RoSHAP: Efficient Uncertainty Quantification for Feature Attributions

Anonymous authors
Paper under double-blind review

Abstract

Feature attribution analysis is critical for interpreting machine learning models and supporting data-driven decisions. However, feature attribution values and rankings can vary substantially across train–test splits, random seeds, or model-fitting procedures. The nature of attribution uncertainty makes it challenging to assess the true importance of a feature. We introduce a robust distributional framework **RoSHAP** that treats feature attributions as random quantities. RoSHAP estimates the attribution distribution by refitting the predictive model across multiple data resamples. The resulting distribution captures both attribution magnitude and variability within a unified framework. We show that, under mild regularity conditions, the aggregated attribution score is asymptotically Gaussian, enabling a computationally efficient approach to achieve the desired estimation precision. Simulations and real-data experiments demonstrate that RoSHAP recovers the known feature ordering, quantifies uncertainty in feature importance, and achieves competitive signal identification and predictive performance relative to single-fit and refit-based approaches. By incorporating attribution stochasticity and a computationally efficient approach, RoSHAP provides a reliable and robust metric for assessing feature importance under uncertainty.

1 Introduction

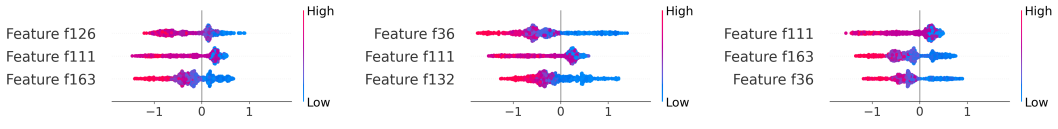
Feature attribution provides a quantitative basis for interpreting machine learning (ML) models by assigning contribution scores to individual predictors. Post hoc methods such as SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), and DeepLIFT (Shrikumar et al., 2017) are widely used when feature contributions cannot be directly derived from the model structure. Among them, SHAP provides additive instance-level attributions grounded in cooperative game theory and satisfies local accuracy, missingness, and consistency (Lundberg and Lee, 2017). These guarantees, however, apply to a fixed fitted model and do not characterize how attributions vary when the model is refit on different subsets of the data.

Accounting for variation across refitted models is important because both the training data and the fitted model can vary stochastically across runs. A standard workflow splits the data into training and test sets, fits a model on the training set, and computes attributions from the fitted model. Each step introduces variation, through the random partition of the data, random initialization, and stochastic optimization. Consequently, attribution scores are not fixed across model fits, and different train–test splits or random seeds can produce different feature rankings (Shaikhina et al., 2021). This variability is particularly pronounced in high-dimensional settings such as genomics, where the number of features far exceeds the number of samples.

Figure 1 illustrates this issue using the Musk (Version 2) molecular chemistry data (Chapman and Jain, 1994). Three train–test splits yield classifiers with similar test AUC, yet the top three SHAP-ranked molecular features and their attribution magnitudes vary across splits. Such variability can complicate downstream decisions when follow-up measurements or experiments are costly.

This variability can itself be informative, as treating feature attribution as a random quantity allows us to estimate its distribution across refits and quantify the uncertainty associated with each feature's importance and rank. We propose **RoSHAP** (Robust SHAP), a resampling-based framework for estimating the distribution of feature attribution scores across model fits. While existing resampling approaches provide useful summaries such as averages, selection frequencies, or null-reference

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Round 1 (b) Round 2 (c) Round 3

Figure 1: Top SHAP-ranked features in the Musk (Version 2) data across different train–test sets. Although the three runs achieve similar test AUC, the selected top features differ substantially.

assessments, RoSHAP complements these perspectives by characterizing the full attribution distribution, establishing its asymptotic Gaussianity, and providing Monte Carlo precision guarantees to guide the number of refits required. RoSHAP repeatedly resamples the data and refits the model to capture both the magnitude and variability of feature attributions. We show that, under mild regularity conditions, the aggregated feature-level attribution score is asymptotically Gaussian, allowing its distribution to be characterized by its mean and variance, and enabling efficient estimation with fewer refits. We further quantify the Monte Carlo precision of the estimated distribution, providing practical guidance for choosing the number of refits and managing computational cost. Although we use SHAP throughout, the framework applies to any real-valued attribution measure computed on held-out data.

We evaluate RoSHAP in simulations with known signal features and on three real benchmark datasets from indoor positioning, chemistry, and nonlinear classification, covering classification and regression tasks and several learners. Across three real datasets, the estimated attribution distributions reveal uncertainty in features' relative importance that single-fit attributions cannot convey, while models retrained on the top-ranked features retain predictive performance comparable to competing rankings. The experiments also support the Gaussian approximation under its stated conditions, with the attribution distributions of leading features closely approximated using only 30 refits, compared with 1,500–3,500 in the full runs. In simulations, RoSHAP recovers up to 31% more true signals than single-fit SHAP in the high-dimensional setting and performs comparably to or better than refit-based feature selectors.

Our main contributions are threefold: 1) we propose RoSHAP, a distributional framework that characterizes both attribution magnitude and variability across model refits; 2) we establish the asymptotic Gaussianity of attribution scores, enabling efficient approximation of attribution distributions with substantially fewer refits; and 3) we derive Monte Carlo precision guidance for selecting the number of refits needed to achieve a desired estimation accuracy.

The remainder of this paper is organized as follows. Section 2 reviews related work on SHAP and other feature contribution methods. Section 3 presents the RoSHAP framework and the Gaussian approximation. Section 4 evaluates RoSHAP on simulated and real-world datasets, examining attribution uncertainty, the Gaussian approximation, ranking accuracy, and predictive performance relative to existing methods. Section 5 concludes with a discussion of the main findings and future directions.

2 Related Work

Feature attribution and additive attribution. Feature attribution is widely used to interpret predictive models. Common importance measures for ML models include gain-, impurity-, and entropy-based scores, as well as post hoc methods such as SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), and DeepLIFT (Shrikumar et al., 2017). SHAP is grounded in cooperative game theory, assigning each feature its average marginal contribution to the prediction across feature coalitions. Its additive formulation supports broad applicability across model classes. TreeSHAP (Lundberg et al., 2020) enables efficient computation for tree ensembles, while KernelSHAP and DeepSHAP extend SHAP to black-box and deep learning models. These attributions are usually reported as point values from a single fitted model.

Variability of attributions. A growing literature examines how feature attributions vary under estimator, data, and model randomness. With the model and data fixed, Goldwasser and Hooker (2024) quantify uncertainty in SHAP estimation and certify top-ranked features against estimator noise. Across refitted models, Shaikhina et al. (2021) summarize sample-level attributions over training subsamples and show that predictive uncertainty can translate into explanation uncertainty.

Watson et al. (2023) extend SHAP with information-theoretic value functions to attribute epistemic and aleatoric uncertainty.

Resampling-based attribution. Resampling is commonly used to stabilize or assess feature attribution. Some studies summarize SHAP values across repeated resampling or nested cross-validation using averages, selection frequencies, or rank distributions (Scheda and Diciotti, 2022; Halabi et al., 2026). Other methods assess SHAP-based importance against null references generated through repeated fitting. PowerShap (Verhaeghe et al., 2022) introduces a random noise feature, BorutaShap (Keany, 2020) relies on shadow features, and shapicant (Calzolari, 2022) uses target permutations. Across these approaches, repeated attributions are typically reduced to summary statistics, ranks, or test decisions rather than characterized as a distribution with explicit estimation precision.

Overall, prior work has shown that feature attributions vary across data samples and model fits, and has used resampling to summarize or test this variability. We instead make the replicate distribution itself the object of analysis, providing a unified view of attribution magnitude, variability, and estimation precision under practical computational budgets.

3 METHODOLOGY

We introduce RoSHAP, a simple distributional framework that accounts for the stochastic nature of feature attribution and captures variability across data resampling and model refitting. RoSHAP uses repeated resampling to construct an empirical attribution distribution for each feature, allowing feature importance to be characterized through its signal strength and variability. Under mild regularity conditions, the attribution distribution admits a Gaussian approximation, providing a compact parametric characterization and substantially reducing the computational burden of distribution estimation. We further derive the relationship between the number of replicates and the precision of summary statistics, providing practical guidance for allocating the computational budget. The overall pipeline is illustrated in Figure 2.

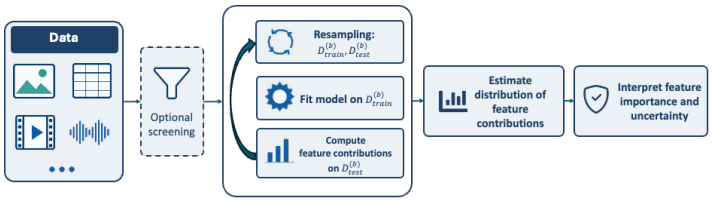


Figure 2: Overall framework for distributional feature attribution estimation.

Resampling Estimation of Feature Attributions. Let $\mathcal{D}_n = \{(x^{(i)}, y_i) : i = 1, \ldots, n\}$ denote a labeled dataset with n observations and d features. For an evaluation set $\mathcal{O} \subset \{1, \ldots, n\}$, let $\hat{f}_{\mathcal{O}}$ denote the model fitted on $\mathcal{D}_n \setminus \mathcal{O}$, and let T_{ij} denote the attribution of feature j under $\hat{f}_{\mathcal{O}}$ at $x^{(i)}$, $i \in \mathcal{O}$.

