

Feature Importance for Applied Research

From model inspection to statistical inference

Lukas Burk

Leibniz Institute for Prevention Research and Epidemiology — BIPS
Bremen, Germany

26 October 2026

DKFZ, Heidelberg

Why feature importance?

- Feature *selection*: which candidate predictors carry most signal?
- *Sanity* check: does a model rely on plausible signal?
- *Hypothesis* generation
- *Inference* on the data-generating process
- **Caveat**: FI $\neq$ causal effect, but associations are still informative (Ewald et al. 2024)

! “Importance” is not a single quantity $\rightarrow$ *method choice* defines the answer

Which features matter?

Example: Predicting bike rentals

Term	Estimate	SE	p
hour	7.8	0.4	< 0.001
humidity	-197.6	16.4	< 0.001
temperature	350.8	22.5	< 0.001
windspeed	18.9	22.6	0.403

- Linear model: compare $|\hat{\beta}_j|$, t-test p -values
- Random forest, XGBoost, neural net: no β to read
- Feature importance (FI) fills into that gap

`lm(count ~ .)` on bike sharing data (more later)



But: Important *for what*, and *how* is it measured?

Importance $\neq$ effect

Effect

- “What happens to $\hat{y}$ *when X_j changes?*”
- Linear model: β_j
- *Signed*, in units of X_j
- For arbitrary prediction methods:
 - Partial dependence plots (*PDP*)
 - Accumulated local effects (*ALE*)

Importance

- “*How much* does the model / learner rely on X_j ?”
- Linear model: $|\beta_j|$ (eyeballing), t-test
- *Unsigned* (generally), one number per feature
- Extends to nonlinear relationships + interactions, depending on learner
- Scale depends on importance method and loss
 - Usually read as a *ranking*

Feature Importance Estimands

A rough taxonomy of interpretability methods

Today: model-agnostic, loss-based, global

4

Model-specific

(G)LM coefficients, Gini impurity, gradient-based

Prediction-based

How much do predictions change without a feature?

Local

One observation's prediction (e.g., SHAP)

vs.

vs.

vs.

Model-agnostic

supports any model / learner

Loss-based

how much worse does the model get without a feature?

Global

the model as a whole, feature-specific

What are we estimating?

Which question?

- **Marginal:** $X_j \not\perp\!\!\!\perp Y$
Is X_j related to Y at all?
- **Conditional:** $X_j \not\perp\!\!\!\perp Y \mid X_{-j}$
Does X_j add information *given the others*?

Which target?

- **Model:** this one fitted model
- **Learner:** the fitting procedure, across datasets
- **Data-generating process** (Williamson et al. 2023)
 - $\approx$ Learner importance + Bayes-optimal predictor

⚠ Popular methods estimate *model* importance first, research questions often target *data*

Feature Importance Methods

Why not just refit?

Leave-one-covariate-out (LOCO)

(Lei et al. 2018)

Calculation

1. Fit model $\hat{f}$ with all features
2. **Refit** $\hat{f}_{-j}$ without X_j
3. Compare test loss

$$\text{LOCO}_j = L(\hat{f}_{-j}) - L(\hat{f})$$

- Effectively: one step of backward selection
- Groups: leave-one-**group**-out (LOGO)
- Leave-one-covariate-**in** (LOCI) = univariate association
- **Cost**: one refit per feature → expensive for
 - Complex algorithms / tuning
 - Huge datasets



Test loss, e.g. mean squared error

$$L(\hat{f}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \rightarrow \text{importance is on the scale of the } \textit{loss}$$

Don't refit, perturb

Permutation feature importance (PFI) (Breiman 2001)

