

Ensemble Machine Learning for Early Prediction of Budget Absorption in Kalimantan: An Explainable Stacking Approach

Sofia Nur Ahda¹, Rezka Norhafizah², Silvia Ratna³, Muhammad Muflih⁴, Haldi Budiman⁵

Faculty of Information Technology, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Banjarmasin, Indonesia

Corresponding author: sofianurahda@gmail.com

ABSTRACT

Indonesian subnational governments concentrate a disproportionate share of annual expenditure in the final quarter of the fiscal year, while fiscal oversight remains largely retrospective. This study asks whether each jurisdiction's December expenditure realization can be predicted from mid-year information, and whether such a model can be trusted where an early-warning system would add the most value. We assemble a panel of 27 subnational budget targets spanning five Kalimantan provinces, two government tiers, and three fiscal years (2023–2025), observed at up to six monthly checkpoints, yielding 160 target–checkpoint observations. A stacking ensemble of a random forest and a histogram-based gradient boosting regressor under a ridge meta-learner is evaluated under two leave-one-group-out regimes, with all preprocessing fitted on training folds only and inner folds grouped by target so that the six checkpoints sharing one label cannot leak into the meta-learner. Under leave-one-year-out validation the ensemble attains a target-level R^2 of 0.910 against 0.859 for a naïve seasonal-scaling baseline, although on percentage error the two are effectively tied (20.5% versus 21.0%). Under leave-one-province-out validation the ensemble is outperformed by that same baseline on all four metrics (R^2 0.574 versus 0.832) and fails on the largest fiscal unit (−0.416) and the smallest (−2.908). Grouped-permutation importance assigns the within-year dynamics family an error increase of 2 IDR billion against a base error near 3,900, while the revenue–expenditure gap and jurisdiction identity dominate. A direct ablation confirms the diagnosis: deleting jurisdiction identity costs 0.019 R^2 temporally but recovers 0.129 R^2 spatially, and an absorption-rate target drives R^2 to −1.485. In small subnational panels, temporal validation alone materially overstates what a predictive fiscal-monitoring model knows.

Keywords: Year-End Expenditure Realization, Stacking Ensemble, Explainable Machine Learning, Fiscal Decentralization, Cross-Validation Design.

1. INTRODUCTION

The rhythm of public spending in Indonesia is markedly uneven. Budget execution crawls through the first half of the fiscal year and then accelerates sharply toward its close, producing a pronounced December spike. For the five provinces of Kalimantan over 2023–2025, December alone absorbs on average 19.5% of annual expenditure more than double the 8.3% a uniform disbursement schedule would imply, while January accounts for barely 2% (Figure 1). Rushed fourth-quarter disbursement has been linked to weaker spending quality and measurable macroeconomic side effects: faster absorption is associated with lower regional inflation (Adrison & Flukeria, n.d.), and micro-level evidence from 13,080 spending units traces the surge to risk-averse responses to budgetary uncertainty, each additional external budget revision raising absorption by 1.697 percentage points (“Yalisman,” n.d.).

The policy problem is less the existence of the rush than the inability to anticipate it. Fiscal monitoring in Indonesia is overwhelmingly retrospective: realization is tabulated after quarterly or annual closes, by which point the levers for intervention have largely lost their purchase. The gap matters in a decentralized architecture where much subnational spending is financed by intergovernmental transfers whose own schedule shapes local execution (Lewis &

Oosterman, n.d.), where reliance on those transfers conditions how provinces absorb fiscal stress (Desdiani et al., 2022), and where dependence on central grants is associated with lower expenditure efficiency (Sugiyarto et al., 2025). The need is not hypothetical: the study quantifying the behavioural mechanism closes by calling on the Ministry of Finance to develop a data-driven system able to track and predict spending behaviour (“Yalisman,” n.d.).

Machine-learning methods suit fiscal forecasting well, from euro-area public-debt projection (Barbier-Gauchard & Sofianos, n.d.) to national government-expenditure forecasting (Yang et al., 2023), smart-budgeting applications (Valle-Cruz et al., 2022), and early-warning tasks such as tax-default prediction (Abedin et al., 2021). Yet the literature is skewed toward national aggregates in advanced economies: of 69 studies comparing budget-forecasting methods over 1983–2024, 43% focus on the United States (Kara & Şengüler, 2025). Two works define the space this paper occupies. (Maličká, 2024) applies decision-tree learning to 2,926 Slovak municipalities, an in-sample explanatory exercise on a large cross-section. (De Oliveira Nonato et al., 2026) estimate a hybrid workflow on Brazilian federal expenditure with $N = 24$ and report that randomised cross-validation yields a mean R^2 of 0.97 while an anti-leakage protocol yields 0.45. Our contribution sits between them: out-of-sample prediction on a small subnational panel, with the validation design itself as the object of study.

