.01	13,136	-1.26	-0.01	274,787	-1.05
Abnormal change in employees	-0.16	2,995	-0.09	13,191	-6.25	-0.09	259,675	-6.64
Operating lease flag	0.80	3,311	0.72	15,142	9.29	0.55	377,604	29.41
Security issue flag	0.90	3,311	0.82	15,142	12.28	0.66	377,604	29.58
Panel C: Benford's law variable								
MAD score	0.03	3,241	0.04	14,566	-6.34	0.04	325,178	-21.58
Panel D: Cecchini et al. (2010) variables								
Lag sales/ lag preferred stock	58.12	3,088	40.59	13,554	5.37	9.82	291,944	36.39
SG&A/investments and adv.	1.74	2,523	1.65	11,443	0.51	0.56	246,696	9.18
Lag total assets/ lag inv. and adv.	14.53	2,870	10.16	12,650	2.02	3.49	281,524	11.29
Lag sales/lag inv. and adv.	10.36	2,856	8.27	12,604	2.39	2.87	279,889	12.62
Total assets/short-term inv.	20.09	3,085	12.37	13,419	5.80	8.94	298,609	3.38

Table 5 (continued)

	Violating firms							
	Violation years		Non-violation years			Non-violating firms		
	Mean	N	Mean	N	t-stat	Mean	N	t-stat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel E: Additional firm characteristics								
Market cap	6.51	3,117	5.74	13,356	13.85	4.43	301,978	48.18
Market to book	2.19	3,240	1.67	14,530	10.00	1.78	324,324	7.52
Leverage	0.60	3,240	0.70	14,530	-7.28	0.70	324,324	-6.06
Employees	0.71	3,311	0.37	15,142	6.28	-1.79	377,604	45.62
Profit margin	-0.41	3,182	-0.50	14,211	1.26	-0.87	305,735	5.34
Basic earning power	0.00	3,240	-0.04	14,530	3.25	-0.11	324,324	8.94
Inventory turnover	11.40	3,311	10.00	15,142	2.90	10.61	377,604	1.59
Intangibles to total assets	0.13	3,240	0.09	14,530	11.65	0.07	324,324	21.93
ALS accruals	-0.01	2,988	0.00	13,041	-1.53	0.00	270,076	-1.11
R&D to sales	0.06	3,182	0.05	14,211	2.70	0.06	305,735	-1.39
ROA	-0.08	3,240	-0.10	14,530	2.23	-0.18	324,324	6.94
Loss flag	0.32	3,311	0.33	15,142	-1.08	0.45	377,604	-15.84
Altman Z-score	-59.88	3,238	-50.93	14,522	-0.11	-56.30	323,940	-0.09
Altman's Z distress flag	0.10	3,311	0.15	15,142	-7.76	0.16	377,604	-9.55
Number of business segments	1.29	3,311	1.13	15,142	11.76	0.72	377,604	45.88
Segment concentration index	-0.10	3,243	0.40	14,587	-7.94	1.73	325,643	-24.34
Average distance to markets	4.22	3,241	3.50	14,566	8.33	1.63	325,177	38.01
Herfindahl 4 index	0.22	3,182	0.22	14,211	-2.33	0.21	305,737	2.42
Number of geographic segments	0.65	3,311	0.52	15,142	8.61	0.21	377,604	48.52
Distance to regulator	3.15	3,311	3.01	15,142	3.06	3.19	377,604	-0.94
Auditor opinion flag	0.03	3,311	0.03	15,142	-0.97	0.04	377,604	-1.84
Big N auditor flag	0.74	3,311	0.73	15,142	1.97	0.59	377,604	18.14
Fama French industry dummies:								
Nondurables	0.06	3,311	0.05	15,142	1.46	0.05	377,604	1.53
Durables	0.03	3,311	0.03	15,142	-0.80	0.02	377,604	0.93
Manufacturing	0.14	3,311	0.13	15,142	0.36	0.12	377,604	3.31
Energy	0.04	3,311	0.05	15,142	-1.68	0.05	377,604	-3.21
Business Equipment	0.24	3,311	0.22	15,142	2.11	0.15	377,604	15.20
Telecom	0.02	3,311	0.02	15,142	-0.15	0.03	377,604	-2.70
Shops	0.11	3,311	0.12	15,142	-1.51	0.09	377,604	3.73
Healthcare	0.11	3,311	0.09	15,142	1.93	0.07	377,604	6.56
Utilities	0.02	3,311	0.03	15,142	-3.40	0.03	377,604	-5.54
Other	0.25	3,311	0.26	15,142	-0.86	0.38	377,604	-15.43

