EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 18(2)
September 2021: 230–234

Rejoinder to the Critique of an Article on Machine Learning in the Detection of Accounting Fraud

Stephen Walker

LINK TO ABSTRACT

In "A Response to 'Critique of an Article on Machine Learning in the Detection of Accounting Fraud'" (Bao et al. 2021), the author group chose not to respond to the fundamental issue raised in the critique (Walker 2021), which was: What was the justification for relabeling identifier values used to identify fraud in their sample? The point is critical. A justified reason would mean that the model beats the prior literature by a handsome margin. Without a justified reason, the model fails to improve upon existing logit-based methodologies. The author group did not provide a reason, much less a justified reason.

I asked the question directly to the authors prior to the publication of my critique. As I reported in my critique, the authors responded in that initial email inquiry by saying that the relabelings were necessary because a consecutive sample was needed for each fraud case (see Walker 2021, 66). Not only was this requirement undisclosed in their original paper published in the *Journal of Accounting Research*, but it also made no sense, which provided the motivation for my critique.

This issue concerns something called *serial fraud*, and it was addressed in Section 3.3 in the authors' original publication. Apparently, this issue came up in the review process because the authors wrote, "We thank the anonymous referee for raising this point," in reference to this section (Bao et al. 2020, 212 n.10). What that section says is that, since frauds can span multiple years, it is possible that the same fraud case could exist both in the training set and the test set used in machine learning. In these instances, the model could learn from its own case and, therefore, overstate results. To address this concern, the authors specifically stated that they

"recode[d] all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods" (ibid., 212).

To illustrate once again, take, as an example, a fraud case for a publicly traded company that covered fiscal years 2000 through 2002. When applying a model to make a prediction as to which companies would be fraudulent based on public financials from 2002, the authors proposed that that the fraud indicator variable in the year 2000 should be recoded to zero because of this serial-fraud issue. Except this was not true for *all* cases. An additional requirement was added within the code that was not explained in the paper: Do this procedure only if *all* observations for 2000–2002 are contained in the sample. Note that training ends in 2000 for the prediction year 2002 because of a two-year gap requirement, which is a separate issue from this discussion. So, the observation for year 2001 is irrelevant here. In cases where all observations are available, the fraud indicator variable would be recoded to zero for year 2000, which is consistent with the description written in their original paper. Consider the alternative scenario which introduces the controversy. If the observation for the irrelevant year 2001 was dropped because of a missing-values problem (i.e., one of the covariates had a missing value), then the authors would relabel the fraud identifier for any fraud year that occurred thereafter. So for this example the fraud identifier for 2002 would be labeled differently from the fraud identifier for the year 2000. Since the two identifiers no longer match (for the same fraud case), the fraud indicator variable *would not* be recoded to zero and would therefore be included in training, which contradicts Section 3.3 of their original paper. Without this relabeling, the model performed no better than a logit-based regression that existed previously in the literature.

Rather than give a rational explanation for relabeling identifiers, the authors chose to write extensively on other topics, and concluded with one that was irrelevant to my critique. Their reply led by addressing the missing-values problem. In their earlier email reply responding to my initial questions, the authors stated that observations with missing values were dropped. However, some observations with missing values were, in fact, not dropped, and those missing values were, instead, filled with zeros. When I asked why this was the case in the critique, the authors' response explained that the treatment was consistent with prior literature and that they applied this logic consistently. So, CompuStat variables *txp*, *ivao*, *ivst*, and *pstk* were recoded to zero because it followed "common practice in the prior accounting literature" (Bao et al. 2021, 72). Observations that had missing values for other variables, such as debt issuance, were dropped from the sample. While one would think that these cases could also be reasonably recoded to zero, the authors chose not to do so and provided no further explanation. Despite their treating the variables differently, I encourage the reader to move beyond this

concern because it is does not address the fundamental issue, which is: *Why would fraud identifiers need relabeling at all?*

