EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 19(1)
March 2022: 30–46

Compared to What? Does Benford's Law Really Detect Corporate Fraud?

Stephen Walker[1]

LINK TO ABSTRACT

The Netflix series *Digits* devoted an episode in 2020 to Benford's law. Benford's law possesses fascinating and mysterious properties that have captured the imaginations of the viewing public, but it is not magic and its uses are limited. Benford's law measures the relative frequency of the first digit of a set of numbers and suggests a non-uniform distribution where the digit 1 "will appear as the first digit 30.1% of the time, 2 will appear 17.6% of the time, and so forth" (Amiram et al. 2015).

In their paper "Financial Statement Errors: Evidence from the Distributional Properties of Financial Statement Numbers," authors Dan Amiram, Zahn Bozanic, and Ethan Rouen (ABR) develop a unique measure based on Benford's law that applies the distribution of first digits obtained from publicly available financial statements, which include the income statement, balance sheet, and cash flow statement. They call their measure the Financial Statement Divergence Score (FSD Score), which they make available for free for non-commercial use.[2] In the abstract to their paper, the authors say that their measure "predicts material misstatements as identified by SEC Accounting and Auditing Enforcement Releases and can be used as a leading indicator to identify misstatements" (ABR 2015). The paper won the 2017 Deloitte Foundation Wildman Medal Award and was featured in the financial press including the *Wall Street Journal* (McGinty 2014). The paper had 131 Google Scholar citations as of 27 February 2022. I seek to

1. I thank anonymous referees for their helpful commentary and suggestions for improvement.
2. The data can be downloaded directly (**link**) or through a form on Professor Bozanic's website (**link**).

analyze both the degree to which their measure really predicts material misstatements and the extent to which it can serve as a leading indicator for them. While the ABR claim is essentially a yes/no question as to *whether* their measure works, I instead ask the question as to *how well* it works.[3] In other words, I ask a question inspired by one of the greatest social theorists and economists of the last half century, Thomas Sowell: Compared to what? Sowell is the Rose and Milton Friedman Senior Fellow on Public Policy at the Hoover Institution, and he frequently asks "Compared to what?" in public discourse.

Analysis of predicting material misstatements

To start this analysis, I construct a dataset based on the following. First, I obtained the SEC Accounting and Auditing Enforcement Release (AAER) database from Professor Patricia Dechow of the USC Marshall School of Business (see Dechow et al. 2011). Next, I obtained company financial statement variables from the Compustat database and constructed the variables described in the ABR paper;[4] for purposes of brevity, I do not describe the other variables here, but these definitions can be found in the original paper (ABR 2015). I then downloaded from Professor Bozanic's website the "FSD Score,"[5] which is based on Benford's law. I filled missing values at the means of each variable by industry and year in order to keep observation drops to a minimum. The combined file contains 103,289 unique firm-year observations spanning 1990–2011, with 353 unique AAERs, or 0.34 percent of the sample.

Replicating results from the original paper and considering an alternative timeframe

I start by reconstructing the results from Column 1 of Table 10 in ABR (2015) and successfully replicated their results shown here in Column 3 of Ta-

3. I thank an anonymous referee for suggesting this clarification.

4. Compustat variables that were recoded zero if missing include: *at act che lct dlc txp ivao lt lct dltt ivst pstk rect invt ppent sale ib sstk*.

5. The authors restrict their analysis to companies that had at least 100 account line items though they say that an unrestricted sample did not alter their inferences (ABR 2015, 1551). For consistency, I followed this approach.

TABLE 1. Replicating ABR (2015) and applying the same specification to the period immediately preceding their study

	Years 2001–2011			Years 1990–2000		
	(1)	(2)	(3)	(4)	(5)	(6)
FSD_Score	−33.219*** (−3.15)	−34.683*** (−3.28)	−34.021*** (−3.18)	7.835 (0.97)	4.221 (0.52)	5.138 (0.62)
ABS_JONES_RESID			−1.564* (−1.93)			0.541 (1.22)
STD_DD_RESID			0.754 (1.36)			0.292 (0.61)
MANIPULATOR			0.217 (0.86)			0.048 (0.25)
F_SCORE		0.437*** (5.47)	0.225* (1.75)		0.559*** (10.49)	0.158 (1.60)
ABS_WCACC			1.405 (1.46)			0.977 (1.55)
ABS_RSST			0.414 (1.10)			0.335 (1.17)
CH_CS			0.093 (0.80)			0.166* (1.91)
CH_ROA			0.124 (0.27)			−0.432 (−1.14)
SOFT_ASSETS			1.108*** (2.58)			1.695*** (4.20)
ISSUE			0.340 (0.86)			1.420** (2.42)
MTB			0.034* (1.88)			0.059*** (4.98)
AT			0.000 (0.27)			0.000*** (2.96)
Constant	−5.050*** (−16.69)	−5.424*** (−17.44)	−6.520*** (−11.51)	−5.643*** (−21.93)	−6.161*** (−23.41)	−8.708*** (−12.72)
Observations	55,875	55,875	55,840	47,414	47,414	44,208