Two research questions guide the study. RQ1: can a subnational government's December expenditure realization be predicted accurately from mid-year fiscal information for jurisdictions represented in training? RQ2: does such a model generalize to jurisdictions absent from training, and if not, what does its failure reveal about the source of its skill? These motivate three contributions. We build and validate a stacking ensemble paired with model-agnostic attribution to keep predictions interpretable for fiscal managers (Wang et al., 2024). We subject it to two deliberately different tests, holding out a fiscal year versus a province that tell sharply divergent stories. And rather than inferring the cause from feature importance alone, we test it directly by ablating jurisdiction identity, adding a structural covariate, and re-estimating on a scale-free target. For an early-warning system the distinction is not a technicality: it determines where such a tool can, and cannot, be trusted.

2. METHODS

2.1 Data Source and Panel Structure

Our data derive from the monthly budget-realization series (APBD) of the five Kalimantan provinces compiled by the Directorate General of Fiscal Balance (DJPK), Ministry of Finance of the Republic of Indonesia. Provincial gross regional domestic product at constant prices, used only in Section 3.5, is published by BPS-Statistics Indonesia. All figures and tables are the authors' own computation. For each province we observe two government tiers, the consolidated local government and the provincial government on its own across three fiscal years, giving a nominal $5 \times 2 \times 3 = 30$ series, of which 27 survive cleaning (Section 2.2). Within each target-year we sample the cumulative expenditure position at six monthly checkpoints, January through June, so the model forecasts the December total from mid-year information alone, yielding 160 target checkpoint observations. This is a small sample by machine-learning standards, a constraint shared by other public-budget studies (De Oliveira Nonato et al., 2026). Table 1 summarizes the panel.

Table 1 Composition of the subnational budget panel.

Dimension	Levels	Count
Provinces	East, South, Central, West, North Kalimantan	5
Government tiers	Consolidated local; provincial	2
Fiscal years	2023, 2024, 2025	3
Nominal targets	$5 \times 2 \times 3$	30
Valid targets (post-cleaning)	Three truncated provincial series removed	27
Monthly checkpoints	January–June (cumulative cut-offs)	6
Target–checkpoint observations	Two impossible cumulative positions removed	160

2.2 Data Cleaning and Leakage-Safe Preprocessing

Two classes of reporting artifact are screened before any modelling, identically for every model and for the baseline. First, three provincial-government series; South Kalimantan 2024, East Kalimantan 2024, and Central Kalimantan 2025 report a December value identical to their November cumulative, a signature of reporting truncation; these targets are removed. Second, a cumulative expenditure series cannot exceed its own December total. The condition fires for exactly two observations, both in the North Kalimantan provincial series for 2023, whose May and June figures (17,867 and 25,257 IDR billion) stand against a year-end realization of 3,242. They are removed rather than imputed, so no label information reaches a predictor: December serves purely as a consistency check on the reported input data.

Residual month-on-month decreases are floored at zero, consistent with panel-data guidance: conclusions in panel settings can reverse depending on how outliers are handled (Fan & Peng, 2024). We therefore re-estimated with the rule disabled and target-level R^2 is unchanged at 0.910 temporally and 0.574 spatially. No feature is normalized against a full-year total: every predictor at checkpoint month k uses only cumulative quantities observed through month k . Preprocessing and the naïve baseline are fitted on training folds only, so no held-out unit contributes to its own prediction. All monetary values are in IDR billion.

2.3 Feature Engineering

From each checkpoint we construct eleven predictors in four families plus six identity indicators (Table 2). Level features capture fiscal scale: cumulative own-source revenue (PAD), cumulative intergovernmental transfers (TKDD), and cumulative total expenditure. Gap and ratio features summarize fiscal posture: the cumulative revenue–expenditure gap, the routine- and capital-expenditure ratios, and transfer dependence. Dynamics features track momentum through absorption velocity and month-on-month growth in revenue and transfers. Identity indicators, with the checkpoint month, let the model condition on unit-specific fixed effects. The design deliberately mixes slow-moving scale variables with fast-moving flow variables so the attribution analysis of Section 2.6 can adjudicate which family drives year-end outcomes.

Table 2 Engineered feature set.

