

Reporting Uncertainty in Permutation Feature Importance: Why the Usual Dispersion Measure Is the Wrong One

Ofieroher Ufuoma Earnest

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

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

Corresponding Author

Abstract: Permutation feature importance is usually reported as one mean value per feature and ranked, even though the standard implementation produces a dispersion measure alongside every mean. This paper argues that the dispersion normally reported is the wrong quantity, and demonstrates the consequences on a real model. For a support vector regression forecasting Nigerian FDI one quarter ahead, evaluated on 13 held-out quarters, we report permutation importance three ways: the mean, the standard deviation across 30 permutation repetitions, and a bootstrap 95% confidence interval from 1,000 resamples of the test observations. The three disagree in a way that matters. A screening rule based on the ratio of standard deviation to mean, which some authors have proposed, would clear four of six features as reliable. The bootstrap intervals show that not one of the six has an importance whose sign is established, including current FDI, which is the strongest variable in the companion econometric model. The reason is that permutation dispersion measures Monte Carlo noise in the shuffling procedure, which the analyst controls through the repetition count, while the bootstrap measures sampling uncertainty over the test observations, which the analyst does not. We recommend reporting the repetition count, and reporting bootstrap intervals wherever the holdout permits.

Keywords: Permutation Importance; Uncertainty Quantification; Bootstrap; Model Interpretation.

JEL Classification: C52; C53; C15.

1. INTRODUCTION

Permutation importance measures a feature's contribution by shuffling its values and recording the loss in accuracy. Because the shuffling is random, standard implementations repeat it and report a mean and a standard deviation (Breiman, 2001; Pedregosa et al., 2011). Applied papers usually report and rank the mean alone. The obvious fix is to report the standard deviation too, and to flag features whose dispersion exceeds their mean. This paper argues that the obvious fix does not work, and shows what happens when it is applied to a real model.

1.1 What this paper contributes

- It distinguishes permutation Monte Carlo noise from sampling uncertainty, a distinction routinely elided in applied work.
- It shows that a screening rule built on the first quantity is anti-conservative on a real model, clearing four of six features that a proper interval leaves undetermined.
- It implements a bootstrap over test observations and reports intervals for every feature.
- It reports the repetition count and random seed, which most applied papers omit and without which the dispersion figures cannot be interpreted.

2. LITERATURE REVIEW

2.1 Permutation importance

The measure originates with Breiman (2001) as an out-of-bag diagnostic and was generalised by Fisher, Rudin and Dominici (2019) as model-class reliance, with formal treatment of estimation uncertainty. Strobl et al. (2007) document bias under correlated predictors, and Hooker, Mentch and Zhou (2021) show that permuting correlated features forces the model to evaluate combinations never observed in training, so scores can pick up extrapolation and not reliance.

2.2 Two sources of variability

This distinction is the paper's core and is easy to miss. The standard deviation returned by common implementations is computed across the repetition count: it measures how much the score bounces around when the same feature is shuffled repeatedly on the same data. That is Monte Carlo noise, and the analyst controls it. Raising the repetition count from 5 to 500 leaves the reported standard deviation roughly unchanged while shrinking the standard error of the mean by a factor of ten.

A rule based on that standard deviation is therefore partly a function of a tuning choice. The quantity that determines whether a feature genuinely contributes is sampling uncertainty over the test observations, which permutation repetitions do not capture at all. Fisher et al. (2019) develop intervals addressing this directly; resampling the test set (Efron & Tibshirani, 1993) is a practical alternative.

3. METHOD

The dataset was originally assembled as 144 quarterly observations from 1990Q1. Everything before 2008 proved to be linearly interpolated from annual figures, with each series changing by a constant amount within the year and jumping at the boundary, so that period has been dropped. Trade openness was 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. The working sample is 2008Q2 to 2024Q4, 67 quarters.

The source model is a support vector regression selected in a companion forecasting study, using six predictors to forecast FDI one quarter ahead, evaluated on 13 held-out quarters from 2021Q3 to 2024Q3. Permutation importance was computed with 30 repetitions and random seed 42, using RMSE as the scoring metric.