Notes: Coefficients are based on a logistic regression and reported in log odds. Following ABR, AAER is an indicator variable equal to 1 for the first year of the misstatement. For other variable definitions, see ABR 2015. $AAER_{i,t} = a + \beta_1 FSD\_SCORE_{i,t} + \beta_2 ABS\_JONES\_RESID_{i,t} + \beta_3 STD\_DD\_RESID_{i,t} + \beta_4 MANIPULATOR_{i,t} + \beta_5 F\_SCORE_{i,t} + \beta_6 ABS\_WCACC_{i,t} + \beta_7 ABS\_RSST_{i,t} + \beta_8 CH\_CS_{i,t} + \beta_9 CH\_ROA_{i,t} + \beta_{10} SOFT\_ASSETS_{i,t} + \beta_{11} ISSUE_{i,t} + \beta_{12} MTB_{i,t} + \beta_{13} AT_{i,t}$. Z-statistics in parentheses. ***p<0.01, **p<0.05, *p<0.1.

ble 1. In this specification, there were a total of 55,840 observations.[6] I also report additional specifications in Columns 1 and 2 that use the measure as a standalone

6. For robustness, alternative specifications included (1) dropping the missing constructed variables based on the zero recodings for the Compustat variables and (2) dropping the missing constructed variables that did not include these recodings. Including the zero recodings returns 39,713 observations and not including them returns 34,663 observations, a quantity which is closer to ABR's reported 27,805 observations. For these alternative specifications, inferences are similar.

and as an incremental variable to the F-Score, which is the prior detection model from the accounting literature. The FSD Score's coefficients are similar across Models 1, 2 and 3, and are directionally consistent with the −40.691 reported in the ABR paper. Evaluating the marginal effect in Model 3 from the mean to the 25th percentile yields a 0.057 percent change in probability, which was similar to the 0.046 percent estimate from their original regression (I use this range because it is the range provided by ABR's descriptive statistics table). I also considered an alternative period in similar length immediately preceding their study period. Evaluating the same regressions for the years 1990–2000, the coefficient on the FSD Score turns positive and insignificant, which raises questions as to the validity of their FSD Score measure to AAER events outside their study period.

Compared to what? How the FSD Score measures up to the F-Score and a naive sales measure for predicting material misstatements

While the logistic regression shows that there was significance for the FSD Score for ABR's preferred timeframe, interpreting the predictive performance is difficult. For this section, I compare predictive performance in terms of standard metrics from the prediction literature including positive predictive value (precision), AUROC, and NDCG, which were not evaluated in the original paper; for an explanation of these measures, please see the appendix to this paper.

I compare three measures to the FSD Score. The first measure is the *F-Score*, which is constructed based on the coefficients from a material misstatement detection model created by Dechow et al. (2011). The second measure takes Column 2 from Table 1, which simplifies the logistic regression from the ABR (2015) paper taking the only two variables that were reported to be significant at the 0.01 level, the F-Score and the FSD Score; this measure I call *Combined*, and it measures the additive predictive power of the FSD Score relative to the F-Score. I checked the robustness of this simplification to the full ABR regression (from Column 3 in Table 1), and inferences are similar. The third and last measure calculates a four-year geometric sales growth rate, which represents a naive measure. The sales growth rate, a relatively simple measure, was suggested to me in conversation by Professor Howard Schilit, an expert in forensic analysis and author of *Financial Shenanigans*, a popular book among industry professionals.

In their paper, ABR (2015, 1557) acknowledge that their results show that divergence from Benford's law is associated with a *decrease* in likelihood of material misstatement, which was opposite of what they expected, represented by the

negative coefficient from Table 1 for the main sample period (but not by the alternative timeframe). They address the reasons why they believe they found this result. While their argument is up for debate, I do not address it here because my analysis is a predictive exercise, where the 'why' is less important than the 'what.' The 'what' here is the predictive output. Therefore, when comparing Tables 2 and 3, examine Decile 1 for the FSD Score while comparing it to Decile 10 of the others. For Table 2, the dependent variable is an indicator variable coded "1" for the first year of the misstatement (but not for the duration for multi-year misstatements), following ABR's approach.