Table 6. Replications based on four prior prediction models

This table reports the results of our replications of four previous financial misreporting prediction models using our data on 3,311 firm-years in which firms misrepresented their financial statements as determined by subsequent enforcement action for misrepresentation by the SEC and/or DOJ. Panel A reports an exact replication of Beneish's (1998) model. Panel B reports an exact replication of Model 2 in Dechow et al. (2011). Panel C reports results for a prediction model that uses the *MAD score* which is based on Benford's Law. Panel D reports results for a prediction model based on the five most predictive measures used by Cecchini et al. (2010) to predict financial fraud. Each panel reports on the number of firm-years used to estimate the model which is affected by the availability of data for the covariates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A. Replication of the Beneish (1999) model

	Coef.	St.Err	t-value	Sig.
Days sales in receivables index	-0.068	0.026	-2.64	***
Gross margin index	0.058	0.025	2.29	**
Asset quality index	-0.003	0.003	-0.99	
Depreciation index	0.05	0.033	1.54	
Sales growth index	0.162	0.025	6.55	***
Leverage index	-0.007	0.03	-0.22	
Total accruals/total assets	0.193	0.098	1.96	**
SG&A index	0.222	0.05	4.44	***
Constant	-4.916	0.114	-42.96	***

N= 223,984

Panel B. Replication of the Dechow et al. (2011) F-score model

	Coef.	St.Err	t-value	Sig.
RSST accruals	0.243	0.056	4.36	***
Change in receivables	0.595	0.242	2.46	**
Change in inventory	0.948	0.33	2.87	***
% Soft assets	1.368	0.151	9.04	***
Change in cash sales	0.062	0.013	4.91	***
Change in ROA	-0.376	0.075	-5.01	***
Abnormal change in employees	-0.132	0.035	-3.79	***
Operating lease flag	0.682	0.102	6.65	***
Security issue flag	1.13	0.107	10.52	***
Constant	-6.906	0.161	-42.88	***

N = 259,216

Table 6 (continued)

Panel C. Replication using the MAD score (based on Benford's Law)

	Coef.	St.Err	t-value	Sig.
MAD score	-31.699	2.364	-13.41	***
Constant	-3.489	0.098	-35.72	***

N = 342,985

Panel D: Replication based on Cecchini et al.'s (2010) key predictive measures

	Coef.	St.Err	t-value	Sig.
Lag sales/ lag preferred stock	0.0014	0.0004	3.89	***
SG&A/investments and advances	-0.0126	0.0103	-1.23	
Lag total assets/ lag inv. and adv.	0.0034	0.0015	2.26	**
Lag sales/lag inv. and adv.	0.0006	0.0022	0.28	
Total assets/short-term inv.	-0.0004	0.0006	-0.69	
Constant	-4.5784	0.0454	-100.93	***

N= 217,189

Table 7. Comprehensive logistic prediction model

This table reports the coefficients from logistic regressions based on panel data for all Compustat-listed firms from 1976-2014. The dependent variable equals one if the firm-year is part of the violation period as indicated by the SEC's releases for enforcement actions initiated from 1978-2017 that target financial misrepresentation during the 1976-2014 period. The first model includes all of 50 potential predictor variables summarized in Table 5 (excluding the five Cecchini et al. (2010) variables), which include all the variables used in the prediction models by Beneish (1999) and Dechow et al. (2011), the *MAD score* (based on Benford's Law), and additional firm characteristics as described in the paper and the Appendix. The second model is our final prediction model. It is derived by backward elimination beginning with Model 1, in which covariates with the highest p-values are successively eliminated until all remaining covariates have p-values less than 0.05. Each model is reported in three panels, and each cell reports the coefficient and p-value (in parentheses).

