école normale supérieure paris-saclay

Feature importance measures for Random Forests: the problem of Mean Decrease Impurity, solutions and alternatives

Gaetan De Castellane

September 4, 2025

Outline

Feature Importance

- Explain the impact of a feature on a model
- To understand the output
- Important for biology, finance, patient care
- Subjective notion

Two key considerations

1. **Marginal vs Conditional**: Do we want unique information?
2. **Model vs Data**: Are we explaining the model or the underlying process?

Decision Tree

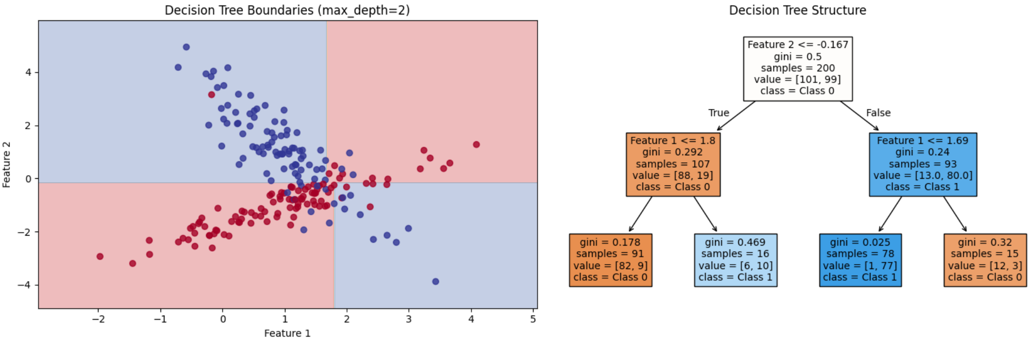


Figure: Visualization of a simple decision tree and its decision function.

école normale supérieure paris–saclay

Mean Decrease in Impurity

Definition (MDI, Breiman (2001))

For feature j:

$$\mathrm{MDI}(j) = \frac{1}{T} \sum_{t \in F} \sum_{\substack{m \in \mathrm{inter}(t) \\ j_m = j}} \left[\omega_m H(m) - \omega_{l_m} H(l_m) - \omega_{r_m} H(r_m)\right]$$

where $\omega_m = \frac{n_m}{n}$ and H is the impurity function.

The Problem with MDI

Three main issues:

1. **Positive bias**: Assigns non-zero importance to irrelevant features
2. **Cardinality bias**: Favors high-cardinality features
3. **Overfitting amplification**: Deeper trees = more bias

Despite that it is widely used, which causes a problem for `scikit-learn`.

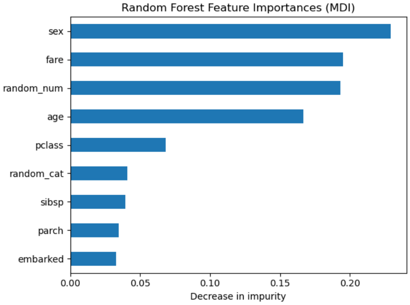


Figure: MDI assigns significant importance to random features

Existing Solutions

1. Conditional Inference Trees (Strobl et al. 2008)

- Replace CART (Breiman et al. 1984) with conditional inference trees (Hothorn, Hornik, and Zeileis 2006)
- Eliminates selection bias
- **Cost**: 25-35x slower training

2. Out-of-Bag Corrections

- UFI (Zhou and Hooker 2021)
- MDI-oob (Li et al. 2019)
- Use oob samples to reduce overfitting bias
- Presented in different ways, we show they are very close

Permutation Importance

Algorithm Permutation Importance

Require: Fitted model f, validation dataset $\mathcal{D}$, scoring function Score

1: Compute reference score $s_0 \leftarrow \text{Score}\big(f, \mathcal{D}\big)$
2: **for** each feature j **do**
3: $\quad \tilde{\mathcal{D}}^{(j)} \leftarrow \text{RandomlyShuffle}(\mathcal{D}, \text{column } j)$
4: $\quad s_j \leftarrow \text{Score}\big(f, \tilde{\mathcal{D}}^{(j)}\big)$
5: **end for**
6: $PI\big(j\big) \leftarrow s_0 - s_j$

- **Pros**: Model-agnostic, suitable for feature selection (Reyero-Lobo, Neuvial, and Thirion 2025)
- **Cons**: Computationally expensive, issues with correlated features