TABLE 2. Misstatement year =1 only in the first year

	Years 2001–2011				Years 1990–2000			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Decile	FSD Score	F-Score	Combined	Sales	FSD Score	F-Score	Combined	Sales
1	**0.4%**	0.1%	0.1%	0.1%	**0.4%**	0.1%	0.1%	0.2%
2	0.3%	0.1%	0.1%	0.2%	0.5%	0.1%	0.2%	0.3%
3	0.2%	0.1%	0.2%	0.1%	0.3%	0.2%	0.2%	0.2%
4	0.4%	0.1%	0.2%	0.3%	0.3%	0.3%	0.3%	0.4%
5	0.3%	0.2%	0.2%	0.2%	0.5%	0.4%	0.4%	0.3%
6	0.2%	0.3%	0.2%	0.4%	0.4%	0.2%	0.3%	0.4%
7	0.1%	0.3%	0.2%	0.3%	0.6%	0.6%	0.4%	0.7%
8	0.2%	0.3%	0.4%	0.2%	0.5%	0.8%	0.8%	0.5%
9	0.2%	0.4%	0.4%	0.3%	0.4%	0.6%	0.5%	0.6%
10	0.1%	**0.5%**	**0.6%**	**0.4%**	0.5%	**1.3%**	**1.3%**	**1.1%**
Other Metrics:								
AUROC	0.57	0.64	0.65	0.60	0.52	0.71	0.69	0.66
Precision at 1%	0.5%	0.9%	1.1%	0.7%	0.6%	2.4%	1.9%	0.9%
NDCG	0.381	0.390	0.397	0.387	0.418	0.460	0.460	0.448
NDCG@10%	0.073	0.104	0.110	0.088	0.053	0.162	0.167	0.131
NDCG@1%	0.015	0.027	0.032	0.022	0.009	0.038	0.036	0.012

Notes: The proportion of material misstatements for the first year in the sample from 2001–2011 was 0.2%. From 1990–2000 it was 0.4%. Deciles of probabilities for each of the four models were created by year. The highest risk decile for the FSD Score is Decile 1. For the other measures, it is Decile 10.

The percentage values reported by decile represent the likelihood of finding a material misstatement and can be compared to the unconditional likelihood, which is the prevalence of material misstatements in the sample. These values in the sample were 0.2 percent for 2001–2011 and 0.4 percent from 1990–2000, which illustrate just how rarely these events occur. Overall, improvements do not move the needle very much. On a relative basis, the FSD Score, as a standalone, performs the worst among the models. Oddly, the standalone scenario was not evaluated in

the ABR paper, as only a fully loaded model with 13 variables was estimated. The next measure, the F-Score, beats the FSD Score in both periods. The FSD Score marginally improves the F-Score in the Combined variable for 2001–2010, but not in the previous period. Interestingly, the naive sales growth screen performs similarly to the complex methodologies at the highest risk decile. While the other measures seem to work in the previous period, the standalone FSD Score appears to fail entirely. In addition to examining these deciles, I produce Precision at 1%, which reports the fraction of true positives in the top 1 percent of each measure by year, similar to what was done in the decile analysis. In addition, I provide other pooled metrics including AUROC, NDCG, and NDCG@k, which are explained in more detail in the Appendix. Regardless of measure, inferences are similar.

TABLE 3. All misstatement years = 1

	Years 2001–2011				Years 1990–2000			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Decile	FSD Score	F-Score	Combined	Sales	FSD Score	F-Score	Combined	Sales
1	**1.1%**	0.4%	0.3%	0.5%	**0.7%**	0.2%	0.2%	0.2%
2	1.2%	0.7%	0.4%	0.6%	1.0%	0.2%	0.4%	0.6%
3	0.9%	0.6%	0.7%	0.8%	0.8%	0.5%	0.5%	0.5%
4	1.0%	0.5%	0.9%	0.7%	0.6%	0.8%	0.7%	0.5%
5	1.0%	0.7%	0.7%	0.7%	0.9%	0.6%	0.7%	0.6%
6	0.9%	0.8%	0.9%	0.8%	0.7%	0.6%	0.6%	0.7%
7	0.8%	1.3%	1.0%	0.9%	0.9%	0.8%	0.7%	1.1%
8	0.6%	1.2%	1.4%	1.0%	1.0%	1.3%	1.3%	0.8%
9	0.8%	1.5%	1.4%	1.3%	0.7%	1.1%	1.0%	1.1%
10	0.6%	**1.0%**	**1.2%**	**1.6%**	0.7%	**2.0%**	**2.0%**	**2.1%**
Other Metrics:								
AUROC	0.55	0.59	0.59	0.60	0.50	0.67	0.66	0.67
Precision at 1%	1.1%	1.1%	1.4%	1.3%	0.8%	2.8%	2.6%	1.7%
NDCG	0.508	0.509	0.513	0.520	0.479	0.520	0.521	0.518
NDCG@10%	0.080	0.074	0.078	0.129	0.053	0.1620	0.171	0.174
NDCG@1%	0.011	0.014	0.016	0.010	0.008	0.037	0.037	0.019

Notes: The proportion of material misstatements for all years in the sample from 2001–2011 was 0.9%. From 1990–2000 it was 0.8%. Deciles of probabilities for each of the four models were created by year. The highest risk decile for the FSD Score is Decile 1. For the other measures, it is Decile 10.