Algorithm 1. Resampling-based attribution distribution estimation

Require: Data $\mathcal{D}_n$; rounds B; evaluation size k
Ensure: Sample- and feature-level SHAP distributions
1: Preprocess $\mathcal{D}_n$; optionally screen predictors
2: **for** $b = 1, \ldots, B$ **do**
3: Draw $\mathcal{O}^{(b)} \subset \{1, \ldots, n\}$, $|\mathcal{O}^{(b)}| = k$, uniformly without replacement (within classes for classification)
4: Fit $\hat{f}^{(b)}$ on $\mathcal{D}_n \setminus \mathcal{O}^{(b)}$
5: Compute $T_{ij}^{(b)}$ for $i \in \mathcal{O}^{(b)}$
6: Set $U_j^{(b)} = \sum_{i \in \mathcal{O}^{(b)}} |T_{ij}^{(b)}|$
7: **end for**
8: Estimate the distributions of T_{ij} and U_j from $\{T_{ij}^{(b)} : i \in \mathcal{O}^{(b)}\}$ and $\{U_j^{(b)}\}_{b=1}^{B}$

In each of B rounds, an evaluation set $\mathcal{O}^{(b)}$ of fixed size k is sampled uniformly without replacement, stratified by class for classification. The model $\hat{f}^{(b)} = \hat{f}_{\mathcal{O}^{(b)}}$ is refitted with learner randomness drawn independently in each round, and attributions $T_{ij}^{(b)}$ are computed only for $i \in \mathcal{O}^{(b)}$. The feature-level attribution is $U_j^{(b)} = \sum_{i \in \mathcal{O}^{(b)}} |T_{ij}^{(b)}|$. Conditional on $\mathcal{D}_n$, the B rounds are i.i.d., so $\{U_j^{(b)}\}_{b=1}^{B}$ forms a sample from the replicate distribution of $U_j = \sum_{i \in \mathcal{O}} |T_{ij}|$, and, for each observation i, the values $T_{ij}^{(b)}$ from the rounds with $i \in \mathcal{O}^{(b)}$ form a sample from the distribution of

T_{ij}. Algorithm 1 summarizes the procedure. The framework applies to any real-valued attribution computed on observations excluded from model fitting, with SHAP used as a representative example.

Gaussian Approximation of Feature-level Attributions. Under suitable conditions, the replicate distribution of U_j admits a Gaussian approximation. Conditional on $\mathcal{D}_n$, U_j varies across draws of $\mathcal{O}$, which determine the corresponding training and evaluation sets. For randomized learners, U_j may also vary with the internal randomness of model fitting. For deterministic learners, only the draw of $\mathcal{O}$ varies across replicates. Define $\mu_j = E[U_j]$, $\sigma_j^2 = \mathrm{Var}(U_j)$, and let

$$\widetilde{U}_j = \mu_j + \sum_{i=1}^{n} \left(E[U_j \mid \mathbf{1}\{i \in \mathcal{O}\}] - \mu_j\right)$$

denote the Hájek projection of U_j onto the inclusion indicators.

Assumption. *For a fixed feature j, as $n \to \infty$:*

- *(A1)* Sampling. *$\mathcal{O}$ is sampled uniformly without replacement, $k/n \to \lambda \in (0,1)$.*
- *(A2)* First-order accuracy. $\mathrm{Var}(E[U_j \mid \mathcal{O}] - \widetilde{U}_j) = o(\sigma_j^2)$, *where* $\sigma_j^2 > 0$.
- *(A3)* Non-dominance. $\max_{i \le n} \mathrm{Var}(E[U_j \mid \mathbf{1}\{i \in \mathcal{O}\}]) = o(\mathrm{Var}(\widetilde{U}_j))$.
- *(A4)* Gaussian learner noise. *Either* $E[\mathrm{Var}(U_j \mid \mathcal{O})] = o(\sigma_j^2)$, *as for a learner that is deterministic given $\mathcal{O}$, or, conditional on $\mathcal{O}$,*

$$\frac{U_j - E[U_j \mid \mathcal{O}]}{\mathrm{Var}(U_j \mid \mathcal{O})^{1/2}} \xrightarrow{d} \mathcal{N}(0,1) \qquad \textit{and} \qquad \frac{\mathrm{Var}(U_j \mid \mathcal{O})}{E[\mathrm{Var}(U_j \mid \mathcal{O})]} \xrightarrow{p} 1.$$

Theorem. *Under (A1)–(A4), conditional on $\mathcal{D}_n$,*

$$\frac{U_j - \mu_j}{\sigma_j} \xrightarrow{d} \mathcal{N}(0,1).$$

A proof is given in Appendix A.1. Condition (A1) holds by design. Conditions (A2) and (A3) require that, after averaging over learner randomness, the first-order resampling effect dominate the variation in U_j, with no single observation dominating. Condition (A4) requires approximately Gaussian learner randomness with split-invariant variance, and holds trivially for deterministic learners. The replicate distribution of U_j is then approximately Gaussian, with μ_j and σ_j^2 estimated by the sample mean and variance of $\{U_j^{(b)}\}_{b=1}^{B}$. (A2)–(A3) can be assessed from the replicates, and (A4) from repeated fits at a fixed split.

Replicate precision of the location and spread estimates. Conditional on $\mathcal{D}_n$, the replicates $U_j^{(1)}, \ldots, U_j^{(B)}$ are i.i.d. draws from the replicate distribution of U_j, since both the resampling sets and any learner randomness are independent across rounds. The location μ_j and spread σ_j are estimated by

$$\widehat{\mu}_j = \frac{1}{B}\sum_{b=1}^{B} U_j^{(b)}, \qquad \widehat{s}_j^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(U_j^{(b)} - \widehat{\mu}_j\right)^2.$$

Proposition (Replicate precision)**.** *Conditional on $\mathcal{D}_n$, let $\sigma_j^2 = \mathrm{Var}(U_j) > 0$ and $\mu_{4,j} = E(U_j - \mu_j)^4$. Then $\mathrm{Var}(\widehat{\mu}_j) = \sigma_j^2/B$, and as $B \to \infty$ with $\mathcal{D}_n$ fixed,*

$$\sqrt{B}\left(\widehat{\mu}_j - \mu_j\right) \xrightarrow{d} N\left(0, \sigma_j^2\right), \qquad \sqrt{B}\left(\widehat{s}_j - \sigma_j\right) \xrightarrow{d} N\left(0, \tfrac{\mu_{4,j} - \sigma_j^4}{4\sigma_j^2}\right).$$

The Monte Carlo standard errors are estimated by

$$\widehat{\mathrm{SE}}(\widehat{\mu}_j) = \frac{\widehat{s}_j}{\sqrt{B}}, \qquad \widehat{\mathrm{SE}}(\widehat{s}_j) = \frac{\sqrt{\widehat{\mu}_{4,j} - \widehat{s}_j^4}}{2\,\widehat{s}_j\sqrt{B}},$$

where $\widehat{\mu}_{4,j} = B^{-1}\sum_{b=1}^{B}(U_j^{(b)} - \widehat{\mu}_j)^4$.

The proposition does not require a Gaussian replicate distribution. Conditional on the observed data, the required moments are finite, and $\sigma_j^2 > 0$ excludes only a constant replicate distribution. The Monte Carlo standard errors for both location and spread decrease at the $B^{-1/2}$ rate, allowing B to be chosen to achieve a desired estimation precision.

4 EXPERIMENTS

The experiments establish three main results. First, the RoSHAP distribution result reveals attribution variability and uncertainty that are concealed by any single fitted model. Second, when the Gaussian diagnostics are satisfied, the approximation closely reproduces the full replicate distribution using substantially fewer replicates. This allows the required replicate budget to be evaluated in terms of approximation accuracy and estimation precision. Third, the resulting distributional summaries are useful for feature ranking, as demonstrated by recovery of the true feature ordering in simulation and by downstream predictive performance on real datasets.

4.1 EXPERIMENTAL SETUP

Real datasets. Three real datasets are used to represent distinct regimes: UJIIndoorLoc (Torres-Sospedra et al., 2014), a WiFi-fingerprint regression task with 19,937 training samples and 520 access-point signal-strength features, using longitude as the response; MADELON (Guyon, 2004), a nonlinear classification benchmark with 2,600 samples and 500 features; and Musk (Version 2) (Chapman and Jain, 1994), an imbalanced classification task with 6,598 conformations, 166 features, and a 15% positive class. All methods and all diagnostics see only the training data, while the held-out rows are reserved exclusively for evaluating downstream performance. Preprocessing and data-split details are provided in Appendix A.2.

Simulated data. The simulation considers binary classification with $Y \sim \mathrm{Bern}(0.5)$ and zero-inflated Gaussian predictors. Signal features follow $X_j \mid Y{=}y \sim \pi\delta_0 + (1-\pi)\mathcal{N}(\mu_{jy}, \sigma^2)$, with $\mu_{j1} = -\mu_{j0}$, the signal magnitudes μ_{j1} decrease evenly from 1.5 to 0.4, inducing a known ordering of feature importance. With the sample size fixed at $n = 600$, the total number of features d and the number of signal features s define three regimes. The reference setting uses $(d, s) = (100, 10)$, the dense setting uses $(100, 50)$, and the high-dimensional setting uses $(1000, 50)$. Parameters are set to $(\sigma, \pi) = (3, 0.3)$ for signal features and $(1, 0.2)$ for noise features, keeping individual signal effects relatively weak. Each regime is evaluated over thirty independently generated datasets, with the dataset serving as the unit of replication.