Family	Features	n
Level (scale)	Cumulative PAD; cumulative TKDD; cumulative total expenditure	3
Gap & ratio	Revenue–expenditure gap; routine-expenditure ratio; capital-expenditure ratio; transfer dependence	4

Family	Features	n
Dynamics	Absorption velocity (m/m); PAD growth (m/m); transfer growth (m/m)	3
Information cut-off	Checkpoint month	1
Identity	Province indicators (5); government-tier indicator	6
Total		17

2.4 Predictive Model

We adopt a stacked generalization architecture (Wolpert, n.d.) with two base learners of complementary inductive bias: a random forest (Breiman, 2001), robust on a modest heterogeneous panel, and a histogram-based gradient boosting regressor (Ke et al., n.d.), well behaved on small samples. Because their relative ranking varies across fiscal-forecasting applications we combine rather than select between them; ensembles of this form have repeatedly outperformed single models and linear specifications on government-expenditure tasks (Ivaşcu & Ştefoni, 2023). Their out-of-fold predictions are combined by a ridge-regularized meta-learner. The target is strictly positive and spans more than an order of magnitude, so the model is fitted on its natural logarithm and back-transformed inside each training fold. Hyperparameters were fixed a priori (random forest: 300 trees, depth 6, minimum leaf 2; gradient boosting: 300 iterations, depth 3, learning rate 0.05, L2 regularization 1.0; ridge meta-learner $\alpha = 1.0$). No search was performed, so no tuning information can leak across folds.

2.5 Validation Design and Metrics

Because a single random split would flatter a model on such a small, structured panel, we evaluate under two leave-one-group-out regimes (Roberts et al., 2017). Leave-one-year-out (LOYO) holds out one fiscal year and asks how well the model transfers to a new year for jurisdictions it has already seen. Leave-one-province-out (LOPO) holds out an entire province and asks how well it transfers to a jurisdiction it has never seen. A model that has learned genuine spending dynamics should survive both; one that has memorized fiscal scale will pass the first and fail the second. In sample-selection terms the LOPO prediction sample is structurally unobserved during training, since the held-out province's indicator is identically zero throughout (Brewer & Carlson, 2023).

Stacked generalization requires the meta-learner's inputs to be genuine out-of-sample guesses (Wolpert, n.d.). Because each budget target contributes six checkpoint rows carrying an identical label, a random inner split would place rows sharing that label on both sides of the fold and leak it into the level-1 features. All inner folds therefore use GroupKFold grouped by target province $\times$ tier $\times$ fiscal year. Both regimes use the identical feature set; under LOPO the held-out province's indicator is structurally zero in training, which is the object of the experiment rather than a defect, and is examined directly in Section 3.5.

As a reference point we implement a naïve seasonal-scaling baseline projecting the December total by inflating the latest observed cumulative with the December-to-checkpoint ratio, computed from training folds only. Performance is reported with R^2 , MAE, RMSE, and MAPE, the four metrics dominating this literature (Kara & Şengüler, 2025), at the honest target level ($N = 27$), where the checkpoint predictions for each target are averaged before scoring because the six rows are not independent. MAPE is interpreted against the conventional scale: 10–20% good, 20–50% reasonable forecasting.

2.6 Interpretability and Multicollinearity Diagnostics

To open the ensemble to inspection we combine three diagnostics, all computed on out-of-fold predictions so they measure generalization rather than in-sample fit. Permutation importance records the increase in target-level mean absolute error when a feature is shuffled in the held-out fold. Because several level and gap features are mechanically related, single-feature permutation can distribute importance arbitrarily among collinear predictors, and additive attribution methods are known to violate consistency in such settings (Lundberg et al., n.d.). We therefore also report grouped-permutation importance and variance-inflation factors, following the practice of triangulating machine-learning attribution against econometric diagnostics established for local-government fiscal data (Maličká, 2024) and the growing use of interpretable machine learning to explain rather than merely predict subnational fiscal outcomes (Wang et al., 2024).

3. RESULTS AND DISCUSSION

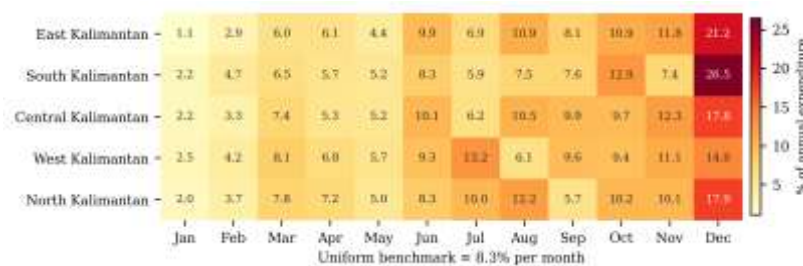


Figure 1 Mean monthly share of annual expenditure by province, Kalimantan, 2023-2025.
December absorbs 19.5% on average against a uniform benchmark of 8.3%

3.1 Predictive Performance under Temporal Validation

Under leave-one-year-out validation the ensemble outperforms the seasonal-scaling baseline at the honest target level, attaining $R^2 = 0.910$ against 0.859 (Table 3). Mean absolute error falls from 4,501 to 3,925 IDR billion, about 13%, and root mean squared error from 6,681 to 5,328, about 20%. Figure 2(a) shows the fit directly: predicted and observed year-end expenditure track the identity line across the full range of provincial and consolidated fiscal scales, with no systematic curvature.