The next section of their reply was entitled "Walker's approach to dealing with serial fraud" (Bao et al. 2021, 72). I do not know why the authors chose to attribute my name to their own written methodology as described in Section 3.3 of their original paper. They started with the argument that I did not recalibrate the number of trees in the RUSBoost parameters. To provide some background on this issue, machine learning algorithms contain several parameters which could be adjusted so that a better fit can be obtained for the training sample. However, it is unknown *ex ante* how this would affect the out-of-sample test. Specifically, the authors wrote that I did not properly tune the number-of-trees parameter to optimize performance (ibid.), and, when they did so, their performance improved on what I reported in my critique. Upon inspection of their results in the reply shown as their Table 1, taking their "re-optimized" parameters (ibid., 73) at face value, results were virtually identical to the results I showed in my critique making it obvious that tuning does not matter. For example, for the main sample period of 2003–2008, I reported 9 hits in the critique (Walker 2021, 68 Table 2 Panel A) while they reported 10 hits in the reply (Bao et al. 2021, 74 Table 1 Panel B). So, there was one additional hit from this tuning. Furthermore, this result is far from the purported improvement where they reported 16 hits in the original publication (Bao et al. 2020, 204, 223). Second, the authors do not state how their parameter tuning was implemented, nor do they provide the code for this process either with their published paper or with this reply. This is typically a requirement since parameter tuning must be done only on the training sample. If it was done to maximize out-of-sample performance, then the procedure would be invalid. Generally, tuning does not alter machine learning performance significantly. In fact, recent literature in the computer sciences reported that leaving models at their default parameter values was non-inferior to optimization (Weerts, Mueller, and Vanschoren 2020).

The second issue raised in this section was that I only published Table 3 from their original paper for the years 2003–2008, while ignoring the results from the following Table 5, which included three alternative test samples (Bao et al. 2021, 73). The implication was that I cherry-picked results. The reason I chose Table 3 was that it was their *main result*. In their original paper, they stated "we use the years 2003–2008 as our primary test sample" (Bao et al. 2020, 211). Table 5 was included only for robustness in a section entitled "Supplemental Analyses" (ibid., 224ff.). Regardless, equivalent comparisons for the alternate period 2003–2005 could be easily calculated since I provided cross-sectional results by year in the analysis, whereas they only reported the overall average. Incremental results beyond 2008 by their analysis were essentially the same between the logit-based model and the

RUSBoost model. For example, expanding the test period through 2014 from 2008, their RUSBoost model picked up an additional three cases while the logit-based model also picked up three cases. Lastly, the authors concluded this section by writing that the RUSBoost "always outperforms" (Bao et al. 2021, 75). This statement is in contradiction with their own reported Table 1, where results for their primary test sample show the value for NDCG@k was 0.0273 with the logit-based model whereas their RUSBoost model underperformed at 0.0237, so the logit-based model performed 15 percent better (ibid., 74). We also know that NDCG@k was their preferred metric because, in the original paper, the authors wrote that, relative to the AUC (Area-under-the-curve), the "NDCG@k is more useful to regulators and other monitors" (Bao et al. 2020, 205). The authors also concluded in the reply that results did not alter inferences (Bao et al. 2021, 71, 75). How could this be true? Their new table showed results far from the purported 70 percent improvement shown in their original publication.

Finally, the last section in their response was titled "What is the optimal approach for dealing with serial fraud?" (Bao et al. 2021, 75). They concluded this section by saying, "Walker's approach of relying solely on p_AAER ID to define serial fraud could be inappropriate" (ibid., 76). Again, why is the approach taken in their original paper declared to be my approach? What other way is there to identify the fraud except with the fraud identifier? This identifier is the same number given by the Securities and Exchange Commission which issued the AAER—the Accounting and Auditing Enforcement Release. In their Figure 1, they illustrate an example titled "a serial fraud example with a key fraud revelation event during the training period" (ibid.). While timing of fraud revelation might make an interesting discussion, it was never addressed in my critique. What they label as "Walker's approach" is precisely their approach, applied to all observations without a missing intervening variable.

In summary, the authors provided no justification for relabeling identifiers with their reply. In my critique, I made a clear case that their code was not consistent with how their paper described the implementation of the solution to the serial fraud problem (in Bao et al. 2020, Section 3.3, which appears to have been drafted originally in response to an anonymous referee during the review process). Without a reasonable explanation, results do not hold up to scrutiny and I can only conclude that their RUSBoost model does not outperform the logit-based model for the detection of corporate fraud. While the authors wrote otherwise, their updated data supported this conclusion, which showed that the logit-based model from prior literature outperformed their RUSBoost model for the main sample period 2003–2008 in terms of their preferred metric NDCG@k.

References

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2020. Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research* 58(1): 199–235.

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2021. A Response to "Critique of an Article on Machine Learning in the Detection of Accounting Fraud." *Econ Journal Watch* 18(1): 71–78. **Link**

Walker, Stephen. 2021. Critique of an Article on Machine Learning in the Detection of Accounting Fraud. *Econ Journal Watch* 18(1): 61–70. **Link**

Weerts, Hilde J. P., Andreas C. Mueller, and Joaquin Vanschoren. 2020. Importance of Tuning Hyperparameters of Machine Learning Algorithms. Working paper, July 15. **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., Stephen worked in sellside equity research for Sanford C. Bernstein in New York City. He also holds an MBA from Columbia Business School. He can be reached via his personal website at stephenwalker.me (**link**).

Go to archive of Comments section
Go to September 2021 issue