(Continued on following page)

Table 7 (continued)

	Initial model (all variables)	Final model		Initial model (all variables)	Final model		Initial model (all variables)	Final model
	(1)	(2)		(1)	(2)		(1)	(2)
Beneish Variables:			***Benford's Law Variable***					
Days sales in receivables index	-0.0765		MAD score	-3.6453		Herfindahl 4 index	-0.4057	
	(0.110)			(0.179)			(0.113)	
Gross margin index	0.0821	0.0886				Number of geographic segments	0.1204	0.1253
	(0.055)	(0.036)	***Additional Variables***				(0.136)	(0.035)
Asset quality index	-0.0012		Market cap	0.3578	0.3923	Distance to regulator	-0.0258	
	(0.762)			(<0.001)	(<0.001)		(0.189)	
Depreciation index	0.0123		Market-to-book	0.0103		Auditor opinion flag	0.8191	0.8046
	(0.796)			(0.476)			(<0.001)	(<0.001)
Sales growth index	0.1247	0.1862	Leverage	-0.1142		Big N auditor flag	-0.6283	-0.6007
	(0.005)	(<0.001)		(0.348)			(<0.001)	(<0.001)
Leverage index	0.0448		Employees	0.0379		Fama-French Industry:		
	(0.291)			(0.167)		Nondurables	(base)	
Total accruals/total assets	0.0620		Profit margin	0.0044				
	(0.739)			(0.769)		Durables	-0.1832	
SG&A index	0.3729	0.3800	Basic earning power	-0.1425			(0.537)	
	(<0.001)	(<0.001)		(0.447)		Manufacturing	0.0334	
			Inventory turnover	0.0018			(0.877)	
DGLS Variables:				(0.186)		Energy	0.0110	
RSST accruals	0.0851		Intangibles to total assets	-0.2195			(0.970)	
	(0.437)			(0.362)		Business equipment	0.4546	0.3239
Change in receivables	0.8964	0.8102	ALS accruals	-0.4400	-0.4552		(0.026)	(0.002)
	(0.046)	(0.030)		(0.001)	(<0.001)	Telecom	-0.5575	-0.7688
Change in inventory	2.0463	2.0788	R&D to sales	0.1434			(0.105)	(0.009)
	(<0.001)	(<0.001)		(0.676)		Shops	0.2267	
% Soft assets	1.7916	1.7154	ROA	-0.1847			(0.314)	
	(<0.001)	(<0.001)		(0.238)		Healthcare	0.3843	
Change in cash sales	0.0317		Loss flag	0.3936	0.4153		(0.108)	
	(0.159)			(<0.001)	(<0.001)	Utilities	(empty)	
Change in ROA	-0.0192		Altman Z-Score	0.0017				
	(0.902)			(0.373)		Other	0.2506	
Abnormal change in employees	-0.0785		Altman's Z distress flag	-0.1593			(0.227)	
	(0.100)			(0.237)				
Operating lease flag	0.3249	0.3420	Number of business segments	0.1236		Constant	-8.5553	-8.8500
	(0.017)	(0.007)		(0.220)			(<0.001)	(<0.001)
Security issue flag	0.4992	0.4972	Segment concentration index	-0.1129	-0.1304			
	(<0.001)	(<0.001)		(<0.001)	(<0.001)	N	178,670	178,670
			Avg distance to mkts	0.0025		p	(<0.001)	(<0.001)
				(0.857)		AUC	0.7857	0.7813