On percentage error the picture changes. MAPE is 20.5% for the ensemble and 21.0% for the baseline: on this metric the two are effectively tied. Because MAPE weights each jurisdiction equally while R^2 and RMSE are dominated by the largest units, the divergence is informative, the ensemble's advantage comes mainly from fitting the big provinces better, not from serving every jurisdiction better. On the conventional interpretation scale both models sit at the boundary between the good and the reasonable band. We therefore describe the advantage as a modest improvement within a moderate-accuracy regime and report the asymmetry across metrics explicitly. It is stable across held-out years 0.924, 0.908, and 0.891 against baseline values of 0.863, 0.844, and 0.874 (Table 4), so it is not carried by any single year.

Table 3 Target-level predictive performance and robustness experiments (IDR billion for MAE and RMSE; $N = 27$ targets; the rate-target row is expressed in absorption-rate units).

Regime	Specification	R^2	MAE	RMSE	MAPE
LOYO (temporal)	Stacking ensemble	0.910	3,925	5,328	20.5%
LOYO (temporal)	Seasonal baseline	0.859	4,501	6,681	21.0%
LOYO (temporal)	Ensemble, identity ablated	0.891	4,134	5,885	21.1%
LOYO (temporal)	Ensemble, absorption-rate target	-1.485	0.115	0.128	13.6%
LOPO (spatial)	Stacking ensemble	0.574	7,286	11,611	50.9%
LOPO (spatial)	Seasonal baseline	0.832	4,883	7,306	22.2%
LOPO (spatial)	Ensemble, identity ablated	0.703	5,756	9,707	39.2%
LOPO (spatial)	Ensemble + lag-1 GRDP	0.595	7,327	11,325	52.0%

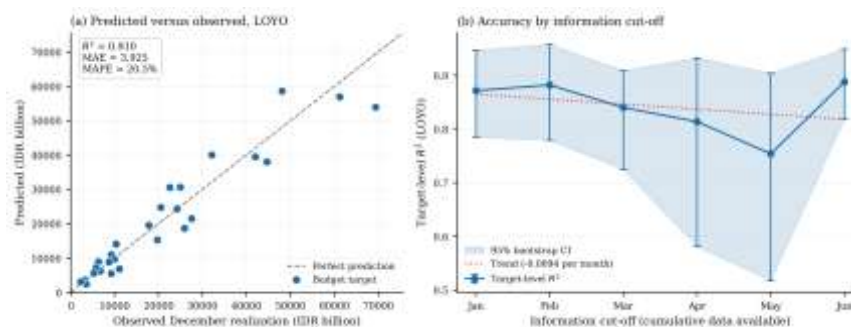


Figure 2 (a) Predictive versus observed December expenditure for the 27 targets under leave-one-year-out validation; each point is the average of that target's monthly checkpoint predictions. (b) Target-level R^2 by information cut-off, each from a model refitted on that single checkpoint, with 95% bootstrap confidence intervals from 2,000 resamples.

3.2 Accuracy Does Not Improve with More Within-Year Data

If the model were reading the tempo of within-year execution, its forecasts should sharpen as more of the year is observed. They do not. Refitting the model at each single information cut-off, target-level R^2 is 0.872 in January, peaks at 0.882 in February, dips to 0.754 in May, and reaches 0.888 in June, with an ordinary-least-squares slope of $-0.0094 R^2$ per month (Figure 2(b)). All six 95% bootstrap intervals overlap substantially, so the differences are not statistically distinguishable: first- and second-quarter information adds no detectable predictive value beyond what is already available in January. This is the first indication that the ensemble is anchored on a quantity fixed early in the year rather than exploiting spending dynamics.

3.3 The Spatial-Generalization Collapse

Leave-one-province-out validation exposes the fragility behind the temporal accuracy. Target-level R^2 falls from 0.910 to 0.574, and the naïve baseline outperforms the ensemble on all four metrics: R^2 0.832 against 0.574, MAE 4,883 against 7,286, RMSE 7,306 against 11,611, and MAPE 22.2% against 50.9% (Table 3). A jurisdiction the model has never seen is better served by a rule that simply inflates its latest observed cumulative by a historical December ratio than by the ensemble.