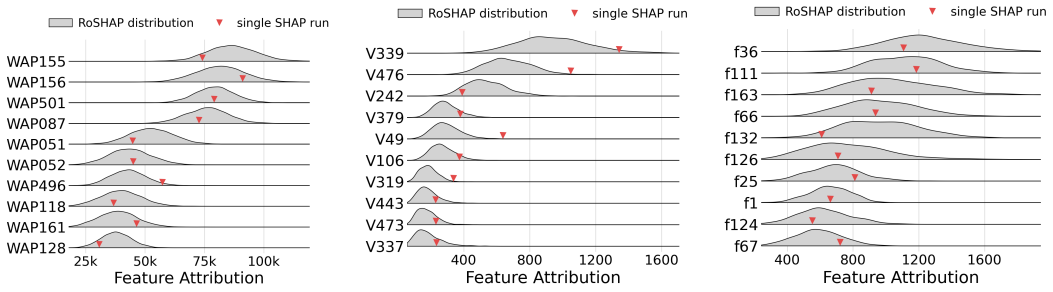
Models and RoSHAP configuration. The simulation uses a fixed XGBoost (Chen and Guestrin, 2016) configuration, while each real dataset is evaluated with five learners, including LightGBM (Ke et al., 2017), XGBoost, CatBoost (Prokhorenkova et al., 2018), random forest (Breiman, 2001), and a penalized linear model. Each RoSHAP round refits the model on a new resample and yields one attribution replicate, and features are ranked by their average within-round rank across replicates. RoSHAP uses $B = 30$ rounds per simulated dataset. For the real data, $B = 3{,}500$ rounds are used for UJIIndoorLoc and $B = 1{,}500$ for MADELON and Musk. For each real dataset–learner pair, the complete set of B rounds is referred to as the *full run*, while a *reduced budget* of $\tilde{B} < B$ replicates is obtained by subsampling the full run. The larger budget for UJIIndoorLoc reflects its substantially larger sample size and provides a sufficiently stable reference for the diagnostic and reduced-budget comparisons. The Gaussian-approximation diagnostics use the same criteria across all datasets and learners. Hyperparameters, diagnostic procedures, and implementation details are provided in Appendix A.3.

Baselines. The comparison includes single-fit SHAP, the learner's native importance, LIME, PowerShap (Verhaeghe et al., 2022), Shapicant (Calzolari, 2022), BorutaShap (Keany, 2020), and the statistical significance ranking method of Goldwasser and Hooker (2024). Features are ranked by each method's native score: summed absolute SHAP values, built-in importance, mean absolute LIME weight, selection p-value for PowerShap and Shapicant, shadow-feature wins for BorutaShap, and the single-fit ranking with its number of verified ranks for Goldwasser and Hooker (2024). Further details are provided in Appendix A.4.

4.2 ATTRIBUTION VARIABILITY AND UNCERTAINTY

Figure 3 visualizes the RoSHAP attribution distributions for one representative learner on each real dataset. Each panel shows the distributions of the top 10 features (grey), ranked by average rank across replicates, together with their corresponding single-fit SHAP attributions (red triangles). Across all three datasets, the displayed replicate distributions appear approximately Gaussian.

The replicate distributions provide information beyond point estimates. Their spread reflects sensitivity to refitting, overlap reflects uncertainty in relative importance, and long right tails indicate that attribution magnitude can depend strongly on the fitted sample. On UJIIndoorLoc, WAP155 and WAP156 have similar replicate means and substantial variability in their within-round difference, with WAP155 exceeding WAP156 in only 61% of refits, despite the single fit ranking them third and first. In contrast, WAP087 and WAP051 are more clearly separated relative to their variability, with WAP087 exceeding WAP051 in 97% of refits. On MADELON, V49 and V379 have nearly identical replicate means relative to the spread of their difference, yielding a 52% exceedance probability, although the single fit assigns 1.7 times greater importance to V49. Similarly, on Musk (Version 2), f66 and f132 have similar means relative to their variability (52%) despite single-fit ranks of third and fourteenth, whereas f36 and f126 show clearer separation across refits (89%).



(a) UJIIndoorLoc, CatBoost (b) MADELON, LightGBM (c) Musk (Version 2), XGBoost

Figure 3: RoSHAP attribution distributions for the ten features with the best average rank, shown for one representative dataset–learner pair per dataset. Grey curves are KDEs over RoSHAP replicates. The number of replicates is $B = 3{,}500$ for UJIIndoorLoc and $B = 1{,}500$ for MADELON and Musk (Version 2). Red triangles show single-fit SHAP attributions from a model fitted and explained on the full training set, rescaled within each panel to match the average total attribution across replicates.

4.3 Gaussian diagnostic and computational efficiency

Assumption diagnostics. Assumptions (A1)–(A4) are evaluated empirically, with normality further assessed using Q–Q plots and the Shapiro–Wilk test. Assumption (A1) holds by construction because each replicate samples half of the training observations uniformly without replacement. The remaining conditions are assessed through the approximate variance decomposition

$$\sigma_j^2 \approx \mathrm{Var}(\widetilde{U}_j) + \mathrm{Var}\Big(E[U_j \mid \mathcal{O}] - \widetilde{U}_j\Big) + E[\mathrm{Var}(U_j \mid \mathcal{O})]\,,$$

corresponding to the first-order split effect, higher-order split effect, and learner randomness. The first-order share and the (A3) no-dominant-observation ratio are estimated directly from stored replicates, while the learner-randomness share is estimated by varying the random seed at fixed splits. The remaining variance provides a diagnostic for (A2). Full definitions and estimation details are given in Appendix A.5.

Taking UJIIndoorLoc with CatBoost as an example, learner randomness accounts for most of the replicate variance, while the higher-order split contribution is not distinguishable from zero, supporting (A2). The (A3) ratios are small, indicating that no individual observation dominates the first-order variation. Seed-only draws are approximately Gaussian with similar variance across the examined splits, supporting (A4). Consistent with these diagnostics, the Q–Q plots in Figure 4 and Shapiro–Wilk tests indicate that the RoSHAP replicate distributions are approximately Gaussian for the leading features.

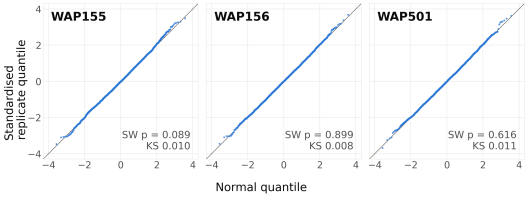


Figure 4: QQ-plot verifies the distributions are Gaussian, and it passed the SW test.

Efficient Gaussian approximation. When the Gaussian approximation holds, the replicate distribution of feature j is characterized by μ_j and σ_j^2, allowing it to be estimated from substantially

fewer replicates. To quantify this gain, Gaussian distributions are fitted using $\tilde{B} = 10$, 30, and 100 randomly sampled replicates and compared with the empirical distribution from all B replicates using the Kolmogorov–Smirnov (KS) distance over 200 random draws. Figure 5 shows the results for WAP156 on UJIIndoorLoc with CatBoost ($B = 3{,}500$). The median KS distance decreases from 0.13 at $\tilde{B} = 10$ to 0.07 at $\tilde{B} = 30$ and 0.04 at $\tilde{B} = 100$, indicating that 30 replicates already provide a close approximation to the full replicate distribution.

Table 1 reports the KS distance at $\tilde{B} = 30$ for the three top-ranked features on each dataset. Median distances range from 0.07 to 0.09, with similar values across datasets and features. The upper 95th-percentile values remain below 0.18, indicating that even with only 30 replicates, the fitted Gaussian closely approximates the full-run empirical distribution for these leading features.

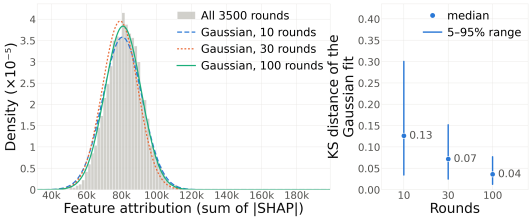


Figure 5: Gaussian approximation using 10, 30, and 100 rounds of resampling. UJIIndoorLoc WAP156 as an example. Left panel shows estimated density overlay with histogram, right panel shows median and 5th–95th percentile.

Table 2 compares rankings from $\tilde{B} = 30$ replicates with the corresponding full-run rankings across learners. Features are ranked by their average rank over the 30 replicates, and agreement with the full-run ranking is measured by Spearman correlation across all features. Results report the median over 200 random draws, with 5th–95th percentiles in brackets. Rank agreement is high across most dataset–learner pairs, with median Spearman correlations generally above 0.9. UJIIndoorLoc and Musk (Version 2) show particularly stable rankings across learners, while MADELON exhibits greater learner dependence, with correlations ranging from 0.80 for Logistic/Ridge to 0.94 for LightGBM and random forest. These results suggest that relatively few replicates can recover the full-run feature ordering in most settings, although the required budget depends on the dataset and learner.

Table 1: Kolmogorov–Smirnov distance between the Gaussian fitted on $\tilde{B} = 30$ replicates and the empirical distribution from all B replicates for the top three features by average rank. The distance ranges from 0 to 1 and lower is better. Values are medians over 200 random draws, with 5th–95th percentiles in brackets.