There is nothing wrong with ABR's approach to only consider the first year of the material misstatement, but the frequency of the dependent variable is significantly reduced as it only captures the first year, thus making it harder to detect. It is also unclear why investors would not also be interested in detecting an ongoing multi-year misstatement. Therefore, I also analyze material misstatements

including the additional years for the multi-year cases, shown in Table 3. This time, the dependent variable is coded '1' for every year there is a material misstatement. The story remains the same, and sales growth shows a slight edge over the other models in terms of top-decile performance. As to the other measures, AUROCs are similar across F-Score, Combined, and Sales models. For both dependent variables, Tables 2 and 3 show that the FSD Score as a standalone performs the worst, and it is not clear if it really adds much value, if at all, as an additional variable to the existing F-Score model.

As a leading indicator?

So far, the results I have provided reflect the contemporaneous relationship between financials and material misstatements; the relationship attempts to predict fraud given a known set of public financial statements. ABR (2015) say that their measure could serve as a leading indicator in the detection of material misstatements. As evidence of such, they show significant coefficients on one-year and two-year lagged FSD Scores; such significance implies that a material misstatement could be detected prior to the year of occurrence based on these leading values. Curiously, they did not lag the control variables, so these published models could never serve as leading indicators as currently specified. Despite this misspecification, to analyze whether the FSD Score could be used as a leading indicator, I repeat the previous analysis for each measure according to their lagged values. Table 4 shows evidence that the FSD Score does not work as a standalone and it does not improve the F-Score in the combined metric. The sales growth screen still performs well in comparison for the top decile. Since positive coefficients were reported on the lagged variables by ABR, the comparable decile for the FSD Score is now Decile 10. As for the other measures, sales outperforms in terms of NDCG, Precision at 1%, AUROC, NDCG@10%, and NDCG@1% for their main sample period of 2001–2011.

TABLE 4. One-year leading indicator (misstatement year = 1 only for the first year)

	Years 2001–2011				Years 1991–2000			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Decile	FSD Score	F-Score	Combined	Sales	FSD Score	F-Score	Combined	Sales
1	0.2%	0.2%	0.1%	0.2%	0.5%	0.1%	0.1%	0.1%
2	0.2%	0.1%	0.2%	0.0%	0.2%	0.1%	0.2%	0.2%
3	0.1%	0.1%	0.1%	0.1%	0.5%	0.2%	0.1%	0.2%
4	0.3%	0.2%	0.2%	0.3%	0.5%	0.2%	0.3%	0.5%
5	0.2%	0.2%	0.3%	0.3%	0.5%	0.4%	0.3%	0.3%
6	0.3%	0.2%	0.3%	0.2%	0.4%	0.4%	0.5%	0.4%
7	0.1%	0.4%	0.2%	0.2%	0.4%	0.5%	0.4%	0.6%
8	0.3%	0.2%	0.3%	0.3%	0.4%	0.7%	0.5%	0.5%
9	0.2%	0.4%	0.4%	0.2%	0.3%	0.6%	0.8%	0.5%
10	**0.2%**	**0.3%**	**0.3%**	**0.4%**	**0.4%**	**1.1%**	**1.0%**	**0.9%**
Other Metrics:								
AUROC	0.53	0.61	0.59	0.61	0.50	0.70	0.69	0.65
Precision at 1%	0.5%	0.5%	0.4%	0.9%	0.3%	2.6%	2.0%	1.0%
NDCG	0.355	0.368	0.364	0.371	0.396	0.443	0.441	0.423
NDCG@10%	0.051	0.071	0.072	0.085	0.059	0.144	0.145	0.122
NDCG@1%	0.006	0.013	0.010	0.018	0.010	0.044	0.043	0.015

Notes: The proportion of material misstatements for all years in the sample from 2001–2011 was 0.2%. From 1991–2000 it was 0.4%. Deciles of probabilities for each of the four models were created by year. Decile 10 shows the highest risk measure for all four measures. Since one-year lagged values are required, the second period starts in 1991 instead of 1990.