SAGE Values

Definition (Shapley Additive Global ImportancE)

$$\mathrm{SAGE}(j) = \frac{1}{p} \sum_{\substack{S \subseteq \{1,\ldots,p\} \\ j \notin S}} \binom{p-1}{|S|}^{-1} \left(v(S \cup \{j\}) - v(S)\right)$$

Satisfies four axioms: Efficiency, Symmetry, Dummy, Linearity

- **Pros**: Additive decomposition, game-theoretic foundation
- **Cons**: Exponential complexity, poor for feature selection (Reyero-Lobo, Neuvial, and Thirion 2025)
- **Note**: Converges to MDI in categorical settings (Sutera et al. 2021)

Rewriting MDI as Loss Decomposition

Key insight: MDI can be written as feature contributions to training loss improvement.

Saabas (2017) show that for any prediction $f_t(x)$:

$$f_t(x) = v_0 + \sum_{j=1}^{p} f_{t,j}(x)$$

This leads to:

$$\mathrm{MDI}(j) = \text{contribution of feature } j \text{ to training score improvement}$$

Definition (Training Score)

$$S_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}\left[l(y_i, v_0) - l(y_i, f_t(x_i))\right] = \sum_{j=1}^{p} S_{\text{train},j} = \sum_{j=1}^{p} \mathrm{MDI}(j)$$

Proposed Method: oob-score

Instead of using training samples, use out-of-bag samples:

$$S_{\text{oob}} = \frac{1}{n'} \sum_{i=1}^{n'} \left[I(y_i', v_0) - I(y_i', f_t(x_i')) \right] = \sum_{j=1}^{p} S_{\text{oob},j}$$

Definition (oob-score)

$$\text{oob-score}(j) = S_{\text{oob},j} = \sum_{\substack{m \in \text{inter}(t) \\ j_m = j}} \omega_m' H'(m) - \omega_{l_m}' H'(l_m) - \omega_{r_m}' H'(r_m)$$

where $H'(m)$ is the cross-impurity: OOB targets with in-bag node values

Advantage: Additive decomposition of risk reduction for single trees: S_{oob} approximates the risk improvement $S := \mathbb{E}_{x,y \sim P}[I(y, v_0) - I(y, f_t(x))]$.

Unifying UFI and MDI-oob

Summary of the Impurity measures

- $H(m) = \frac{1}{n_m} \sum_{\substack{i \in \{1,\dots,n\} \\ x_i \in R_m}} I(y_i, v_m)$
- $H'(m) = \frac{1}{n'_m} \sum_{\substack{i \in \{1,\dots,n'\} \\ x'_i \in R_m}} I(y'_i, v_m)$
- $H''(m) = \frac{1}{n'_m} \sum_{\substack{i \in \{1,\dots,n'\} \\ x'_i \in R_m}} I(y'_i, v'_m)$

Method	Impurity function	Weights
MDI	H	in-bag
oob-score	H'	out-of-bag
naive-oob	H''	out-of-bag
UFI	$\frac{H+H'}{2}$	in-bag
MDI-oob	$\frac{H+H'}{2}$	out-of-bag

Table: Summary of impurity-based methods

Key elements:

- UFI and MDI-oob are nearly identical (different weights)
- All methods converge asymptotically
- UFI has theoretical guarantee: $X_j \perp\!\!\!\perp Y$ in every hyperrectangle $\Rightarrow \mathbb{E}[\text{UFI}(j)] = 0$

Experimental Setup

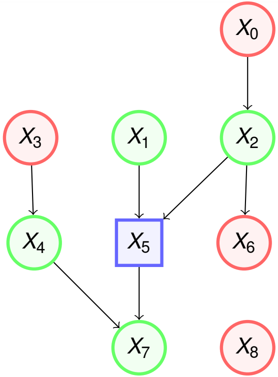


Figure: Feature relationships. Blue = Target, Green = Feature in Markov blanket, Red = Feature not in Markov blanket

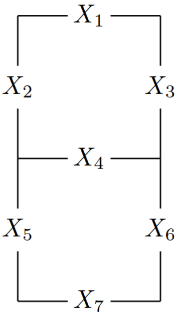


Figure: 7-segment display

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	1	1	1	0	1	1	1
1	0	0	1	0	0	1	0
2	1	0	1	1	1	0	1
3	1	0	1	1	0	1	1
4	0	1	1	1	0	1	0
5	1	1	0	1	0	1	1
6	1	1	0	1	1	1	1
7	1	0	1	0	0	1	0
8	1	1	1	1	1	1	1
9	1	1	1	1	0	1	1