MADELON, LightGBM		UJIIndoorLoc, CatBoost		Musk (Version 2), XGBoost	
Feature	KS distance	Feature	KS distance	Feature	KS distance
V339	0.08 [0.03, 0.17]	WAP155	0.07 [0.02, 0.14]	f36	0.08 [0.04, 0.18]
V476	0.08 [0.03, 0.16]	WAP156	0.07 [0.02, 0.15]	f111	0.08 [0.04, 0.17]
V242	0.08 [0.04, 0.17]	WAP501	0.08 [0.02, 0.16]	f163	0.09 [0.05, 0.16]

Table 2: Rank stability using $\tilde{B} = 30$ replicates under the average-rank criterion. Entries report the median Spearman correlation between the ranking from 30 replicates and the corresponding full-run ranking across 200 random draws, with 5th–95th percentiles in brackets. Full-run budgets are $B = 1{,}500$ for MADELON, $B = 3{,}500$ for UJIIndoorLoc, and $B = 1{,}500$ for Musk (Version 2).

Learner	MADELON	UJIIndoorLoc	Musk (Version 2)
XGBoost	0.93 [0.92, 0.94]	0.99 [0.99, 0.99]	0.99 [0.99, 1.00]
LightGBM	0.94 [0.93, 0.94]	0.99 [0.99, 0.99]	0.99 [0.99, 0.99]
CatBoost	0.89 [0.87, 0.90]	0.95 [0.94, 0.95]	0.98 [0.98, 0.99]
RandomForest	0.94 [0.93, 0.95]	0.99 [0.99, 0.99]	1.00 [0.99, 1.00]
Logistic / Ridge	0.80 [0.77, 0.82]	1.00 [1.00, 1.00]	0.99 [0.98, 0.99]

Precision evaluation. The preceding results show that a small replicate budget can approximate the attribution distribution and recover stable feature rankings. The remaining question is how precisely its moments can be estimated at such budgets. The replicate-precision proposition gives the standard errors of the estimated mean and standard deviation and the replicate budget required for a target precision without assuming Gaussianity.

For the ten top-ranked features of UJIIndoorLoc (CatBoost), MADELON (LightGBM), and Musk (Version 2; XGBoost), 20,000 samples of $\tilde{B}$ replicates are drawn with replacement from the full run for $\tilde{B} = 10, \ldots, 500$. For each sample, the replicate mean $\widehat{\mu}_j$, replicate standard deviation $\widehat{s}_j$ are computed. Coverage is the proportion of nominal 95% intervals that contain the corresponding full-run moment. The empirical spread of $\widehat{\mu}_j$ and $\widehat{s}_j$ across samples is also compared with the standard errors given by the replicate-precision proposition evaluated at the full-run values. Figure 6 reports coverage for the mean (left) and standard deviation (right), with each curve showing the median across the ten features and the shaded region their range.

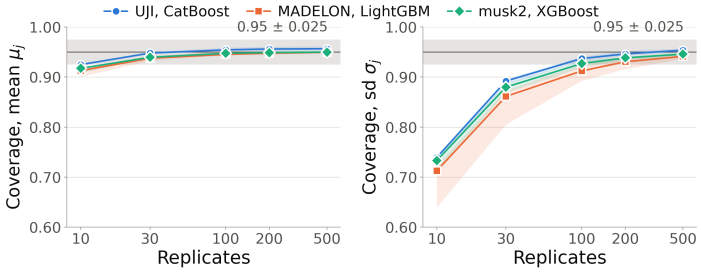


Figure 6: Coverage of nominal 95% intervals for the replicate mean μ_j (left) and standard deviation σ_j (right) as a function of the replicate budget B. Lines show the median across the ten top-ranked features, with shaded bands showing their range. The horizontal line marks the nominal 0.95 coverage level.

The standard-error formulas closely track the empirical variability. For the mean, the observed spread follows the expected $\sigma_j/\sqrt{\tilde{B}}$ rate across budgets. For the standard deviation, agreement is also good by $\tilde{B} = 30$, although interval coverage approaches the nominal level more gradually than for the mean. The distinction is reflected in relative precision. At $\tilde{B} = 30$, the relative standard error is 3–5% for the mean but 13–15% for the standard deviation. At $\tilde{B} = 100$, these decrease to 2–3% and 7–8%, respectively. Thus, 30 replicates can be sufficient for the distributional approximation and ranking results above, while larger budgets are beneficial when precise estimation of the distributional spread is required. Combined with the preceding distributional and rank-stability results, this indicates that $\tilde{B} = 30$ provides a reliable low-budget approximation, while larger budgets improve precision when the distributional spread itself is of primary interest.

Overall, when the Gaussian diagnostics are satisfied, the Gaussian approximation can substantially reduce the number of replicates needed to characterize the attribution distribution and recover stable feature rankings when the diagnostics are satisfied.

4.4 Accuracy and Prediction Performance Evaluation

This section evaluates the accuracy of RoSHAP feature rankings and their downstream predictive utility. Ranking accuracy is assessed in simulation through recovery of the true signal set and agreement with the known importance ordering. Predictive utility is evaluated on held-out data using the top-ranked features. Comparisons include single-fit SHAP, PowerSHAP, Shapicant, BorutaShap, and the rank-verification procedure of Goldwasser and Hooker (2024). RoSHAP uses $B = 30$ replicates throughout this section and the refit-based baselines use their package defaults.

Table 3: Number of true signals ranked in the top s, mean and range (minimum–maximum) over thirty independent data sets per setting, $n = 600$ rows each.

	$(d, s) = (100, 10)$	$(d, s) = (100, 50)$	$(d, s) = (1000, 50)$
Single-fit SHAP	7.2 (5–9)	40.4 (36–43)	31.5 (28–37)
Powershap	8.6 (8–9)	45.2 (44–48)	40.6 (39–44)
Shapicant	8.7 (8–9)	44.8 (43–46)	39.0 (37–42)
BorutaShap	7.9 (7–9)	42.8 (41–45)	36.3 (33–41)
Goldwasser & Hooker	**8.9** (8–9)	44.3 (42–46)	38.7 (36–41)
RoSHAP	8.8 (8–10)	**45.7** (44–48)	**41.4** (39–44)

Accuracy. Table 3 reports the number of true signals ranked among the top s features. Methods based on repeated fitting consistently outperform single-fit SHAP, with the largest gap in the high-dimensional setting $(d, s) = (1000, 50)$. RoSHAP achieves the highest average recovery in the dense and high-dimensional settings and performs comparably to the strongest baselines in the reference

setting. Differences among the leading methods are generally small relative to the variation across independently generated datasets. These results show that the distributional summaries produced by RoSHAP retain strong signal-ranking performance.

Figure 7 evaluates agreement with the true importance ordering in the dense setting. The ten strongest signals differ only slightly in effect size, making their exact order difficult to recover from $n = 600$ observations. RoSHAP tracks the true ordering most closely, with an average median-rank deviation of 2.0 positions and only one of the ten signals deviating by more than three positions. PowerSHAP performs similarly at 2.1, while the remaining methods show larger deviations. Single-fit SHAP is also substantially more variable across datasets. Across all 50 signals, RoSHAP achieves a Spearman correlation of 0.86 with the true ordering, compared with 0.84 for PowerSHAP and 0.68 for single-fit SHAP.

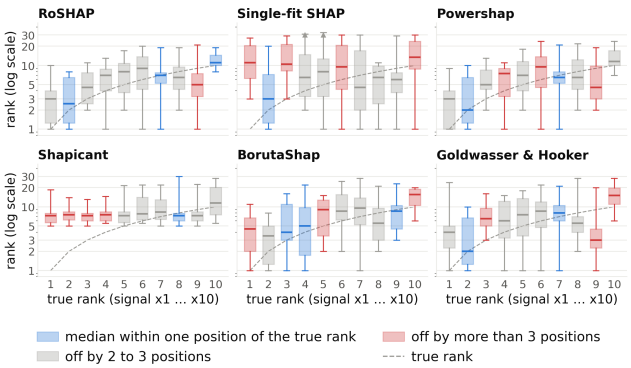


Figure 7: Estimated ranks of the ten strongest signals across thirty $(d, s) = (100, 50)$ simulations. Dashed lines mark true ranks; color indicates median rank deviation (blue: ≤ 1, grey: 2–3, red: > 3). The log-scaled axis is truncated at rank 30, with triangles indicating larger values.

Prediction. Table 4 compares downstream prediction performance using the top $k = 10$ features from each method. RoSHAP is competitive with the baselines on all three datasets: it ties for the best MADELON AP, achieves the best for UJIIndoorLoc errors, and attains the highest Musk (Version 2) F1 and AP. The remaining differences are modest. These results indicate that the distributional summaries retain predictive utility.

Table 4: Held-out performance using the top $k = 10$ features.