The breakdown concentrates at the extremes of the scale distribution. The ensemble transfers acceptably to the mid-sized provinces West Kalimantan 0.809, Central Kalimantan 0.910, South Kalimantan 0.924, but fails on the largest fiscal unit, East Kalimantan (-0.416),

and collapses on the smallest, North Kalimantan (−2.908; Table 4). Such heterogeneity is not peculiar to this setting: comparable studies report strong aggregate accuracy alongside far higher prediction uncertainty for particular units (Agan et al., 2025), and behavioural evidence on Indonesian year-end spending finds the response to budget uncertainty itself varies across islands and sectors (“Yalisman,” n.d.). What the leave-one-province-out design adds is that the heterogeneity here is systematic in fiscal scale rather than idiosyncratic.

Table 4 Coefficient of determination by held-out fold, computed from the same validation run as Table 3.

Regime	Held-out unit	n	Ensemble R ²	Baseline R ²
LOYO	Fiscal year 2023	10	0.924	0.863
LOYO	Fiscal year 2024	8	0.908	0.844
LOYO	Fiscal year 2025	9	0.891	0.874
LOPO	West Kalimantan	6	0.809	0.450
LOPO	South Kalimantan	5	0.924	0.625
LOPO	Central Kalimantan	5	0.910	0.799
LOPO	East Kalimantan (largest)	5	−0.416	0.702
LOPO	North Kalimantan (smallest)	6	−2.908	0.800

Read alongside Sections 3.2 and 3.4, the most plausible interpretation is that the ensemble performs well across years because the jurisdictions themselves persist: a province seen in two training years reappears, at roughly its usual scale, in the held-out year. Remove the province entirely and that anchor is gone. High temporal R² therefore reflects the persistence of province-specific fiscal scale far more than any learned signal about mid-year momentum, precisely the failure mode that leave-one-province-out validation is designed to catch (De Oliveira Nonato et al., 2026; Roberts et al., 2017). Section 3.5 tests this reading directly rather than leaving it as inference.

3.4 What the Model Actually Uses

The attribution analysis corroborates this reading. Single-feature permutation importance (Figure 3(a)) is dominated by the cumulative revenue–expenditure gap, which raises target-level MAE by $5,801 \pm 778$ IDR billion when shuffled more than four times the next feature, the government-tier indicator ($1,336 \pm 330$), followed by transfer dependence (703 ± 227). Because level and gap features are mechanically related, we verify this two ways. Grouped permutation (Figure 3(b)) gives the gap-and-ratio family $5,925 \pm 698$ and identity $2,300 \pm 375$, against 184 for levels, 87 for the information cut-off, and 2 for dynamics; against a base MAE near 3,900, the contribution of within-year dynamics is indistinguishable from zero. The variance-inflation diagnostic (Figure 4(a)) makes the dependence explicit: cumulative transfers (420.7), cumulative expenditure (299.3), the gap (100.6), and own-source revenue (43.2) lie far above the threshold of 10, while ratio and velocity features remain below 5. We therefore read the gap's dominance not as the causal primacy of one variable but as evidence that the model keys on the fiscal-scale subspace these collinear features jointly span (Figure 4(b)).

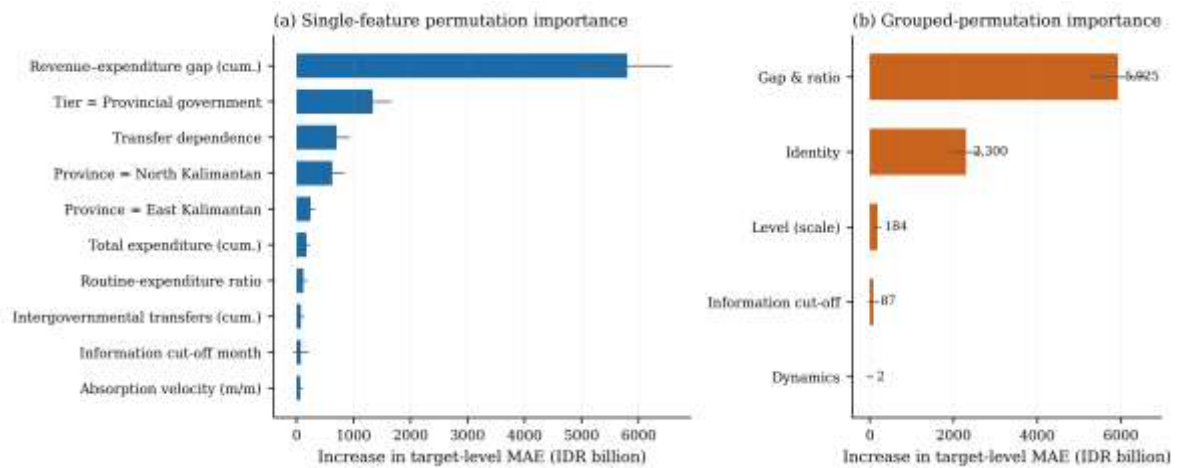


Figure 3 Feature importance for the stacking ensemble. (a) Single-feature permutation importance, top ten features; (b) grouped-permutation importance by family. Computed on out-of-fold predictions; error bars show ± 1 SD across 50 permutations.