Figure: Possible values of $(X_1, \ldots, X_7, Y)$

Noise Detection Results

$$\mathcal{H}_0 : X_8 \text{ has zero importance}$$

vs.

$$\mathcal{H}_1 : X_8 \text{ has non-zero importance.}$$

Method	Mean importance	Rejects H_0 (t-test)
MDI	0.0480	YES
naive-oob	0.0435	YES
UFI	0.0007	NO
MDI-OOB	-0.0048	YES
oob-score	-0.0475	YES
Permutation	0.0003	NO
SAGE	-0.0018	YES

Only UFI and Permutation Importance correctly identify irrelevant features.

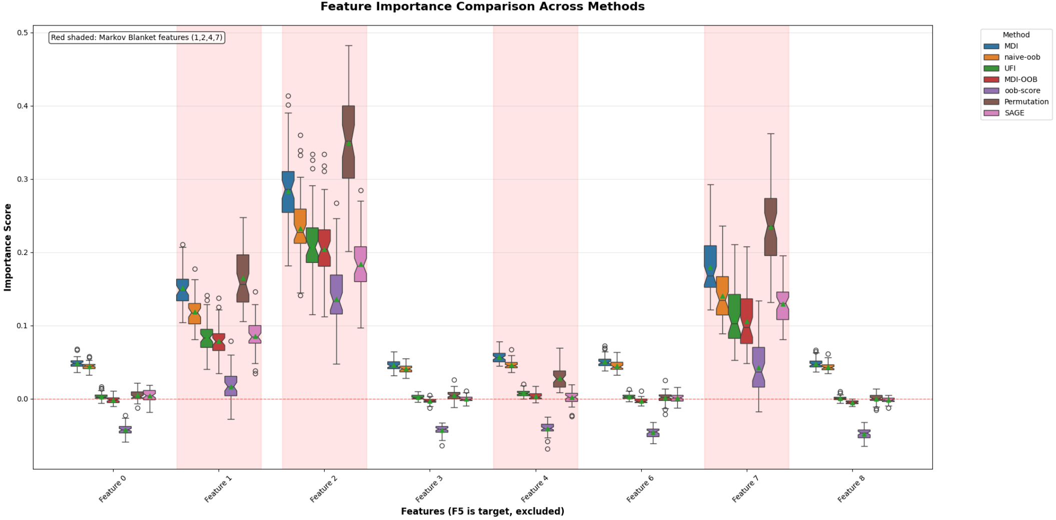
Feature Selection Performance

Task: Rank top 4 features to match Markov blanket

Method	Success rate
MDI	30/50 (60%)
naive-oob	14/50 (28%)
oob-score	15/50 (30%)
UFI	31/50 (62%)
MDI-OOB	33/50 (66%)
Permutation	46/50 (92%)
SAGE	13/50 (26%)

Permutation importance dominates, UFI and MDI-oob improve over MDI

Visualization



Computational Cost

Method	Time (500 pts)	Time (1000 pts)
MDI (retrieval)	16.8 ms	17.4 ms
UFI (high-level)	5506.7 ms	13606.6 ms
UFI (optimized)	192.6 ms	355.1 ms
Permutation	872.8 ms	1294.8 ms
SAGE	2835.7 ms	7028.1 ms

Key takeaway: Optimized UFI is 4x faster than Permutation, 14-20x faster than SAGE