Table 5 reports the two-year lagged values for the key measures and inferences are similar to the one-year lagged analysis. For the main sample period of 2001–2011, the F-Score outperforms in terms of top decile, top 1 percent, AUROC, NDCG, and NDCG@10%.

TABLE 5. Two-year leading indicator (misstatement year = 1 only for the first year)

	Years 2001–2011				Years 1992–2000			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Decile	FSD Score	F-Score	Combined	Sales	FSD Score	F-Score	Combined	Sales
1	0.1%	0.1%	0.1%	0.2%	0.3%	0.1%	0.0%	0.3%
2	0.2%	0.1%	0.2%	0.1%	0.7%	0.0%	0.2%	0.3%
3	0.1%	0.2%	0.2%	0.2%	0.4%	0.2%	0.3%	0.2%
4	0.3%	0.2%	0.1%	0.2%	0.3%	0.3%	0.2%	0.3%
5	0.3%	0.1%	0.1%	0.3%	0.4%	0.4%	0.3%	0.6%
6	0.3%	0.3%	0.2%	0.2%	0.4%	0.6%	0.6%	0.3%
7	0.2%	0.2%	0.4%	0.3%	0.3%	0.5%	0.6%	0.6%
8	0.2%	0.3%	0.2%	0.1%	0.4%	0.5%	0.4%	0.5%
9	0.2%	0.3%	0.3%	0.2%	0.4%	0.7%	0.8%	0.5%
10	**0.4%**	**0.5%**	**0.4%**	**0.4%**	**0.4%**	**0.8%**	**0.7%**	**0.6%**
Other Metrics:								
AUROC	0.56	0.64	0.61	0.58	0.51	0.67	0.66	0.61
Precision at 1%	0.2%	0.3%	0.2%	0.0%	0.6%	0.9%	1.4%	0.0%
NDCG	0.350	0.363	0.361	0.362	0.384	0.411	0.408	0.396
NDCG@10%	0.065	0.103	0.092	0.097	0.062	0.115	0.099	0.092
NDCG@1%	0.000	0.005	0.006	0.013	0.007	0.024	0.012	0.000

Notes: The proportion of material misstatements for all years in the sample from 2001–2011 was 0.2%. From 1992–2000 it was 0.4%. Deciles of probabilities for each of the four models were created by year. Since two-year lagged values are required, the second period starts in 1992 instead of 1990.

Evaluating the predictive margins from the original paper

While ABR (2015) did not report predictive margins—that is, to what extent the probability changes from extreme changes in the underlying metric—their readers could apply a back-of-the-envelope method using the output from their reported logistic regression and their summary statistics table and come to the conclusion that their model does not really move the needle in absolute terms. Their logistic regression from their first model is reproduced here in Table 6, Column 1. Their output was reported in log odds, and, to the layperson, it might seem that all others are beaten by the FSD Score, which has the greatest absolute coefficient value. The real effect on probability is tiny, which can be evaluated by calculating the predictive margins for the coefficient of interest. The table walks through the calculation evaluating the change in probability for a move in the

FSD Score from its mean value to the 25th percentile based on ABR's descriptive statistics table. A change from the mean to the 25th percentile is the best the reader could do since ABR's descriptive statistics table does not report values beyond this level. The result is that the change in FSD Score increases the probability of material misstatement detection by 0.046 percent according to their sample—indeed an extremely small change.

TABLE 6. Evaluating the predictive margins from the original logistic regression

	Model 1 coefficients from ABR	Means of variables	FSD Score at 25th percentile; means otherwise	Log odds at means	Log odds varying FSD Score only
	(1)	(2)	(3)	(1) × (2)	(1) × (3)
FSD Score	**−40.691**	**0.030**	**0.023**	**−1.204**	**−0.952**
ABS_JONES_RESID	−1.078	0.184	0.184	−0.198	−0.198
STD_DD_RESID	0.011	0.123	0.123	0.001	0.001
MANIPULATOR	0.122	0.143	0.143	0.017	0.017
F_SCORE	1.980	0.401	0.401	0.793	0.793
ABS_WCACC	−1.233	0.054	0.054	−0.067	−0.067
ABS_RSST	0.401	0.138	0.138	0.055	0.055
CH_CS	0.004	0.146	0.146	0.001	0.001
CH_ROA	1.339	−0.002	−0.002	−0.003	−0.003
SOFT_ASSETS	−0.121	0.545	0.545	−0.066	−0.066
ISSUE	−0.341	0.915	0.915	−0.312	−0.312
MTB	0.166	1.360	1.360	0.226	0.226
AT	0.000	3228.380	3228.380	0.000	0.000
Constant	−5.686			−5.686	−5.686
Sum of Log Odds				−6.442	−6.189
Odds Ratio (OR): e^(log odds)				0.002	0.002
Probability (OR / 1+OR)				0.159%	0.205%
Change in Probability				**0.046%**	