	MADELON, CatBoost		UJI, LightGBM		Musk (Version 2), Random forest	
Method	F1 ↑	AP ↑	RMSE ↓	MAE ↓	F1 ↑	AP ↑
SHAP	0.885	0.953	103.6	87.5	0.685	0.859
Native	0.869	0.951	136.6	123.0	0.688	0.863
LIME	0.825	0.914	130.8	116.4	0.689	0.863
PowerSHAP	0.881	**0.955**	**91.0**	73.9	0.685	0.859
BorutaShap	**0.886**	**0.955**	103.6	87.5	0.688	0.863
Shapicant	0.819	0.924	101.6	86.7	0.685	0.859
Goldwasser & Hooker	0.885	0.953	103.6	87.5	0.685	0.859
RoSHAP	0.881	**0.955**	**91.0**	**73.9**	**0.708**	**0.876**

5 Discussion

This study addresses a key limitation of existing feature attribution methods: attribution scores are often treated as fixed despite their stochastic nature. We propose RoSHAP, a robust distributional framework for estimating attribution distributions over repeated resampling and refitting. By characterizing both the magnitude and variability of feature attributions, the framework reveals when strong attributions are accompanied by substantial uncertainty, while also identifying near-ties and interpretations sensitive to the particular data split. The Gaussian approximation and accompanying diagnostics make this distributional view computationally practical by identifying when the attribution distribution can be summarized parametrically, allowing reliable estimates with substantially fewer resampling rounds. Although repeated model fitting remains more costly than single-run attribution, this reduction can substantially lower the computational burden when the approximation is appropriate. Future work could develop adaptive resampling strategies and evaluate the framework across broader data modalities and attribution methods. RoSHAP therefore complements existing feature attribution methods by adding uncertainty information to support more reliable interpretation.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with wording, grammar, and language refinement of text written by the authors. We did not use generative AI to generate research ideas, formulate mathematical claims, develop theoretical results, construct proofs, design the methodology or experiments, implement methods, analyze data, or interpret results. All AI-assisted edits were reviewed and revised by the authors to ensure accuracy and consistency with the intended scientific content. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the RoSHAP procedure, assumptions, ranking criteria, and experimental settings in the main text and appendices. Complete proofs of the theoretical results and empirical diagnostics for the stated assumptions are included in the appendix. The experimental section specifies the datasets, preprocessing procedures, learners, hyperparameters, resampling design, evaluation metrics, and replicate budgets used in the simulations and real-data studies. Additional results and diagnostic analyses are provided in the supplementary material. Code for reproducing the experiments will be released with the paper.

REFERENCES

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

Torgyn Shaikhina, Umang Bhatt, Roxanne Zhang, Konstantinos Georgatzis, Alice Xiang, and Adrian Weller. Effects of uncertainty on the quality of feature importance explanations. In *AAAI workshop on explainable agency in artificial intelligence*. AAAI Press Washington, DC, USA, 2021.

David Chapman and Ajay Jain. Musk (Version 2). UCI Machine Learning Repository, 1994.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Jeremy Goldwasser and Giles Hooker. Statistical significance of feature importance rankings. *arXiv preprint arXiv:2401.15800*, 2024.

David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic shapley values. *Advances in Neural Information Processing Systems*, 36:7330–7350, 2023.

Riccardo Scheda and Stefano Diciotti. Explanations of machine learning models in repeated nested cross-validation: an application in age prediction using brain complexity features. *Applied Sciences*, 12(13):6681, 2022.

Ramzi Halabi, Benoit H Mulsant, Mirkamal Tolend, Daniel M Blumberger, Alexandra DeShaw, Arend Hintze, Christina Gonzalez-Torres, Muhammad I Husain, Helena K Kim, Claire O'Donovan, et al. A systematic exploration of digital biomarkers for the detection of depressive episodes in bipolar disorder. *npj Mental Health Research*, 5(1):13, 2026.

Jarne Verhaeghe, Jeroen Van Der Donckt, Femke Ongenae, and Sofie Van Hoecke. Powershap: a power-full shapley feature selection method. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 71–87. Springer, 2022.

Eoghan Keany. Borutashap: A wrapper feature selection method which combines the boruta feature selection algorithm with shapley values. *Zenodo*, 2020.

Manuel Calzolari. shapicant: Feature selection based on shap and target permutation. https://github.com/manuel-calzolari/shapicant, 2022. Version 0.4.0.

Joaquín Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Tomàs Arnau, and Joan P. Avariento. UJIIndoorLoc. UCI Machine Learning Repository, 2014. Accessed from the UCI Machine Learning Repository.

Isabelle Guyon. Madelon. UCI Machine Learning Repository, 2004. DOI: https://doi.org/10.24432/C5602H.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Jaroslav Hájek. Limiting distributions in simple random sampling from a finite population. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének közlemenyei*, 5(3):361–374, 1960.

A Appendix

A.1 Proof of the Gaussian approximation

Throughout, $\mathcal{D}_n$ and the feature j are fixed: all probabilities, expectations and variances are conditional on $\mathcal{D}_n$, and the subscript j is dropped. Write $I_i = \mathbf{1}\{i \in \mathcal{O}\}$,

$$A = U - E[U \mid \mathcal{O}], \qquad v(\mathcal{O}) = \mathrm{Var}(U \mid \mathcal{O}), \qquad s^2 = \mathrm{Var}(\widetilde{U}),$$

$$\tau^2 = s^2 + E[v(\mathcal{O})], \qquad \gamma = \frac{E[v(\mathcal{O})]}{\tau^2},$$

so that

$$U - \mu = (\widetilde{U} - \mu) + (E[U \mid \mathcal{O}] - \widetilde{U}) + A, \tag{1}$$

where $E[A \mid \mathcal{O}] = 0$, so that A is uncorrelated with every square-integrable function of $\mathcal{O}$, and

$$\mathrm{Var}(A) = E[v(\mathcal{O})].$$

Convergence in distribution conditional on $\mathcal{O}$, as in (A4), means convergence in probability of the conditional characteristic function,

$$E[e^{iuX} \mid \mathcal{O}] \xrightarrow{p} e^{-u^2/2} \qquad \text{for every real } u;$$

the ratio in (A4) is set to zero when $v(\mathcal{O}) = 0$.

We prove the theorem under the following slightly more general version of (A1), where $H = 1$ recovers the original assumption. For classification, the groups are the classes.

(A1′) The observations are partitioned into a fixed number $H \geq 1$ of groups $\mathcal{N}_1, \ldots, \mathcal{N}_H$ of sizes $n_1, \ldots, n_H$. For each h, $\mathcal{O} \cap \mathcal{N}_h$ is a uniformly drawn subset of $\mathcal{N}_h$ of fixed size k_h, independently across groups, with $n_h \to \infty$ and $f_h = k_h/n_h \to \lambda_h \in (0, 1)$.

Let $h(i)$ be the group of observation i.

Lemma 1. *Let $D_i = E[U \mid I_i = 1] - E[U \mid I_i = 0]$. Under (A1′), for n large enough that $0 < k_h < n_h$ for every h:*

(i) We have

$$E[U \mid I_i] - \mu = (I_i - f_{h(i)})D_i;$$

hence

$$\mathrm{Var}(E[U \mid I_i]) = f_{h(i)}(1 - f_{h(i)})D_i^2, \qquad \mathrm{Cov}(U, I_i) = f_{h(i)}(1 - f_{h(i)})D_i.$$

(ii) $\sum_{i\in\mathcal{N}_h} D_i = 0$ *for every* h*. Hence*

$$\widetilde{U} - \mu = \sum_h T_h, \qquad T_h = \sum_{i\in\mathcal{O}\cap\mathcal{N}_h} D_i,$$

with the T_h *independent across* h*, and* $s^2 = \sum_h s_h^2$ *with*

$$s_h^2 = \mathrm{Var}(T_h) = \frac{n_h}{n_h - 1} f_h(1 - f_h) \sum_{i\in\mathcal{N}_h} D_i^2.$$

(iii) We have

$$\mathrm{Cov}\big(\widetilde{U},\, E[U \mid \mathcal{O}] - \widetilde{U}\big) = -\sum_h \frac{s_h^2}{n_h},$$

and therefore

$$\sigma^2 = \tau^2 - 2\sum_h \frac{s_h^2}{n_h} + \mathrm{Var}\big(E[U \mid \mathcal{O}] - \widetilde{U}\big). \tag{2}$$

Proof. (i) With $f = f_{h(i)} = P(I_i = 1)$,

$$E[U \mid I_i] = I_i\, E[U \mid I_i = 1] + (1 - I_i)\, E[U \mid I_i = 0] \qquad \text{and} \qquad \mu = E\big[E[U \mid I_i]\big];$$

subtracting gives $(I_i - f)D_i$, and $\mathrm{Var}(I_i) = f(1 - f)$. Also $\mathrm{Cov}(U, I_i) = \mathrm{Cov}(E[U \mid I_i], I_i)$.
(ii) $\sum_{i\in\mathcal{N}_h} I_i = k_h$ is constant, so by (i)

$$0 = \mathrm{Cov}\Big(U, \sum_{i\in\mathcal{N}_h} I_i\Big) = f_h(1 - f_h) \sum_{i\in\mathcal{N}_h} D_i.$$

Then

$$\widetilde{U} - \mu = \sum_i (I_i - f_{h(i)}) D_i = \sum_h T_h,$$

and T_h depends on $\mathcal{O} \cap \mathcal{N}_h$ only. T_h is the sum of a simple random sample of size k_h from the numbers $\{D_i : i \in \mathcal{N}_h\}$, whose mean is zero, so its variance is $k_h(1 - f_h)$ times their variance with divisor $n_h - 1$.
(iii) For $l \in \mathcal{N}_h$,

$$\mathrm{Cov}(I_i, I_l) = \begin{cases} f_h(1 - f_h), & i = l, \\ -f_h(1 - f_h)/(n_h - 1), & i \in \mathcal{N}_h \setminus \{l\}, \\ 0, & \text{otherwise.} \end{cases}$$

With $\sum_{i\in\mathcal{N}_h\setminus\{l\}} D_i = -D_l$, this gives

$$\mathrm{Cov}(\widetilde{U}, I_l) = \frac{n_h}{n_h - 1} f_h(1 - f_h) D_l = \frac{n_h}{n_h - 1} \mathrm{Cov}(U, I_l),$$

so

$$\mathrm{Cov}(U - \widetilde{U}, I_l) = -\frac{f_h(1 - f_h) D_l}{n_h - 1}$$

and

$$\mathrm{Cov}(U - \widetilde{U}, \widetilde{U}) = \sum_l D_l\, \mathrm{Cov}(U - \widetilde{U}, I_l) = -\sum_h \frac{f_h(1 - f_h)}{n_h - 1} \sum_{l\in\mathcal{N}_h} D_l^2 = -\sum_h \frac{s_h^2}{n_h}.$$

As A is uncorrelated with $\widetilde{U}$, the left side equals $\mathrm{Cov}(E[U \mid \mathcal{O}] - \widetilde{U}, \widetilde{U})$, and the variance of equation 1 gives equation 2.