Table 8: Performance of the base logistic model and 14 machine learning models

Summary metrics of the fifteen supervised machine learning classifiers using 100 randomized training/testing trials of classifiers from Scikit-learn 0.23.1 and Imbalanced-learn Python libraries the on the financial misrepresentation test sample. The table presents the average metrics of each machine learning classifier in a 100-trial bootstrap procedure that randomly split the dataset using a 60/40 split, training on the 60 percent training data and calculating fit performance on the 40 percent withheld test data. Hyperparameters were set for each classifier as presented in Table 7. The training data was used to calculate parameter estimates and the predicted probabilities or decision function scores used to determine the optimum threshold cut-point according to Youden (1950). The optimum cut-point threshold is the value for which the continuous forecast outputs from each classifier are mapped into a binary 0/1 classification. The coefficient estimates were applied to the withheld test data and the continuous forecast outputs were mapped into a 0/1 classification using the optimum cut-point threshold from the training data. The attributes were standardized before splitting the datasets using sklearn.preprocessing.StandardScaler. The metrics include: balanced accuracy (Bal Acc) which is the average of how well the classifier predicts each class; sensitivity (Sens) which is the percentage of samples predicted positive of the known condition positive samples; specificity (Spec) which is the percentage of samples predicted negative from the known condition negative samples; geometric mean (GMean) which is the geometric mean of sensitivity and specificity; Type I (false-positive) error rate is the percentage of samples predicted positive of the known condition negatives; Type II (false-negative) error rate is the percentage of samples predicted negative of the known condition positives; area under receiver operating characteristic curve (AUC) which is a measure of how well the classifier can distinguish between the classification values where 1.0 represents perfect classification and 0.5 no better than a random coin toss for classification; Rank which is the rank order of the classifiers based upon the aera under the receiver operating characteristic curve (AUC); and standard deviation (SD) which is the standard deviation of the AUC over the 100-trials as a measure of the consistency for each classifier.

Classifier	Bal Acc	Sens	Spec	GMean	Type I	Type II	AUC	SD	Rank
Logistic regression model (base case)	0.7073	0.7157	0.6989	0.7063	0.3011	0.2843	0.7797	0.0055	2
14 Machine learning models:									
Linear									
Stochastic gradient descent	0.6955	0.7059	0.6851	0.6942	0.3149	0.2941	0.7601	0.0114	7
Gaussian naive Bayes	0.6753	0.6171	0.7335	0.6725	0.2665	0.3829	0.7178	0.0089	13
Nearest Neighbors									
k-Nearest neighbor	0.5740	0.1792	0.9688	0.4165	0.0312	0.8208	0.5744	0.0051	15
Decision Trees									
Decision tree	0.6680	0.6634	0.6726	0.6651	0.3274	0.3366	0.7182	0.0074	12
Ensemble Methods									
Random forest	0.7007	0.6736	0.7277	0.6995	0.2723	0.3264	0.7654	0.0060	6
Extra trees	0.6957	0.6670	0.7245	0.6941	0.2755	0.3330	0.7585	0.0068	8
Random under-sampling boosting	0.6847	0.6855	0.6840	0.6826	0.3160	0.3145	0.7437	0.0374	10
Neural Network									
Multi-layer perceptron	0.6363	0.5904	0.6823	0.6320	0.3177	0.4096	0.6842	0.0103	14
Discriminate Analysis									
Quadratic discriminant analysis	0.6825	0.6294	0.7356	0.6798	0.2644	0.3706	0.7318	0.0092	11
Support Vector Machines									
LinearSVC	0.7069	0.7213	0.6925	0.7056	0.3075	0.2787	0.7796	0.0055	4
SVC(linear)	0.7074	0.7272	0.6875	0.7060	0.3125	0.2728	0.7797	0.0055	3
SVC(sigmoid)	0.7032	0.7288	0.6776	0.7011	0.3224	0.2712	0.7725	0.0057	5
SVC(polynomial)	0.7010	0.6941	0.7080	0.7001	0.2920	0.3059	0.7516	0.0071	9
SVC(radial basis function)	0.7103	0.7391	0.6814	0.7086	0.3186	0.2609	0.7832	0.0055	1

Table 9. Prediction model confusion matrix

This table reports the confusion matrix for our base case model. We use the parsimonious model from column 2 in Table 7 with 1/1 error cost ratio and a three-year violation duration. Panel A reports the outcomes as described in Section 4.2 for this scenario. Panel B illustrates calculation of specific outcomes from the model.