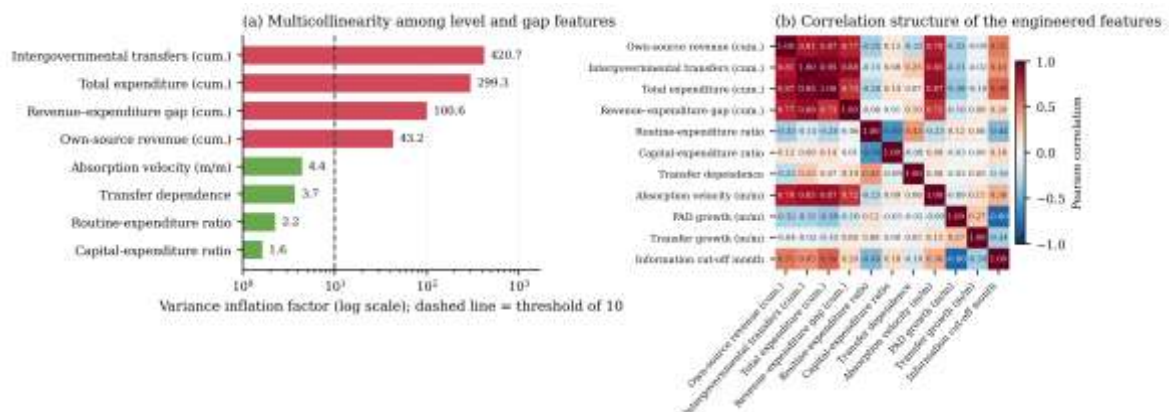


Figure 4 Multicollinearity diagnostics. (a) Variance-inflation factors on a logarithmic scale; the dashed line marks the conventional threshold of 10. (b) Pearson correlation matrix of the engineered features.

In substantive terms, the model predicts year-end spending chiefly from how large a jurisdiction is and how wide its running revenue–expenditure gap is, not from how fast it is spending in a given month. This is a known regularity: in euro-area public-debt forecasting the most influential predictors are past values of public debt itself (Barbier-Gauchard & Sofianos, n.d.); on a far larger expenditure dataset both regression and random forests struggled to disentangle sector-specific effects (Guariso et al., 2023); and in Slovak municipalities the size category of the jurisdiction is itself a significant determinant (Maličká, 2024). The finding is consistent with evidence that fiscal scale and intergovernmental-transfer structure govern subnational spending outcomes (Lewis & Oosterman, n.d.) and with Kalimantan's heavy reliance on central transfers (Sugiyarto et al., 2025).

3.5 Robustness: Identity Ablation, Structural Covariate, and Rate Target

Three experiments test the diagnosis directly rather than by inference (Table 3). Identity ablation: if the ensemble's temporal skill is anchored on jurisdiction identity, deleting the province and tier indicators should cost little temporally and help substantially under spatial holdout. It does. LOYO R^2 falls only from 0.910 to 0.891, while LOPO R^2 rises from 0.574 to 0.703; the collapse on North Kalimantan moderates from -2.908 to -0.366 and East Kalimantan recovers from -0.416 to -0.001 . Removing the anchor makes the model worse where the anchor is available and better where it is not the signature of a model leaning on identity rather than execution dynamics.

Structural covariate: adding lag-one gross regional domestic product improves LOPO R^2 only from 0.574 to 0.595, and once identity has been ablated it does not help at all (0.703 to 0.698); continuous information about economic size is a poor substitute for the identity anchor. Rate target: re-estimating on an absorption-rate target removes the scale information the model relies upon and yields $R^2 = -1.485$ under leave-one-year-out. Once scale is stripped away, predictive skill vanishes entirely direct evidence for the paper's central claim. The ceiling is published for the provincial tier only, so this experiment rests on a smaller sample.

Figure 5(a) reports the realized absorption rate itself, December expenditure divided by the approved ceiling for the same fiscal year at the provincial tier. Absorption clusters between 71% and 100%, with North Kalimantan spending its full ceiling in FY2024 but only 71% in FY2025, the widest swing in the panel; truncated series are excluded because their numerator is not a genuine year-end figure. Figure 5(b) shows the ablation: the temporal cost is small, the spatial gain is large, and the two bars move in opposite directions.

