

Explanation versus Prediction in Macroeconomic Modelling: Reconciling Econometric and Machine-Learning Evidence on Nigerian FDI

Ofiero hor Ufuoma Earnest

ORCID: <https://orcid.org/0009-0002-7468-8082>

Department of Finance and Banking, Faculty of Management Sciences, University of Port Harcourt, Nigeria

Abstract: A variable can be statistically significant in a structural model and carry no weight in a predictive one, or the reverse. This paper works through that problem using two companion analyses of the same Nigerian quarterly dataset, 2008Q2 to 2024Q4: an ARDL bounds-testing model of FDI and a leakage-free machine-learning forecasting pipeline. The interest rate makes the point cleanly. In the ARDL model it is significant and negative under both classical and HAC standard errors (coefficient -57.24, HAC $p = 0.005$). In the forecasting model its permutation importance has a bootstrap 95% confidence interval of [-29.34, 42.94], which spans zero, so its predictive contribution is not established at all. Exchange-rate volatility fails on both axes: insignificant in the ARDL model ($p = 0.70$) and with a bootstrap interval of [-12.35, 44.06]. Neither pattern is a contradiction. The two frameworks answer different questions, one about a stable structural association over the full sample and one about whether a variable improves forecasts on held-out data, and neither licenses a causal claim. We set out a three-way distinction between association, predictive usefulness and causation, and draw practical guidance for reading the two kinds of output side by side.

Keywords: *Explanation versus Prediction; ARDL; Machine Learning; Causality; Nigeria.*

JEL Classification: C52; C53; B41; F21.

1. INTRODUCTION

Applied macroeconomic work increasingly runs both a structural econometric model and a machine-learning model on the same data. The two traditions were built for different purposes. ARDL and its relatives estimate structural associations and test for long-run equilibrium, with significance read as evidence for or against a hypothesis. Machine-learning regression minimises out-of-sample error, with feature importance measuring contribution to that task (Shmueli, 2010; Mullainathan & Spiess, 2017).

When the two disagree about the same variable, researchers face a choice: treat it as a failure to be resolved, or recognise that the frameworks were never answering the same question. This paper takes the second view and grounds it in a worked case.

1.1 What this paper contributes

- It grounds the explanation-prediction distinction in two rigorously conducted analyses of identical data, not in a conceptual argument alone.
- It separates statistical association, predictive usefulness and causation, and shows a single variable occupying a different position on each.
- It uses bootstrap confidence intervals for predictive importance instead of point estimates, which turns out to change the picture substantially.
- It offers policy analysts a way to read both kinds of output together without over-claiming from either.

2. LITERATURE REVIEW

2.1 The distinction

Shmueli (2010) gives the canonical statement. Explanatory modelling tests a theoretically motivated structure and is judged on its parameters; predictive modelling minimises expected error on new data and is judged on held-out performance. The two diverge because the bias-variance trade-off that optimises prediction may require deliberately

biased estimators, while unbiased structural estimation can forecast poorly. Breiman (2001a) frames the same split as two cultures of statistical modelling.

Within economics, Mullainathan and Spiess (2017) and Varian (2014) set out what machine learning can and cannot contribute, stressing that predictive performance carries no identification content. Kleinberg et al. (2015) separate policy problems that genuinely need prediction from those that need causal estimates.

2.2 Feature importance and what it measures

Permutation importance measures the loss in accuracy when a feature is shuffled (Breiman, 2001b; Fisher, Rudin & Dominici, 2019). It describes a fitted model on a specific sample, not the data-generating process. Hooker, Mentch and Zhou (2021) show that permuting correlated features forces the model to extrapolate to combinations never observed, so scores can reflect extrapolation behaviour and not reliance. None of this supports reading a high importance score as evidence of a causal channel.

A practical point follows and is central to this paper. Importance is usually reported as a mean across permutation repetitions. That mean has sampling error, and on a short holdout the error can be large enough to leave the sign undetermined. Reporting the point estimate alone hides this.

3. METHOD

This is a comparative methodological paper, not a new estimation. It draws on two companion analyses of one dataset, 2008Q2 to 2024Q4, so that any divergence reflects method and not data.

The dataset behind this paper was originally assembled as 144 quarterly observations running from 1990Q1. Inspection showed that everything before 2008 was linearly interpolated from annual figures: FDI, GDP, the exchange rate, inflation and the trade series all change by a constant amount each quarter and then jump at the year boundary. Interpolation of that kind imposes autocorrelation by construction, so the earlier period has been dropped. Trade openness was also excluded, because the available series divides nominal trade by real GDP and tests as I(2). GDP is rebased in 2025 with an unadjusted splice, so 2025 is dropped as well. The working sample is 2008Q2 to 2024Q4.