Three quantities are reported per feature. The mean importance. The standard deviation across the 30 repetitions, with its ratio to the mean, which is the screening rule under examination. And a bootstrap 95% interval: the 13 test observations were resampled with replacement 1,000 times, permutation importance recomputed on each resample, and the 2.5th and 97.5th percentiles reported. A feature whose interval excludes zero has an importance whose sign is established by the data.

4. RESULTS

Table 1. Permutation importance three ways (n_repeats = 30, seed 42, bootstrap B = 1,000)

Feature	Mean	SD across repeats	SD / mean	Screening rule verdict	Bootstrap 95% CI	Sign established?
FDI	-14.993	7.728	0.52	Reliable	[-31.27, 5.46]	No
EXRVOL	15.890	11.841	0.75	Reliable	[-12.35, 44.06]	No
OILP	-15.363	9.636	0.63	Reliable	[-67.35, 26.64]	No
GDPG	-21.309	17.492	0.82	Reliable	[-65.30, 23.29]	No
INTEREST	9.741	18.970	1.95	Unreliable	[-29.34, 42.94]	No
INF	-0.153	26.474	172.49	Unreliable	[-51.77, 58.40]	No

The screening rule clears four of six features. The bootstrap establishes the sign of none of them.

5. DISCUSSION

The two columns disagree systematically, and in the dangerous direction. A ratio of standard deviation to mean below one is meant to indicate that a feature's contribution is large relative to its noise. Four features clear that bar: current FDI at 0.52, exchange-rate volatility at 0.75, oil price at 0.63 and GDP growth at 0.82. A paper applying the rule would report these four as reliable and flag only inflation and the interest rate.

The bootstrap intervals span zero for all six. Current FDI, which is the strongest and most significant variable in the companion econometric model, has an interval from -31.27 to 5.46. Its permutation importance is not distinguishable from zero on this holdout, and neither is anything else.

The reason is the one set out in Section 2.2. Thirty repetitions of shuffling the same 13 observations produces a tight distribution, because the observations themselves never change. That tightness says nothing about whether a different 13 quarters would give the same answer, and it is the second question that determines whether the feature matters. The screening rule measures how stable a procedure is, not how reliable its result is, and on this model it clears four features that a proper interval leaves undetermined. Anti-conservative is the right word for it.

A reader might object that all six intervals spanning zero simply means the model is weak, not that the method is wrong. That is correct, and it is the point. The companion forecasting study found that eleven of twelve model specifications lost to a mean benchmark, so a model with no established feature contributions is exactly what should be expected. The screening rule failed to detect this and would have reported four confident-looking rankings from a model with essentially no signal.

None of this is a criticism of permutation importance, which is doing what it claims. The problem is entirely in what gets reported alongside the mean. Two changes fix it and neither is expensive: state the repetition count and seed, without which the dispersion figures cannot be interpreted at all; and compute bootstrap intervals over the test observations, which for a holdout this size takes seconds.

6. CONCLUSION

For a forecasting model of Nigerian FDI, a screening rule based on permutation dispersion cleared four of six features as reliable while bootstrap intervals established the sign of none. The two measure different things: dispersion across repetitions is Monte Carlo noise controlled by the analyst, while the bootstrap captures sampling uncertainty over the observations, which is what determines whether a feature contributes. Applied work should report the repetition count and seed, and should report bootstrap intervals wherever the holdout permits. Importance rankings computed on the short holdouts unavoidable in macroeconomic work should be read as suggestive and not as findings.

7. LIMITATIONS

The holdout is 13 observations. Wide intervals are partly a consequence of that, and the bootstrap over so few points is itself imprecise. A moving-block bootstrap would respect the serial dependence that the i.i.d. version ignores. Importance is computed for one model on one holdout. Variability across model refits and across split boundaries is not captured and is likely substantial.

Several predictors are correlated, and permuting correlated features forces extrapolation to unobserved combinations (Hooker et al., 2021). Conditional or grouped permutation would give a cleaner reading. Only permutation importance is examined. Comparing against SHAP values or model-class reliance intervals would show whether the instability is method-specific or reflects an absence of signal.

REFERENCES

- [1]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [2]. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- [3]. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful. *Journal of Machine Learning Research*, 20(177), 1–81.
- [4]. 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>
- [5]. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.).
- [6]. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [7]. Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures. *BMC Bioinformatics*, 8, 25.