Notes: Column 1 reports the original logistic regression coefficient values from ABR (2015). Columns 2 and 3 are values sourced from ABR's descriptive statistics. Coefficients are reported in log odds and probabilities are evaluated accordingly. For the logistic regression, AAER was an indicator variable equal to 1 for the first year of the misstatement. For other variable definitions, see ABR 2015. $AAER_{i,t} = a + \beta_1 FSD\_SCORE_{i,t} + \beta_2 ABS\_JONES\_RESID_{i,t} + \beta_3 STD\_DD\_RESID_{i,t} + \beta_4 MANIPULATOR_{i,t} + \beta_5 F\_SCORE_{i,t} + \beta_6 ABS\_WCACC_{i,t} + \beta_7 ABS\_RSST_{i,t} + \beta_8 CH\_CS_{i,t} + \beta_9 CH\_ROA_{i,t} + \beta_{10} SOFT\_ASSETS_{i,t} + \beta_{11} ISSUE_{i,t} + \beta_{12} MTB_{i,t} + \beta_{13} AT_{i,t}$.

Conclusion

Benford's law has captured the imaginations of fraud researchers worldwide. The results here should curb some of the enthusiasm, particularly for the detection of material misstatements based on publicly available financial statements. Our best models in accounting research, applying financial statement variables do very little to move the needle in terms of the predictive margins based on what is observable in the financial statements. Anecdotally, I've heard that industry professionals have found little use with the models published in accounting research and the evidence here explains why this is the case. ABR (2015, 1541), as motivation for their research, noted that previous methods for analyzing firm-level financial statements "have deficiencies that limit their usefulness." Given the evidence here, their FSD Score joins that club for the purpose of detection of material misstatements. The present critique started with: Compared to what? The FSD Score based on Benford's law does not meaningfully detect material accounting misstatements nor does it perform as a leading indicator as a standalone measure. As an additive measure to the F-Score, the evidence here shows a range of outcomes from very little benefit to potential worsening of predictive power. The naive sales growth screen performs surprisingly well compared to advanced statistical methods. Finally, the FSD Score does not appear robust to an alternative period of time immediately preceding the sample period chosen by ABR.

Appendix

The purpose of this appendix is to explain the available choices of metrics used in classification modeling and to provide the reasoning for the selection of metrics presented in the present paper. To start this discussion, Table A1 shows an example of a classification matrix typically used when analyzing model performance. For example, the logistic regression can be used to estimate a fitted probability for each observation. However, the researcher must choose at which point along this probability continuum to classify cases in the positive or negative. This cutoff choice can either be made at a specific value or standardized through examining top deciles or percentiles. The cutoff choice can be optimized by weighting the relative cost of false positives and false negatives. For an e-commerce company, declining a potentially fraudulent transaction must be weighed against the possibility of lost margin from a sale. For the purposes of corporate fraud detection, there are real-world constraints. Investigative hours are limited and consumed by existing investigations. Detection models, if they work, would act as

a screen or as a watch list to launch new investigations at the margin. Therefore, when analyzing performance, it is accuracy at the top that matters the most, which is why I presented the decile performance for each model considered. This measure is also known as Precision@k, which is a common metric used to analyze true positives out of the predicted positives at some threshold *k*. In this case, the top decile would be precision at 10 percent. I also added precision at 1 percent to summarize results at the more extreme level of the top percentile. The downside with going more extreme is that most of the corporate fraud cases will be missed entirely as so few cases have fitted probability values at that level. However, for corporate fraud detection, reliable investigative tips typically come from other channels. Therefore, these models may be useful if precision is high enough at the margin for the highest risk cases. In addition to precision, other measures can be calculated from the classification matrix. Table A1 shows a classification matrix which includes the four quadrants of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in terms of the dependent variable in this paper (AAERs).

TABLE A1. Example of a classification matrix

	Prediction based on threshold (e.g., top 1 percent)		
	YES: AAER	**NO: AAER**	**Total**
AAER	True Positive (TP)	False Negative (FN)	Total AAERs
Non-AAER	False Positive (FP)	True Negative (TN)	Total Non-AAERs
Total	Total Predicted Positive	Total Predicted Negative	Total Sample (N)

From this classification matrix, the following values can be calculated:

- Prevalence (Total AAERs / N): These are the AAERs in the sample that is known before the classification exercise.
- Classification accuracy (TP+TN)/N: These are the correct classifications of true positives (both AAER and predicted positive) and true negatives (both non-AAER and predicted negative) out of the total observations in the sample.
- Sensitivity (TP/Total AAERs): Also known as the true positive rate, these are the true positives out of the total AAERs in the sample.
- Specificity (TN/Total Non-AAERs): Also known as the true negative rate, these are the true negatives out of the total non-AAERs in the sample.
- Precision (TP/Total Predicted Positive): These are the true positives out of those that were predicted positive.

