

Overfitting and Extrapolation in Tree Ensembles: Two Distinct Failures in Short-Sample FDI Forecasting

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: ufuoma.earnest@uniport.edu.ng

Abstract: Tree ensembles are recommended for macroeconomic forecasting because they capture non-linear structure, and they are known in principle to be unable to predict outside the range of target values seen in training. This paper documents both properties in one short-sample forecasting problem. Comparing training and test error for three models forecasting Nigerian FDI one quarter ahead on 62 observations, gradient boosting achieves a training RMSE of 2.80 against a test RMSE of 983.64, a ratio of 351 to 1. Random forest and support vector regression degrade far less (2.24 and 2.61 times respectively) from much looser training fits. The extreme ratio for gradient boosting is straightforward overfitting on 49 training observations. Separately, the training targets are strictly positive, running from 58.72 to 3,084.90, while the test period contains four negative quarters and five observations below the training minimum. A tree ensemble trained only on positive targets cannot emit a negative prediction, because its output is a weighted average of training-set leaf values. Both effects are present and they are distinct. We separate them, and note that on a full walk-forward comparison gradient boosting on an enhanced feature set was the best of twelve specifications, so neither effect supports a general claim about tree ensembles in this domain.

Keywords: Gradient Boosting; Random Forest; Overfitting; Extrapolation; Short Samples.

JEL Classification: C53; C45; F21.

1. INTRODUCTION

Random forests (Breiman, 2001) and gradient boosting (Friedman, 2001) are strong general-purpose regressors and are routinely recommended over linear models for macroeconomic forecasting. They also share a structural limit that is well known in principle and rarely checked in practice: because every tree assigns a constant value within each partition, and those constants come from training-set observations, the ensemble cannot produce a prediction outside the range of training targets (Hastie, Tibshirani & Friedman, 2009).

This paper examines a case where that limit binds, and separates it from a second and more mundane problem that occurs in the same models at the same time.

1.1 What this paper contributes

- It decomposes model performance into training and test error instead of reporting test metrics alone.
- It documents a training-to-test RMSE ratio of 351 to 1 for gradient boosting on 49 training observations, which is overfitting rather than extrapolation failure.
- It documents a separate and genuine extrapolation problem: strictly positive training targets against a test period containing negative values.
- It stops short of generalising: the same model class was the best performer under a different feature set and evaluation protocol.

2. LITERATURE REVIEW

2.1 The structural limit

A regression tree partitions the feature space and predicts a constant within each region, computed from the training observations falling there. Averaging (random forest) or additive boosting over such trees produces an output bounded

by the convex hull of training target values. No input can induce a prediction outside that range (Hastie et al., 2009). This is the standard reason for differencing or detrending a trending series before applying tree methods, and it is a hard constraint rather than a tendency.

Kernel methods including support vector regression fit a continuous function defined by support vectors (Vapnik, 1995; Smola & Schölkopf, 2004), which is not bounded in the same way, though extrapolation remains unreliable in practice.

2.2 Overfitting in short samples

A separate problem arises when the number of effective parameters approaches the number of observations. Boosting with several hundred trees on fewer than fifty training points can drive training error arbitrarily low while learning nothing generalisable. The two problems are often conflated in applied work, because both show up as a large train-test gap, but the remedies differ: overfitting responds to regularisation and shallower trees, while the extrapolation limit does not respond to tuning at all.

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 target is FDI one quarter ahead, predicted from six core macroeconomic variables. Both feature sets are restricted to a common window of 62 observations from 2009Q2 to 2024Q3, split chronologically into 49 training and 13 test observations. Three models were tuned by four-fold expanding-window cross-validation, fitted once on the training set, and used to predict both training and test observations so that the gap can be computed directly.

The train-test decomposition necessarily uses single-holdout fits, because a training-set prediction requires one fitted model. Walk-forward refitting does not provide one. Figures here therefore differ from walk-forward figures reported elsewhere in this series, and the two should not be compared directly.

4. RESULTS

Table 1. Training against test RMSE, single-holdout fits

Model	Train RMSE	Test RMSE	Ratio
Support vector regression	345.49	901.95	2.61×
Random forest	331.17	741.97	2.24×
Gradient boosting	2.80	983.64	351.35×

Gradient boosting fits the 49 training observations almost exactly and generalises worst of the three. Random forest and SVR degrade by comparable, unremarkable factors.

Table 2. Where the test observations sit relative to the training range

Statistic	Training set (n = 49)	Test set (n = 13)
Minimum target	58.72	-558.45
Maximum target	3,084.90	1,223.18
Negative values	0 of 49	4 of 13
Below the training minimum	—	5 of 13
Above the training maximum	—	0 of 13

The test maximum sits inside the training range, so the upper tail required no extrapolation. The lower tail did: five test observations fall below the training minimum, and four are negative when no training observation was.

5. DISCUSSION

Two distinct things are happening and they are easy to confuse because both widen the train-test gap.

The first is overfitting, and it is specific to gradient boosting here. A training RMSE of 2.80 on data whose target has a standard deviation in the hundreds means the model has memorised 49 observations. The tuning grid allowed up to 200 trees at depth 3, which is ample capacity to interpolate a sample that small. Random forest, with the same number of trees but bootstrap averaging, kept its training RMSE at 331.17 and generalised better. This one has a tuning solution: fewer trees, shallower depth, stronger shrinkage, or a minimum-leaf constraint scaled to the sample.

The second is the extrapolation limit, and it applies to both tree models regardless of tuning. Every one of the 49 training targets is positive, with a minimum of 58.72. The test period contains four negative quarters, the lowest at -558.45. A tree ensemble cannot produce a negative prediction from strictly positive training leaves; it will compress toward the lower boundary of its training range and take an error roughly the size of the observation. With 13 test points, a handful of such errors dominates RMSE. No amount of regularisation changes this, because it is a property of how the prediction is formed.

The second problem needs a transformation, and no hyperparameter will touch it. Modelling FDI in first differences, or in a signed-log transformation, removes the sign barrier and lets a tree ensemble represent the relevant range. This is testable on these data and is the obvious next step.

One caution against over-reading all of this. Under a full walk-forward comparison on an enhanced feature set, gradient boosting was the best of twelve model-specification pairs and the only one to beat a mean benchmark. The failures documented here are real, but they are properties of this feature set under this evaluation protocol, not evidence that tree ensembles are unsuited to Nigerian FDI forecasting. Model-class rankings in short samples are specification-dependent, and a single comparison does not settle them.

6. CONCLUSIONS

Gradient boosting overfitted 49 training observations to a training RMSE of 2.80 and generalised worst of three models. Separately, both tree ensembles faced a test period containing values below and outside the sign of everything they had seen in training, which their prediction mechanism cannot represent. The two problems are distinct and have different remedies: regularisation for the first, target transformation for the second. Comparative forecasting studies should report training as well as test error, and should check whether test targets fall outside the training range before attributing tree-ensemble underperformance to weak signal.

7. LIMITATIONS

The extrapolation mechanism is argued from the structure of the estimators and the observed ranges, not demonstrated experimentally. Refitting on differenced or signed-log targets would test it directly and is cheap.

One split boundary is examined. Repeating across several would show how much is specific to this 80/20 point.

The test set is 13 observations, so all the metrics here carry wide sampling error.

Only two ensemble methods are examined. Implementations with linear leaf models, or hybrids fitting trees to the residuals of a linear trend, are not bounded in the same way.

REFERENCES

- [1]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- [2]. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155-161.
- [3]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
- [4]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- [5]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), e0194889.
- [6]. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222.
- [7]. Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.