Panel A. Confusion matrix assuming an equal error cost ratio and 3-period violation duration

	Condition		
Predicted	Positive	Negative	Total
Positive	1,405	38,484	39,889
Negative	934	137,847	138,781
Total	2,339	176,331	178,670

Predicted positive if Pr(condition) >= 0.0137

Panel B. Illustration of model outcomes

Statistic	**Formula**	**Calculation**	**Value**
P(violate)	(TP+FP)/N	=39,889/178,670	22.3%
P(caught|violate)	TP/(TP+FP)	=1,405/39,889	3.5%
Sensitivity	TP/(TP+FN)	=1,405/2,339	60.1%
Specificity	TN/(FN+TN)	=137,847/176,331	78.2%

Table 10. Sensitivity of model estimates to changes in the error cost ratio and assumed violation duration

This table reports the sensitivity of our model estimates to variations in the error cost ratio and the assumed duration of the violation when calibrating the model. All results are based on the parsimonious logistic prediction model summarized in column 2 of Table 7. Each calibration reflects a given error cost ratio from 1/1.5 to 1.5/1 and an assumed violation duration from one to three years. An error cost ratio of 1.5/1 indicates that Type I errors (false positives) are weighted 50% more than Type II errors (false negatives). The highlighted columns report the implied probability of misrepresentation (i.e., prevalence) and probability of enforcement action for each calibration. The highlighted row reports our base case with a 1/1 error cost ratio and three-year violation duration.

Key assumptions		Model results								Model quality measures	
Error cost ratio (Type I/Type II)	Violation duration	True negatives	False positives	False negatives	True positives	N	Total model-predicted violation years	Probability of misrepre-sentation	Probability of enforcement action	Sensitivity	Specificity
1/1.5	1-period	98,080	78,251	384	1,955	178,670	80,206	44.9%	2.4%	83.6%	55.6%
	2-period	105,329	71,002	485	1,854	178,670	72,856	40.8%	2.5%	79.3%	59.7%
	3-period	111,807	64,524	585	1,754	178,670	66,278	37.1%	2.6%	75.0%	63.4%
1/1.25	1-period	110,268	66,063	505	1,834	178,670	67,897	38.0%	2.7%	78.4%	62.5%
	2-period	117,142	59,189	623	1,716	178,670	60,905	34.1%	2.8%	73.4%	66.4%
	3-period	122,950	53,381	725	1,614	178,670	54,995	30.8%	2.9%	69.0%	69.7%
1/1	1-period	127,024	49,307	718	1,621	178,670	50,928	28.5%	3.2%	69.3%	72.0%
	2-period	133,118	43,213	835	1,504	178,670	44,717	25.0%	3.4%	64.3%	75.5%
	3-period	137,847	38,484	934	1,405	178,670	39,889	22.3%	3.5%	60.1%	78.2%
1.25/1	1-period	139,122	37,209	908	1,431	178,670	38,640	21.6%	3.7%	61.2%	78.9%
	2-period	144,327	32,004	1,022	1,317	178,670	33,321	18.6%	4.0%	56.3%	81.9%
	3-period	148,022	28,309	1,105	1,234	178,670	29,543	16.5%	4.2%	52.8%	83.9%
1.5/1	1-period	145,508	30,823	1,029	1,310	178,670	32,133	18.0%	4.1%	56.0%	82.5%
	2-period	150,097	26,234	1,151	1,188	178,670	27,422	15.3%	4.3%	50.8%	85.1%
	3-period	153,312	23,019	1,230	1,109	178,670	24,128	13.5%	4.6%	47.4%	86.9%