□

Proof of the theorem. We first compare σ^2 with τ^2. In equation 2,

$$0 \le \sum_h \frac{s_h^2}{n_h} \le \frac{\tau^2}{\min_h n_h} \qquad \text{and} \qquad \mathrm{Var}\big(E[U \mid \mathcal{O}] - \widetilde{U}\big) = o(\sigma^2)$$

by (A2), so $\sigma^2 = \tau^2\{1 + o(1)\} + o(\sigma^2)$ and

$$\frac{\sigma^2}{\tau^2} \to 1. \tag{3}$$

It therefore suffices to show $(U - \mu)/\tau \xrightarrow{d} \mathcal{N}(0, 1)$, and $o(\sigma^2)$ may be read as $o(\tau^2)$. Put

$$W_h = \frac{T_h}{\tau}, \quad W = \sum_h W_h = \frac{\widetilde{U} - \mu}{\tau}, \quad R = \frac{E[U \mid \mathcal{O}] - \widetilde{U}}{\tau}, \quad c_h = \frac{s_h^2}{\tau^2}, \quad c = \sum_h c_h = 1 - \gamma;$$

then, by equation 1,

$$\frac{U - \mu}{\tau} = W + R + \frac{A}{\tau}.$$

We next show that

$$E\, e^{itW} - e^{-ct^2/2} \to 0 \qquad \text{for every real } t. \tag{4}$$

Since H is fixed and $c_h \in [0, 1]$, consider an arbitrary subsequence and extract a further subsequence along which every c_h converges, say $c_h \to c_{h,0}$. Along this further subsequence, since the W_h are independent, it suffices to show $E\, e^{itW_h} \to e^{-c_{h,0}t^2/2}$ for each h; then

$$E\, e^{itW} = \prod_h E\, e^{itW_h} \to e^{-\sum_h c_{h,0}t^2/2}, \qquad e^{-ct^2/2} \to e^{-\sum_h c_{h,0}t^2/2}.$$

If $c_{h,0} = 0$, this follows from $E\, W_h^2 = c_h \to 0$. If $c_{h,0} > 0$, then by Lemma 1(i), (A3) and $s^2 \le \tau^2$,

$$\frac{\max_{i \in \mathcal{N}_h} D_i^2}{s_h^2} = \frac{\max_{i \in \mathcal{N}_h} \mathrm{Var}(E[U \mid I_i])}{f_h(1 - f_h)\, c_h\, \tau^2} = \frac{o(\tau^2)}{f_h(1 - f_h)\, c_h\, \tau^2} \to 0,$$

so for every $\varepsilon > 0$ the set

$$\{i \in \mathcal{N}_h : D_i^2 > \varepsilon^2 s_h^2\}$$

is eventually empty. This maximum-term condition implies the finite-population Lindeberg condition of Hájek (1960) for the sum T_h of a simple random sample of size k_h from $\{D_i : i \in \mathcal{N}_h\}$. Since $k_h \to \infty$ and $n_h - k_h \to \infty$ by (A1′), Hájek's theorem gives

$$\frac{T_h}{s_h} \xrightarrow{d} \mathcal{N}(0, 1), \qquad \text{that is} \qquad W_h \xrightarrow{d} \mathcal{N}(0, c_{h,0}).$$

Since the original subsequence was arbitrary, equation 4 holds for the full sequence.

Moreover,

$$E\, R^2 = \frac{\mathrm{Var}\big(E[U \mid \mathcal{O}] - \widetilde{U}\big)}{\tau^2} \to 0 \tag{5}$$

by (A2) and equation 3, so $R \xrightarrow{p} 0$.

It remains to handle the learner's randomness. Let $\psi(t) = E[e^{itA/\tau} \mid \mathcal{O}]$. We show

$$\psi(t) - e^{-\gamma t^2/2} \xrightarrow{p} 0 \qquad \text{for every real } t. \tag{6}$$

If $E[v(\mathcal{O})] = o(\sigma^2)$, then $\gamma \to 0$, $e^{-\gamma t^2/2} \to 1$ and

$$E|\psi(t) - 1| \le |t|\, \frac{E|A|}{\tau} \le |t|\sqrt{\gamma} \to 0.$$

Under the second case of (A4), let

$$Z = \frac{A}{v(\mathcal{O})^{1/2}} \quad (Z = 0 \text{ if } v(\mathcal{O}) = 0), \qquad \rho = \frac{v(\mathcal{O})}{\tau^2},$$

so that $A/\tau = \sqrt{\rho}\, Z$, $E[Z^2 \mid \mathcal{O}] \le 1$, and $\varphi(u) = E[e^{iuZ} \mid \mathcal{O}]$ satisfies

$$|\varphi(u) - \varphi(u')| \le |u - u'|\, E[|Z| \mid \mathcal{O}] \le |u - u'|.$$

By (A4), $\varphi(u) \xrightarrow{p} e^{-u^2/2}$ for every fixed u, and

$$|\sqrt{\rho} - \sqrt{\gamma}| \le |\rho - \gamma|^{1/2} = \gamma^{1/2} \left| \frac{v(\mathcal{O})}{E[v(\mathcal{O})]} - 1 \right|^{1/2} \xrightarrow{p} 0.$$

Hence

$$|\psi(t) - e^{-\gamma t^2/2}| = |\varphi(t\sqrt{\rho}) - e^{-\gamma t^2/2}| \le |t|\, |\sqrt{\rho} - \sqrt{\gamma}| + \sup_{|u| \le |t|} |\varphi(u) - e^{-u^2/2}|,$$

and the supremum tends to zero in probability: for $\varepsilon > 0$ take finitely many points $u_1, \dots, u_m$ at spacing at most ε in $[-|t|, |t|]$; both φ and $u \mapsto e^{-u^2/2}$ are 1-Lipschitz, so

$$\sup_{|u| \le |t|} |\varphi(u) - e^{-u^2/2}| \le 2\varepsilon + \max_{l \le m} |\varphi(u_l) - e^{-u_l^2/2}|,$$

and the maximum over finitely many points tends to zero in probability. This proves equation 6.

Finally, since $|\psi(t) - e^{-\gamma t^2/2}| \le 2$, convergence in probability and boundedness imply $E|\psi(t) - e^{-\gamma t^2/2}| \to 0$. As W is a function of $\mathcal{O}$,

$$E\, e^{it(W + A/\tau)} = E[e^{itW} \psi(t)] = e^{-\gamma t^2/2}\, E\, e^{itW} + o(1) = e^{-(\gamma + c)t^2/2} + o(1) = e^{-t^2/2} + o(1)$$

by equation 4. By Lévy's continuity theorem $W + A/\tau \xrightarrow{d} \mathcal{N}(0, 1)$; by equation 5 and Slutsky's lemma $(U - \mu)/\tau \xrightarrow{d} \mathcal{N}(0, 1)$, and by equation 3

$$\frac{U - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

□

A.2 Real datasets, preprocessing, and splits.

UJIIndoorLoc (Torres-Sospedra et al., 2014) contains $n = 19{,}937$ training fingerprints with received signal strengths from $d = 520$ WiFi access points. We predict longitude and recode undetected access points from $+100$ to -105,dBm to retain the RSSI scale. We use the official split with $1{,}111$ held-out fingerprints collected four months later.

MADELON (Guyon, 2004) is a balanced nonlinear classification benchmark with $n = 2{,}600$ observations and $d = 500$ features. Five features define the signal, fifteen are linear combinations of them, and the remaining 480 are random probes.

Musk (Version 2) (Chapman and Jain, 1994) is an imbalanced binary classification dataset with $n = 6{,}598$ molecular conformations and $d = 166$ continuous shape features. Of these observations, 1,017 (15%) belong to the musk class.