Classification accuracy, while highly intuitive, is a poor measure to analyze model performance for fraud detection because fraud is a rare event. For rare events, classification accuracy is essentially defined by the true negative rate since non-fraud events dominate. To see why, we can write classification accuracy as the sum of the proportions that were true positive and true negative:

$$ClassificationAccuracy = \frac{TP}{N} + \frac{TN}{N}$$

These terms can be re-written in terms of prevalence, sensitivity, and specificity:

$$ClassificationAccuracy = [Prevalence*Sensitivity] + [(1 - Prevalence)*Specificity]$$

When prevalence is extremely low, for example 0.5 percent, the other term would be weighted accordingly, in this example 99.5 percent. Furthermore, the measure is hard to interpret for rare events because it can be arbitrarily maximized through an uninformative decision rule. For rare events, the strategy would be to classify all cases in the negative. In this case, sensitivity would be zero as no AAERs would be captured, but specificity would be 100 percent as all negatives would be classified correctly. Therefore, for an event occurring 0.5 percent of the time, classification accuracy would equal 99.5 percent—which sounds high, but contains no information beyond what was known before the classification exercise, which was the prevalence rate.

Precision, on the other hand, essentially follows from Bayes' rule. Bayes' rule is defined as:

$$P(A \mid B) = \frac{P(A)xP(B \mid A)}{P(B)}$$

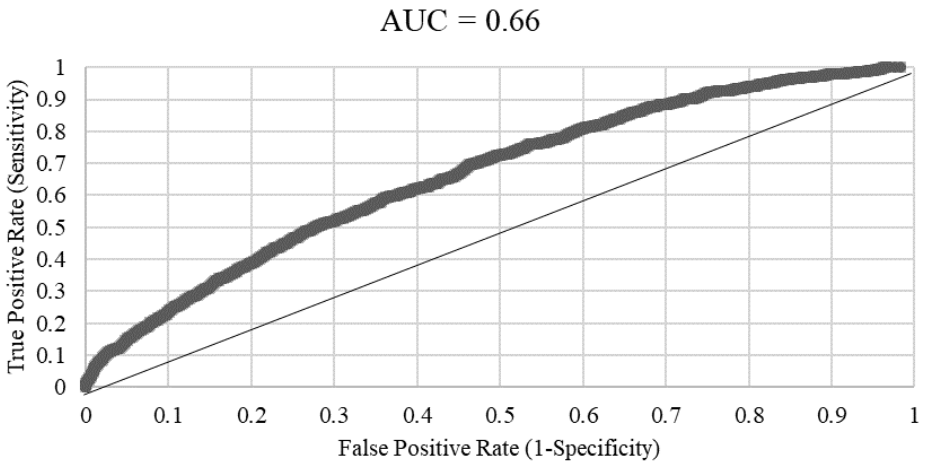
Bayes' rule states that the prior probability P(A) can be updated to the posterior probability P(A|B) through learning about P(B) and P(B|A). In the context of the AAER sample, the prior P(A) is the prevalence of AAERs in the sample. By analyzing financial statement values through complex modeling, we can learn both the probability of a positive classification P(B) as well as the probability of a positive classification given that an AAER occurred P(B|A). This equation can be rewritten in terms of prevalence, sensitivity, and specificity:

$$Precision = \frac{Prevalence*Sensitivity}{Prevalence*Sensitivity + (1 - Prevalence)*(1 - Specificity)}$$

The second alternative metric presented was AUROC, or area under the receiver operating characteristic curve, which has been described as the de facto

standard for measuring classification performance (Fawcett 2006). Some prefer this rule because it measures the area under a curve generated by mapping sensitivity and (1 − specificity) across all possible cutoff points. The AUROC is a summary statistic that is standardized between 0.5 (where the model fails completely) and 1.0 (which would represent a perfect classifier). Figure A1 shows an example of an AUROC.

Figure A1. Area under the curve example



While AUROCs provide a standard statistic, they are more difficult to interpret. A Google machine-learning crash course says that "one way of interpreting AUROC is as the probability that the model ranks a random positive example more highly than a random negative example" (**link**).