7

Calculation

1. Fit model $\hat{f}$ once
2. *Shuffle* X_j in the test data: $\tilde{x}_j$
3. Compare test loss before and after

$$\text{PFI}_j = L(\hat{f}, \tilde{x}_j) - L(\hat{f})$$

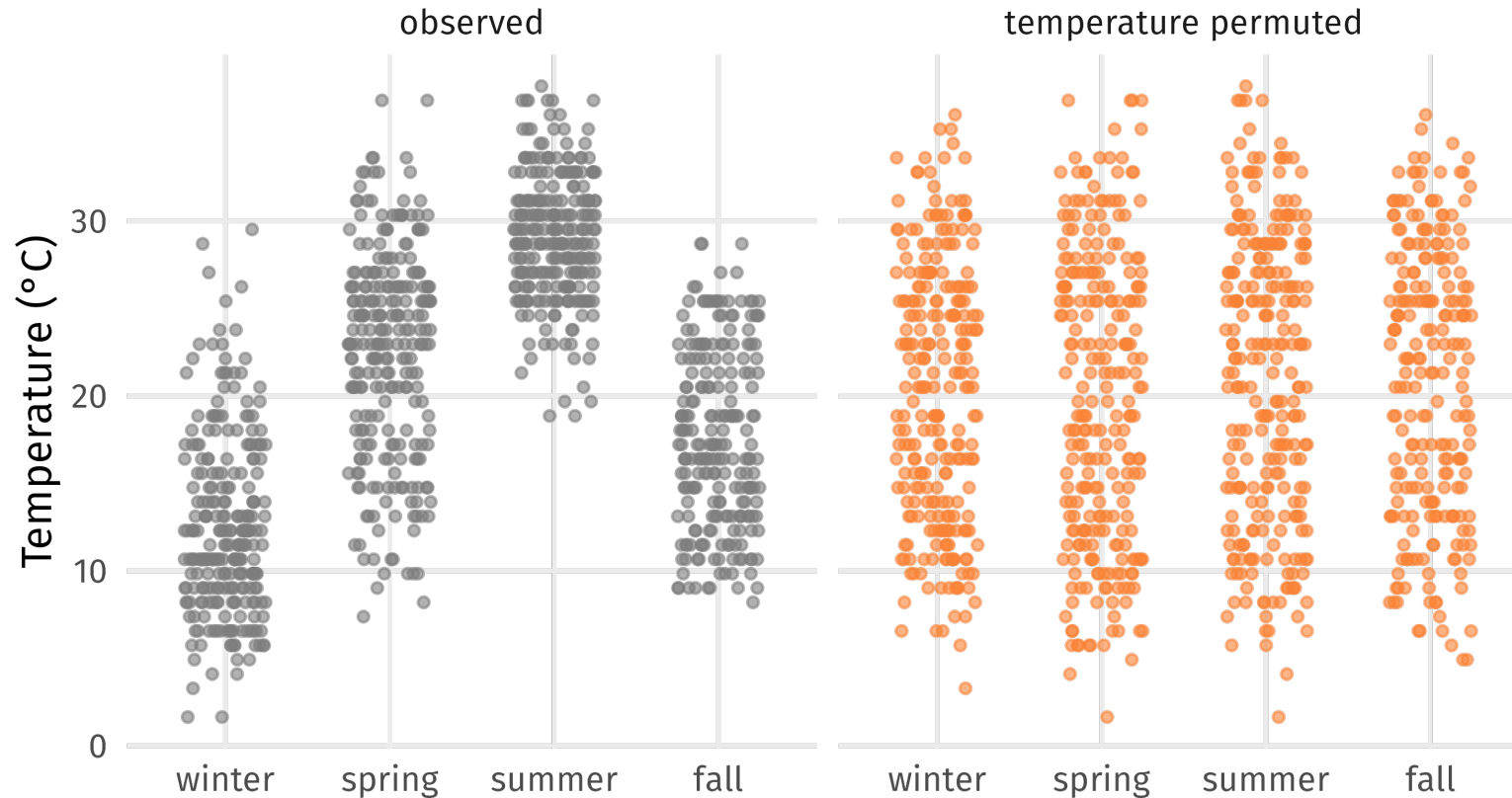
4. Repeat k times and average for stability

- No refits, works with any learner
- Breaks the link to Y *and* to the other features
- Correlated features → implausible combinations

→ A 20-year-old patient with 30 years of smoking

Shuffling leaves the data

Model is evaluated where it has never seen data (Hooker et al. 2021)



Perturb, but keep the dependence

Conditional feature importance (CFI)

- Sample replacement values $\tilde{x}_j$ from $X_j \mid X_{-j}$ instead of marginal shuffling
- Removes only the part of X_j *not* explained by the others
- → Stays on the data manifold, no extrapolation
- Otherwise: Same procedure as PFI:

$$\text{CFI}_j = L(\hat{f}; \tilde{x}_j) - L(\hat{f}), \quad \tilde{x}_j \sim X_j \mid X_{-j}$$

! Needs samples $\tilde{x}_j$ from $X_j \mid X_{-j}$, which are not easy to get

Where do conditional samples come from?