A.3 Implementation details

Models and hyperparameters. Table 5 lists every configuration. The simulations use XGBoost with 150 trees in all settings, so that differences between settings are attributable to the data alone. On the real data sets each learner uses one configuration per data set; the penalised linear model is ℓ_2-penalised logistic regression for the classification tasks and ridge regression for UJIIndoorLoc. On Musk (Version 2) the features are standardised once on the full data set before the test split, and every learner weights the positive class by the class ratio of about 5.5. Attributions are TreeSHAP values from each library's own implementation (`shap.TreeExplainer` for random forests) and `shap.LinearExplainer` for the linear models, on the log-odds scale for classifiers. In every round the learner's seed is held fixed and the columns are randomly permuted, which redraws the learner's internal random choices; the evaluation set is half of the training pool. XGBoost on UJIIndoorLoc was fitted on a GPU; every other fit ran on CPUs. Software: XGBoost 3.2.0, LightGBM 4.6.0, CatBoost 1.2.10, scikit-learn 1.7.2, shap 0.49.1.

Table 5: Learner configurations. T: trees; η: learning rate; d: maximum depth; L: leaves; m: minimum rows per leaf; h: minimum child weight; r, c: row and column subsampling fractions per tree; α, λ: ℓ_1, ℓ_2 penalties; w: weight of the positive class relative to the negative class (inverse class frequency). Values not listed are the library defaults.

Data set	LightGBM	XGBoost	Random forest	CatBoost	Linear
Simulations	—	T=150, η=0.5, d=20, h=0.2, r=0.8, c=0.6	—	—	—
MADELON	T=300, η=0.1, d=8, L=63, m=10, r=0.8, c=0.5	T=300, η=0.1, d=8, r=0.8, c=0.5	T=300, m=2	T=300, η=0.1, d=8	logistic, C=1
UJIIndoorLoc	T=300, η=0.03, d=7, L=63, m=30, r=0.8, c=0.7, α=0.01, λ=1	T=300, η=0.03, d=6, h=10, r=0.8, c=0.7, α=0.01, λ=2	T=100, d=10, m=5, $0.3p$ features per split	T=300, η=0.03, d=6, λ=5, c=0.7	ridge, α=10
Musk (Version 2)	T=150, η=0.1, d=8, L=31, m=15, r=0.8, c=0.8, w=5.5	T=400, η=0.07, d=5, r=0.8, c=0.5, γ=0.3, λ=1.5, w=5.5	T=200, d=8, m=3, w=5.5	T=200, η=0.1, d=6, w=5.5	logistic, C=1, w=5.5

Diagnostics procedure. Notation is that of Appendix A.1. Let $\eta = s^2/\sigma^2$ be the first-order share of the replicate variance and $\bar{\gamma} = E[v(\mathcal{O})]/\sigma^2$ the share of the learner's internal randomness. By Lemma 1(iii), $\mathrm{Var}(E[U \mid \mathcal{O}] - \widetilde{U})/\sigma^2 = 1 - \eta - \bar{\gamma} + r$ with $0 \le r \le 2\eta/\min_h n_h$, so (A2) asks that $1 - \eta - \bar{\gamma}$ be close to zero. The diagnostics use the B replicates $U^{(b)}$ of Algorithm 1 with their evaluation sets $\mathcal{O}^{(b)}$, regenerated from the round seeds, and, for a randomised learner, *seed-only refits*: at the evaluation set of a stored round, the learner is refitted m times with only its seed changed, at two rounds.

First-order share. For rounds $b \neq b'$ whose evaluation sets share $m_{bb'}$ rows, $E[(U^{(b)} - \mu)(U^{(b')} - \mu) \mid m_{bb'}] = S^2(m_{bb'} - k^2/n)$ to first order, where $S^2 = \sum_i D_i^2/(n-1)$ and $s^2 = k(1-f)S^2$. The least-squares slope of the cross-products $(U^{(b)} - \bar{U})(U^{(b')} - \bar{U})$ on $m_{bb'} - k^2/n$ over all $B(B-1)/2$ pairs estimates S^2 (one slope per class under a stratified split), and $\hat{\eta} = k(1-f)\hat{S}^2/\hat{\sigma}^2$, with $\hat{\sigma}^2$ the sample variance of the replicates. The learner's randomness is independent across rounds and does not enter the slope. The standard error is the delete-one-round jackknife, and a permutation test (the $U^{(b)}$ shuffled against the $\mathcal{O}^{(b)}$) tests $s^2 > 0$.

Seed-only share. $\hat{\bar{\gamma}}$ is the sample variance of the seed-only draws divided by $\hat{\sigma}^2$, averaged over the two rounds; $\bar{\gamma} = 0$ for a learner that is deterministic given its training set.

(A2) ratio. $1 - \hat{\eta} - \hat{\bar{\gamma}}$, with standard error from those of its two parts; (A2) is judged met when its 95% interval contains zero.

(A3) ratio. D_i is estimated by $\hat{D}_i$, the mean of the replicates whose evaluation set contains i minus the mean of those whose evaluation set does not, within the class of i; the ratio is $\max_i f_{h(i)}(1 - f_{h(i)})\hat{D}_i^2/\hat{s}^2$, equal to $1/n$ under equal contributions. Its upper 95% limit replaces $\max_i |\hat{D}_i|$ by $\max_i |\hat{D}_i| + 1.64\,\hat{\sigma}\,(1/B_i + 1/(B - B_i))^{1/2}$, B_i the number of rounds with $i \in \mathcal{O}^{(b)}$. The ratio is not estimable when s^2 is not above its permutation null; (A3) is judged met when the upper limit is below 0.10.

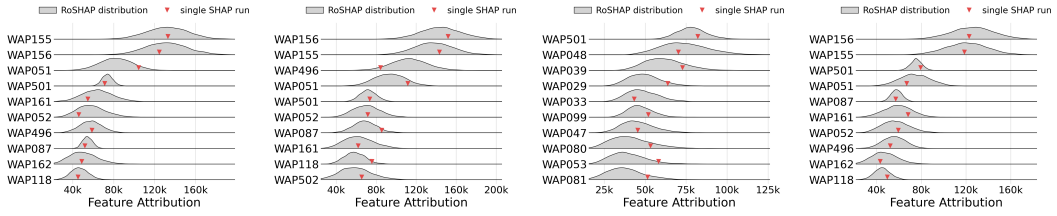
(A4). The Shapiro–Wilk test of the seed-only draws at each round assesses the conditional normality of U given $\mathcal{O}$, and the Brown–Forsythe test of equal variance between the two rounds the concentration of $v(\mathcal{O})$; (A4) is judged met when both have $p > 0.05$. It holds by construction for a deterministic learner. The Shapiro–Wilk p-value of the B replicates is reported alongside as the conclusion of the theorem, not as a condition.

A.4 BASELINE CONFIGURATIONS

Single-fit SHAP fits one model to all training rows and explains the same rows. Native importance is total gain for the boosters, mean decrease in impurity for the random forest, and absolute coefficients for the linear models. The rank-verification procedure of Goldwasser and Hooker (2024) is applied to a single XGBoost fit on all n rows, since it verifies the ranking of a fixed model rather than producing one across refits. LIME explains 50 randomly chosen training rows of the single fit, with up to 50 features per explanation and the library defaults otherwise, and ranks features by mean absolute local weight. PowerShap, Shapicant, and BorutaShap are implemented according to their published definitions using the same learner and SHAP estimator as the other methods. PowerShap ranks features by the p-value obtained against an injected uniform random feature, while Shapicant uses permutation p-values from models fit to shuffled targets, with ties broken by mean $|\mathrm{SHAP}|$. BorutaShap augments the data with a permuted shadow copy of each feature and ranks features by the number of iterations in which their importance exceeds the best shadow feature, with ties again broken by mean $|\mathrm{SHAP}|$. All three methods are run using their published packages at default settings: powershap 0.1.0.1, shapicant 0.4.0, and BorutaShap 1.0.17. The simulation uses the same XGBoost learner throughout.

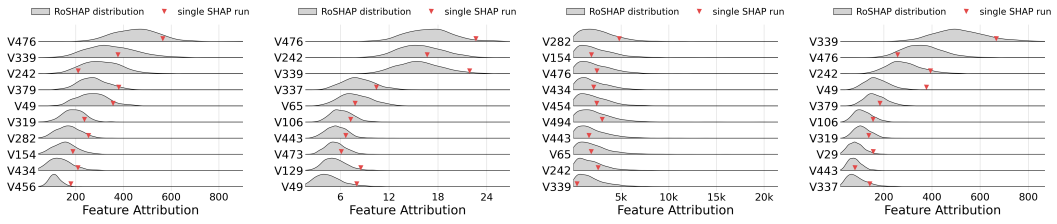
A.5 ADDITIONAL RESULTS

Attribution distribution across data–learner pairs. Additional RoSHAP attribution distributions across learners are shown in Figures 8, 9, and 10 for UJIIndoorLoc, MADELON, and Musk (Version 2), respectively. These results show that the unstable single-run SHAP patterns are consistent across learners and are not limited to the representative models shown in the main text.



(a) LightGBM (b) Random forest (c) Ridge regression (d) XGBoost

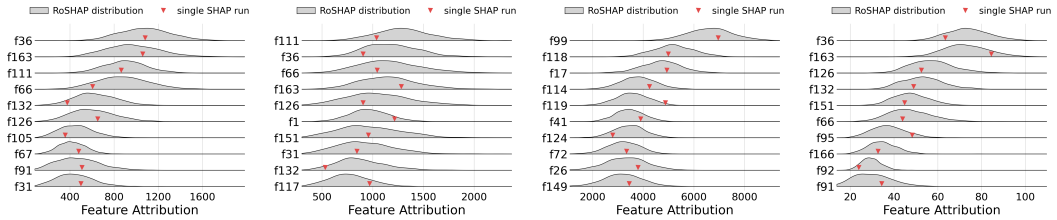
Figure 8: RoSHAP attribution distributions for UJIIndoorLoc across four learners.