The AUROC measures the entire area below the curve. If the curve went up in a straight line to where sensitivity is one and the false positive rate is zero (upper left of the graph), the AUROC value would shade the entire graph and equal 1.0 which would imply a perfect classifier (not likely in the real world). The minimum AUROC possible is 0.5, which is represented by the diagonal. Where sensitivity equals the false positive rate, the classifier contributes no information and is not different from a random guess in the sample. The reason why the AUROC cannot be below the diagonal is that the decision rule where the classifier does worse could simply be reversed.

The third measure reported comes from the information retrieval literature and applies a log discounting to the rank order. This measure is called normalized discounted cumulative gain (NDCG) and can be measured across the entire sample, or at a certain cutoff k (NDCG@k). It takes the following form:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

$$NDCG@k = \frac{DCG@k}{IdealDCG@k}$$

$$rel_i = RelevanceScoreinrank_i$$

While relevance offers a way to weight observations, for fraud detection this value simplifies to the binary indicator for fraud in the dependent variable. Therefore, the numerator $2^{rel_i} - 1$ simplifies to zero or one matching the AAER indicator value. Table A2 illustrates a toy example for NDCG@k where there are 10 observations and 3 positive cases. Column 1 ranks the probability outcome in rank order from 1 to 10. Column 2 identifies the AAERs, which are shown to be in the second, fourth and fifth rank. DCG@k sums the discounted cumulative gain values applying the discount factor of $1 \,/\, \log_2(i+1)$ as shown in the formula above. To normalize this value, an ideal ranking must be computed. To do so, the cumulative gain column is sorted in descending order. The final column discounts this ranking applying the same factor as before. With the DCG@k and the Ideal DCG@k values, the normalized value NDCG@k can be calculated. In this example, the value is 0.68. A perfect classifier would have the value of 1.0. For more theoretical background on NDCG@k, see Yining Wang et al. (2013).

TABLE A2. Normalized discounted cumulative gain example (NDCG@k)

$Rank_i$	Cumulative gain	Discount factor[7]	Discount cumulative gain (DCG@k)	Ideal ranking	Ideal DCG@k
(1)	(2)	(3)	(2) × (3)	(4)	(3) × (4)
1	0	1.00	0.00	**1**	1.00
2	**1**	0.63	0.63	**1**	0.63
3	0	0.50	0.00	**1**	0.50
4	**1**	0.43	0.43	0	0.00
5	**1**	0.39	0.39	0	0.00
6	0	0.36	0.00	0	0.00
7	0	0.33	0.00	0	0.00
8	0	0.32	0.00	0	0.00
9	0	0.30	0.00	0	0.00
10	0	0.29	0.00	0	0.00
Total	**3**		**1.45**	**3**	**2.13**
NDCG@k	**0.68**				

7. Discount factor is $1/\log_2(rank_i+1)$.

NDCG@k is similar to precision without the log discounting. Observe that Column 2, labeled Cumulative Gain, is the sum of the number of true positives above the cutoff threshold *k*. This value divided by the total observations in the subgroup is 30 percent for this example. Precision@k would also be 30 percent. So, NDCG@k simply adds a discount factor to the rank position for observations in this subgroup. The benefit to NDCG@k is that it adds information as to the order within the group, but at the cost of a loss to interpretability.

Data and code

Data and code used in this research is available from the journal website (**link**).

References

Amiram, Dan, Zahn Bozanic, and Ethan Rouen (ABR). 2015. Financial Statement Errors: Evidence from the Distributional Properties of Financial Statement Numbers. *Review of Accounting Studies* 20(4): 1540–1593.

Dechow, Patricia M., Weili Ge, Chad R. Larson, and Richard G. Sloan. 2011. Predicting Material Accounting Misstatements. *Contemporary Accounting Research* 28(1): 17–82.

Fawcett, Tom. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters* 27(8): 861–874.

McGinty, Jo Craven. 2014. Accountants Increasingly Use Data Analysis to Catch Fraud. *Wall Street Journal*, December 5. **Link**

Wang, Yining, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A Theoretical Analysis of NDCG Type Ranking Measures. Presented at the 26th Annual Conference on Learning Theory (Princeton, N.J.), June. *Proceedings of Machine Learning Research* 30: 25–54. **Link**

About the Author

Stephen Walker earned his Ph.D. at the University of California Haas School of Business in May 2021. Prior to his Ph.D. studies, Stephen worked in equity research at Sanford C. Bernstein in New York City. He also holds an MBA from Columbia Business School. He currently works as an economic consultant for shareholder litigation and can be reached via his personal website at stephenwalker.me.

Go to archive of Comments section
Go to March 2022 issue