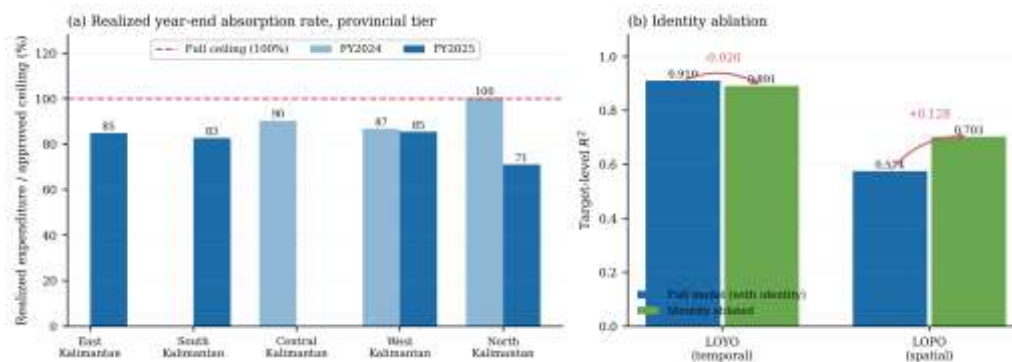


Figure 5 (a) Realized year-end absorption rate at the provincial tier, December realization divided by the approved ceiling for the same fiscal year; truncated series are excluded. (b) Identity ablation: target-level R^2 with and without province and government-tier indicators under both validation regimes.

4. CONCLUSION

We set out to predict the year-end expenditure realization of Kalimantan's subnational governments early enough to be useful, with a model whose reasoning could be inspected. The answer is conditional. Under leave-one-year-out validation the ensemble forecasts December expenditure from mid-year data with a target-level R^2 of 0.910, outperforming a naïve seasonal baseline on R^2 , MAE, and RMSE though on percentage error the two are effectively tied, and both sit at the boundary between good and reasonable forecasting on the conventional scale introduced in Section 2.5 (RQ1). Under leave-one-province-out validation the same ensemble

is outperformed by that baseline on every metric and fails outright on the largest and smallest fiscal units (RQ2). Transparent attribution, a direct identity ablation, and a rate-target re-estimation all locate the same cause: the model's temporal skill rests on persistent fiscal scale and the revenue–expenditure gap, not on within-year spending dynamics.

Pairing temporal with spatial validation is what makes the distinction visible, and the ablation shows the gap between regimes is attributable to a specific, removable component of the feature set. The lesson generalizes: where jurisdictions persist across periods, any evaluation holding out periods rather than jurisdictions rewards memorization of unit identity, and the resulting accuracy describes interpolation within a known set of units rather than the extrapolation an early-warning system requires, a lesson recently drawn from Brazilian federal budget data, where tightening the protocol moved reported accuracy from 0.97 to 0.45. For practice this implies a scoped rather than universal tool. The degradation is also not uniform: the ensemble fails precisely at the extremes of the scale distribution, so such a tool would perform worst for the largest and smallest jurisdictions, raising the concerns about trust, inequity and bias that the literature on artificial intelligence in government has flagged. No prospective deployment, decision threshold, or prediction interval has been evaluated here.

The limitations are substantial. The effective sample is small, 27 targets over three years, so estimates are sensitive to individual jurisdictions and years. The pooled checkpoint observations are not independent, since the six checkpoints of each target share one outcome; we therefore report target-level metrics throughout and group the inner folds accordingly. The targets are nominal expenditure levels, which couples accuracy to fiscal scale, though the rate-target experiment quantifies exactly this coupling. The feature set also omits the budget-revision channel identified as a proximate driver of year-end spikes, which may explain why the dynamics family contributes so little: the trigger may lie outside the variables observable in monthly realization reports. Future work should widen the panel beyond Kalimantan, model absorption rates alongside nominal levels, incorporate budget-revision, electoral-cycle, and procurement covariates, and adopt higher-frequency execution data.

Data and Code Availability.

The panel is built from monthly APBD realization records published by DJPK, Ministry of Finance of the Republic of Indonesia (five Kalimantan provinces, FY2023–2025), and provincial GRDP from BPS-Statistics Indonesia. All figures and tables are the authors' own computation. The complete pipeline is available at [<https://github.com/sofianurahda/kalimantan-yearend-spending>] and runs end to end with one command under a single seed (42), using Python 3.11.15, scikit-learn 1.8.0, NumPy 2.4.4, and pandas 2.3. Tables 3 and 4 are emitted by the same function call.