Asymptotic Convergence of Impurity methods

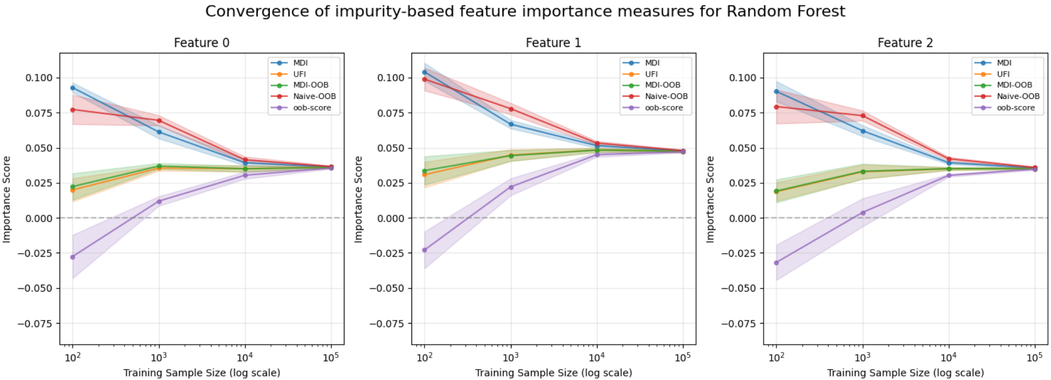


Figure: Evolution of the impurity based feature importance measures on the `noised_led` dataset as sample size increases, for the first 3 features.

Asymptotic Convergence of MDI to SAGE

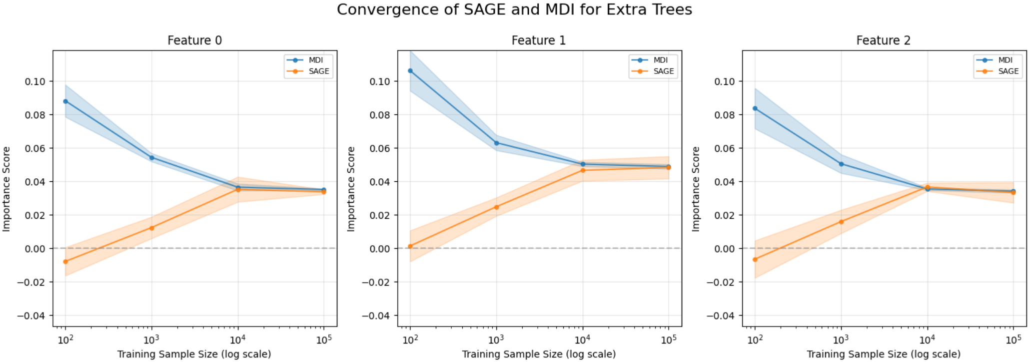


Figure: Convergence of the feature importance of SAGE and MDI in the categorical setting for Totally randomized trees, on the `noised_led` dataset, for the first 3 features.

Main Contributions

1. **Unified framework** for all impurity-based methods
2. **New method (oob-score)** with additive decomposition property
3. **Extended UFI/MDI-oob** to arbitrary loss functions
4. **Evaluation** of feature selection capability
5. **Fast implementation** of UFI in Cython

Conclusion

For scikit-learn's replacement of MDI

UFI is the best choice:

- Fast computation during training with Cython implementation
- Theoretical guarantee for noise detection
- Significant improvement over MDI

For feature selection tasks

Permutation Importance:

- Best performance for feature selection
- Already available in scikit-learn
- Worth the computational cost for critical applications

Future Work

- Prove or disprove $\text{UFI}(j) = 0 \implies X_j \perp\!\!\!\perp Y | X_{-j}$
- Formal proof that MDI is strictly positive in finite samples
- Adapt UFI to Gradient Boosting

Thank you for your attention.

Bibliography I

Breiman, Leo et al. (1984). *Classification and regression trees*. Wadsworth.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). “Unbiased recursive partitioning: A conditional inference framework”. In: *Journal of Computational and Graphical statistics* 15.3, pp. 651–674.

Strobl, Carolin et al. (2008). “Conditional variable importance for random forests”. In: *BMC bioinformatics* 9.1, p. 307.

Saabas, Ando (2017). “Interpreting random forests. 2014”. In: *URL: https://blog.datadive.net/interpreting-random-forests/*.

Li, Xiao et al. (2019). “A debiased MDI feature importance measure for random forests”. In: *Advances in Neural Information Processing Systems* 32.

Sutera, Antonio et al. (2021). “From global to local MDI variable importances for random forests and when they are Shapley values”. In: *Advances in Neural Information Processing Systems* 34, pp. 3533–3543.

Zhou, Zhengze and Giles Hooker (2021). “Unbiased measurement of feature importance in tree-based methods”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.2, pp. 1–21.

Bibliography II

Reyero-Lobo, Angel, Pierre Neuvial, and Bertrand Thirion (2025). “A principled approach for comparing Variable Importance”. In: *arXiv preprint arXiv:2507.17306*.