(a) CatBoost (b) Random forest (c) Logistic regression (d) XGBoost

Figure 9: RoSHAP attribution distributions for MADELON across four learners.

Diagnostics results. Table 6 provides an example of the diagnostics for conditions (A2)–(A4), using the ten highest-ranked UJIIndoorLoc features under CatBoost. The results illustrate how the proposed diagnostics assess first-order adequacy, dominance by individual observations, and the Gaussianity and variance stability of learner randomness. The (A3) ratios are small for the leading features, indicating that no single observation dominates the first-order contribution. The seed-only Shapiro–Wilk and Brown–Forsythe tests provide complementary checks of Gaussianity and variance stability. Although some features show evidence of departures from these conditions, the corresponding QQ plots suggest that deviations from Gaussianity are generally modest.



(a) CatBoost (b) LightGBM (c) Logistic regression (d) Random forest

Figure 10: RoSHAP attribution distributions for Musk (Version 2) across four learners.

Table 6: Diagnostics of conditions (A2)–(A4) for the ten features of smallest expected rank, in that order; UJIIndoorLoc, CatBoost, $n = 19{,}937$, $B = 3{,}500$ rounds, seed-only refits at two rounds ($m = 100$ and 60). $\hat{\eta}$: first-order share (jackknife s.e.); $\hat{\tilde{\gamma}}$: seed-only share, mean of the two rounds; (A2): $1 - \hat{\eta} - \hat{\tilde{\gamma}}$ (s.e.) [upper 95% limit]; (A3): ratio [upper 95% limit], n.e. when the first-order share is not above its permutation null; (A4) SW: the smaller Shapiro–Wilk p of the seed-only draws over the two rounds; (A4) BF: Brown–Forsythe p for equal seed-only variance at the two rounds; SW: Shapiro–Wilk p of the B replicates.

Feature	$\hat{\eta}$ (s.e.)	$\hat{\tilde{\gamma}}$	(A2) (s.e.) [upper]	(A3) [upper]	(A4) SW	(A4) BF	SW
WAP155	0.28 (0.05)	0.83	-0.11 (0.12) [0.13]	0.034 [0.049]	0.331	0.387	0.089
WAP156	0.33 (0.05)	0.72	-0.05 (0.10) [0.14]	0.024 [0.036]	0.645	0.767	0.899
WAP501	0.18 (0.05)	1.06	-0.24 (0.19) [0.14]	0.016 [0.030]	0.724	0.376	0.616
WAP087	0.15 (0.05)	0.76	+0.09 (0.10) [0.28]	0.025 [0.045]	0.073	0.813	0.080
WAP051	0.12 (0.04)	0.89	-0.01 (0.15) [0.28]	0.033 [0.058]	0.022	0.221	0.101
WAP052	0.15 (0.04)	1.04	-0.20 (0.33) [0.44]	0.047 [0.072]	0.045	< 0.001	0.100
WAP496	0.19 (0.04)	0.70	+0.11 (0.09) [0.29]	0.026 [0.043]	0.533	0.514	0.013
WAP118	0.06 (0.04)	1.06	-0.12 (0.23) [0.35]	0.075 [0.130]	0.362	0.105	0.010
WAP161	0.16 (0.04)	0.66	+0.18 (0.09) [0.36]	0.024 [0.043]	0.233	0.508	0.010
WAP128	0.11 (0.04)	0.89	-0.00 (0.23) [0.45]	0.051 [0.082]	0.334	0.087	0.011

Figures 11–13 provide corresponding QQ plots for the five highest-ranked features of MADELON with LightGBM, UJIIndoorLoc with CatBoost, and Musk (Version 2) with XGBoost. These plots visually assess the Gaussian approximation used in the main analysis and complement the numerical diagnostics.

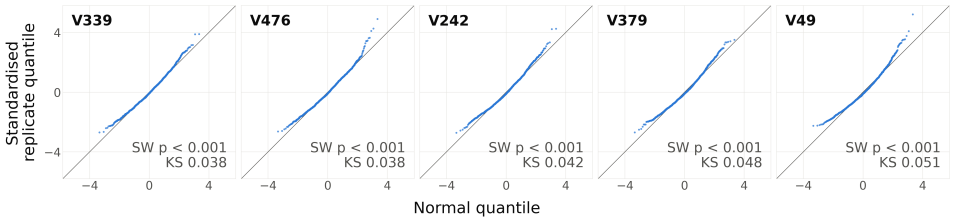


Figure 11: Gaussian QQ plots for the five highest-ranked features of MADELON under LightGBM.

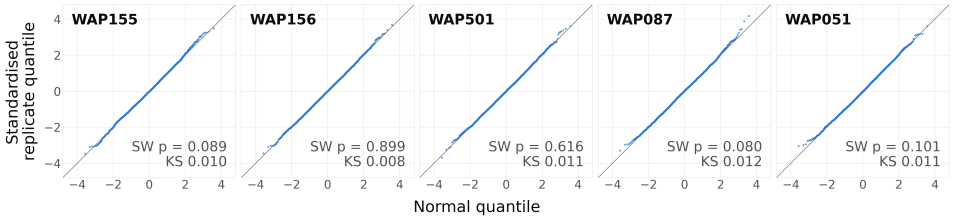


Figure 12: Gaussian QQ plots for the five highest-ranked features of UJIIndoorLoc under CatBoost.

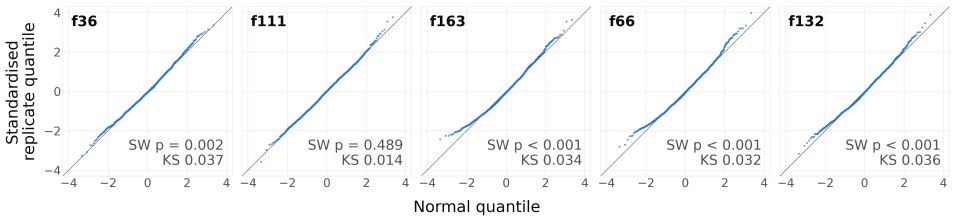


Figure 13: Gaussian QQ plots for the five highest-ranked features of Musk (Version 2) under XGBoost.

Accuracy evaluation other two simulation settings. Figures 14 and 15 show the ranking comparisons for the reference and high-dimensional simulation settings, respectively. These results complement the reference-setting analysis in the main text and show how ranking accuracy changes as the number of signal features and the ambient dimension increase.

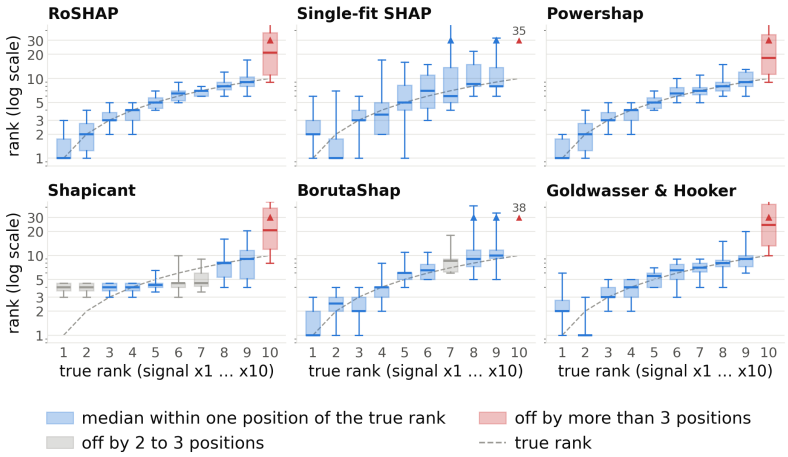


Figure 14: Feature-rank comparison across methods in the reference simulation setting, $(d, s) = (100, 10)$. Box plots summarize results across thirty independently generated datasets.

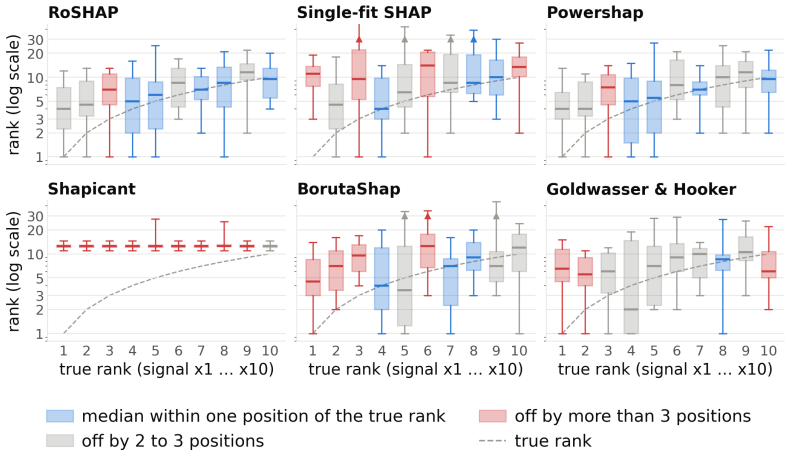


Figure 15: Feature-rank comparison across methods in the high-dimensional simulation setting, $(d, s) = (1000, 50)$. Box plots summarize results across thirty independently generated datasets.