The explanatory model is ARDL(1,0,0,0,0), selected by BIC, with coefficients reported under classical and Newey-West standard errors. The predictive model is a support vector regression forecasting FDI one quarter ahead from six core predictors, tuned by expanding-window cross-validation and evaluated on a 13-quarter holdout. Permutation importance is computed with 30 repetitions and accompanied by bootstrap 95% intervals from 1,000 resamples of the holdout observations.

Each variable is then placed on three axes: whether its ARDL coefficient is significant, whether permuting it reliably degrades forecast accuracy, and whether either result supports a causal reading given that both designs are observational.

4. RESULTS

Table 1. The same variables under the two frameworks

Variable	ARDL coefficient (HAC p)	Permutation importance, bootstrap 95% CI	Agreement?
INTEREST	-57.2374 (p = 0.0052)	9.741 [-29.34, 42.94]	No — significant, not predictive
FDI (lagged / current)	0.5009 (p < 0.0001)	-14.993 [-31.27, 5.46]	No — significant, not predictive
EXRVOL	-336.16 (p = 0.6449)	15.890 [-12.35, 44.06]	Yes — neither
OILP	3.6347 (p = 0.2283)	-15.363 [-67.35, 26.64]	Yes — neither
INF	11.2903 (p = 0.4048)	-0.153 [-51.77, 58.40]	Yes — neither
GDPG	-2.3751 (p = 0.7782)	-21.309 [-65.30, 23.29]	Yes — neither

Not one feature has a bootstrap interval excluding zero, including lagged FDI, which is the strongest variable in the ARDL model.

Table 2. Three-way classification

Variable	Significant association?	Predictive contribution established?	Causal claim supported?
INTEREST	Yes (HAC $p = 0.005$)	No (interval spans zero)	No
FDI lagged	Yes (HAC $p < 0.0001$)	No (interval spans zero)	No — persistence, not a channel
EXRVOL	No	No	No
INF, GDPG, OILP	No	No	No

5. DISCUSSION

The interest rate is the cleanest illustration. Its ARDL coefficient is negative, sizeable and significant under a covariance estimator robust to heteroskedasticity and autocorrelation. On the predictive side, shuffling it produces a mean accuracy change of 9.74 with a bootstrap interval running from -29 to +43. The structural evidence says the interest rate belongs in a model of Nigerian FDI. The predictive evidence says nothing either way about whether knowing it helps forecast next quarter.

These are not competing estimates of one quantity. The bounds test asks whether the interest rate participates in a stable level relationship with FDI across 67 quarters. The permutation test asks whether its current value, combined with five others, reduces one-step-ahead error on 13 held-out quarters. A variable can do the first and not the second, particularly when the holdout is short enough that sampling variation swamps a modest contribution.

Two misreadings should be avoided. Treating the null predictive result as evidence that the ARDL finding is wrong would be a mistake, because the bounds test is a well-specified test of a well-defined hypothesis. Treating ARDL significance as licence to expect forecasting value would be the opposite mistake, and would set a policy analyst up to be disappointed by a monitoring tool built on it.

That every bootstrap interval spans zero, lagged FDI included, is itself the most useful thing here. A conventional report of this model would rank the features by mean importance and present the ordering as a finding. On these data the ordering is not distinguishable from noise. The lesson generalises to any short macroeconomic holdout: report intervals, and expect them to be wide.

Neither result supports a causal claim. Both designs are observational, and the predictive model is agnostic to causal structure by design (Pearl, 2009; Varian, 2014). The interest rate may move with the same conditions that drive investment.

6. CONCLUSIONS

A variable can be structurally significant and predictively unestablished, and this is a feature of the two frameworks asking different questions, not a defect in either. For practice: an insignificant ARDL coefficient is a reason for caution about long-run policy claims and not a reason to drop a variable from a monitoring tool; a strong importance ranking is a reason to keep a variable in a forecasting dashboard and not a reason to infer a policy lever; and importance rankings without intervals should be treated as provisional, because on a short holdout they often do not survive one.

7. LIMITATIONS

The argument rests on one dataset and six variables. Replication across other emerging-market series would strengthen it.

The predictive evidence comes from one model class. A different learner might rank features differently, which would not affect the conceptual point but would change the illustration.

The holdout is 13 quarters. Wide intervals are partly a consequence of that, and a longer holdout might establish some of these contributions.

The paper diagnoses the divergence without resolving which reading should govern a given decision. Formal causal methods, such as a synthetic control around the 2023 exchange-rate unification, would address the question policymakers actually have.

REFERENCES

- [1]. Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231.
- [2]. Breiman, L. (2001b). Random forests. *Machine Learning*, 45(1), 5–32.
- [3]. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- [4]. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful. *Journal of Machine Learning Research*, 20(177), 1–81.
- [5]. Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation. *Statistics and Computing*, 31(6), Article 82. <https://doi.org/10.1007/s11222-021-10057-z>
- [6]. Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review: Papers & Proceedings*, 105(5), 491–495.
- [7]. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
- [8]. Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.
- [9]. Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics*, 16(3), 289–326.
- [10]. Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.
- [11]. Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.