5. REFERENCES

- Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., Khan, M. I., & Hajek, P. (2021). Tax Default Prediction Using Feature Transformation-Based Machine Learning. *IEEE Access*, 9, 19864–19881. <https://doi.org/10.1109/ACCESS.2020.3048018>
- Adrison, V., & Flukeria, M. (n.d.). *Economics and Finance in Indonesia Economics and Finance in Indonesia Lowering Regional Inflation? Improve Budget Absorption Lowering*

Regional Inflation? Improve Budget Absorption Lowering Regional Inflation? Improve Budget Absorption. 62(2), 67–77. <https://doi.org/10.47291/efi.v62i2.531>

Agan, B., Celik, S., Damak, O. I., & Miba'am, B. (2025). Evaluating the machine learning-based models for predicting carbon neutrality in Sub-Saharan African Nations. *Environment, Development and Sustainability*, 27(11), 28185–28219. <https://doi.org/10.1007/s10668-025-06289-y>

Barbier-Gauchard, A., & Sofianos, E. (n.d.). « *Forecasting Public Debt in the Euro Area Using Machine Learning: Decision Tools for Financial Markets* » *Forecasting Public Debt in the Euro Area Using Machine Learning: Decision Tools for Financial Markets*. Retrieved www.beta-economics.fr

Breiman, L. (2001). *Random Forests* (Vol. 45).

Brewer, D., & Carlson, A. (2023). *Addressing Sample Selection Bias for Machine Learning Methods: Supplemental Appendix*.

De Oliveira Nonato, S., Moreira, M. F., Prata, D. N., De Lima, D. V., & Prata, G. R. N. (2026). An exploratory hybrid AI workflow for Brazilian federal budget allocation. *Data and Policy*, 8. <https://doi.org/10.1017/dap.2026.10062>

Desdiani, N. A., Sabrina, S., Husna, M., Budiman, A. C., Afifi, F. A. R., & Halimatussadiyah, A. (2022). Local Budget Resilience in Times of COVID-19 Crisis: Evidence from Indonesia. *Economies*, 10(5). <https://doi.org/10.3390/economies10050108>

Fan, X., & Peng, Z. (2024). A comparative analysis of the outliers influence using GMM estimation based on dynamic panel data model. *Applied Economics Letters*, 31(2), 170–175. <https://doi.org/10.1080/13504851.2022.2129563>

Guariso, D., Castañeda, G., & Guerrero, O. A. (2023). Budgeting for SDGs: Quantitative methods to assess the potential impacts of public expenditure. *Development Engineering*, 8. <https://doi.org/10.1016/j.deveng.2023.100113>

Ivaşcu, C. F., & Ştefoni, S. E. (2023). Modelling the Non-Linear Dependencies between Government Expenditures and Shadow Economy Using Data-Driven Approaches. *Scientific Annals of Economics and Business*, 70(1), 97–114. <https://doi.org/10.47743/saeb-2023-0001>

Kara, B., & Şengüler, H. (2025). A Comparative Analysis of Budget Forecasting Methods: A Systematic Literature Review Covering the 1983–2024 Period. *Public Budgeting and Finance*. <https://doi.org/10.1111/pbaf.70008>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (n.d.). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Retrieved <https://github.com/Microsoft/LightGBM>.

Lewis, B. D., & Oosterman, A. (n.d.). *The Impact of Decentralization on Subnational Government Fiscal Slack in Indonesia*.

- Lundberg, S. M., Allen, P. G., & Lee, S.-I. (n.d.). *A Unified Approach to Interpreting Model Predictions*. Retrieved <https://github.com/slundberg/shap>
- Maličká, L. (2024). Investigating the Determinants of Local Indebtedness in Slovakia: The Machine Learning Approach. *Ekonomicky Casopis*, 72(7–8), 334–359. <https://doi.org/10.31577/ekoncas.2024.07-08.02>
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. In *Ecography* (Vol. 40, Number 8, pp. 913–929). Blackwell Publishing Ltd. <https://doi.org/10.1111/ecog.02881>
- Sugiyarto, A., Lee, B. L., & Wilson, C. (2025). Narrowing the gap: Dynamics of Indonesia's public expenditure performance in economically lagging districts. *Economic Analysis and Policy*, 85, 2299–2318. <https://doi.org/10.1016/j.eap.2025.03.014>
- Valle-Cruz, D., Fernandez-Cortez, V., & Gil-Garcia, J. R. (2022). From E-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation. *Government Information Quarterly*, 39(2). <https://doi.org/10.1016/j.giq.2021.101644>
- Wang, S., Xiao, Y., & Liang, Z. (2024). Exploring cross-national divide in government adoption of artificial intelligence: Insights from explainable artificial intelligence techniques. *Telematics and Informatics*, 90. <https://doi.org/10.1016/j.tele.2024.102134>
- Wolpert, D. H. (n.d.). *STACKED GENERALIZATION*.
- “Yalisman,” “Khoirunurrofik.” (n.d.). *Risk aversion and budget uncertainty: Empirical evidence on year-end government spending spikes in Indonesia*.
- Yang, C. H., Molefyane, T., & Lin, Y. Da. (2023). The Forecasting of a Leading Country's Government Expenditure Using a Recurrent Neural Network with a Gated Recurrent Unit. *Mathematics*, 11(14). <https://doi.org/10.3390/math11143085>