- *Gaussian*: fast, continuous features only
- *Knockoff* framework (Candès et al. 2018)
- Tabular data is usually *mixed*: continuous, categorical, missing values



Adversarial random forests (ARF)

- Generative model for tabular data, estimates conditional density (Watson et al. 2023)
- Categorical features, missing values
- Default sampler for CFI in `{xplainfi}`

Between PFI and CFI

Relative feature importance (RFI) (König et al. 2021)

Sample X_j from $X_j \mid X_G$ for a chosen *conditioning set* G

$$G = \emptyset$$

- PFI
- “Does the model use X_j ?”

$$G = \{\text{age, sex}\}$$

- RFI
- “Beyond age and sex?”

$$G = -j$$

- CFI
- “Beyond all other features?”

Is it more than noise?

Inference for CFI

- CFI is a mean of per-observation loss differences, $\tilde{x}_i: x_i$ with x_{ij} perturbed

$$\Delta_i = \left(y_i - \hat{f}(\tilde{x}_i) \right)^2 - \left(y_i - \hat{f}(x_i) \right)^2, \quad \text{CFI}_j = \overline{\Delta} = \frac{1}{n} \sum_{i=1}^n \Delta_i$$

- One-sided paired test on $\overline{\Delta}$ (e.g., t-test, Wilcoxon)
 - *CPI*: “conditional predictive impact” test using knockoffs (Watson and Wright 2021)
 - *cARFi*: same test with ARF sampling (Blesch et al. 2025)

💡 Holdout set, model fixed → *model* importance

And now for some game theory

Shapley Additive Global importance (SAGE) (Covert et al. 2020)

- Value of a feature set S (“coalition”):
Expected loss reduction when only features in S are known

$$v(S) = \mathbb{E}[\ell(Y, f_{\emptyset})] - \mathbb{E}[\ell(Y, f_S(X_S))]$$

- Shapley value:** average gain from adding j (Shapley 1953)

$$\varphi_j = \sum_{S \subseteq D \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} [v(S \cup \{j\}) - v(S)]$$

- Unknown features: MC integration over marginal or conditional samples ($\approx$ PFI vs. CFI)



SAGE vs. SHAP

SAGE: *global*, explains the *loss* for features
(Covert et al. 2020)

SHAP: *local*, explains *prediction* for observations
(Lundberg and Lee 2017)

Application Example

The bike sharing dataset

Capital Bikeshare, Washington, D.C., 2011–2012

- **Outcome:** bike rentals per hour
- 3000 samples from ~17k
- Gradient boosting (xgboost, pre-tuned) and lm
- Holdout: 80% train, 20% test

9 features

- **Calendar:** hour, weekday, working day, season, year
- **Weather:** temperature, humidity, wind speed, weather type

Setup for code example

{mlr3} objects, defined once

```
library(xplainfi)
library(mlr3learners)
library(mlr3pipelines)

task <- tsk("bike_sharing")
learner <- as_learner(
  po("encode") %>>%
  lrn("regr.xgboost", nrounds = 4932, eta = 0.0023, max_depth = 8)
)
measure <- msr("regr.mse")
resampling <- rsmp("holdout", ratio = 0.8)
```

```
pfi <- PFI$new(  
  task, learner, measure,  
  resampling = resampling,  
  n_repeats = 10  
)  
pfi$compute()  
pfi$importance()
```

Feature	PFI
hour	43478
temperature	8526
working_day	6532
year	6120
humidity	1983
season	1365
weekday	912
weather	673
windspeed	98

```
cfi <- CFI$new(  
  task, learner, measure,  
  resampling = resampling,  
  n_repeats = 10,  
  sampler = ConditionalARFSampler$new(task)  
)  
cfi$compute()  
cfi$importance()
```

Feature	CFI
hour	36061
year	5537
temperature	4817
working_day	3268
humidity	1180
season	1020
weekday	735
weather	567
windspeed	89

LOCO

```
loco <- LOCO$new(  
  task, learner, measure,  
  resampling = resampling  
)  
loco$compute()  
loco$importance()
```

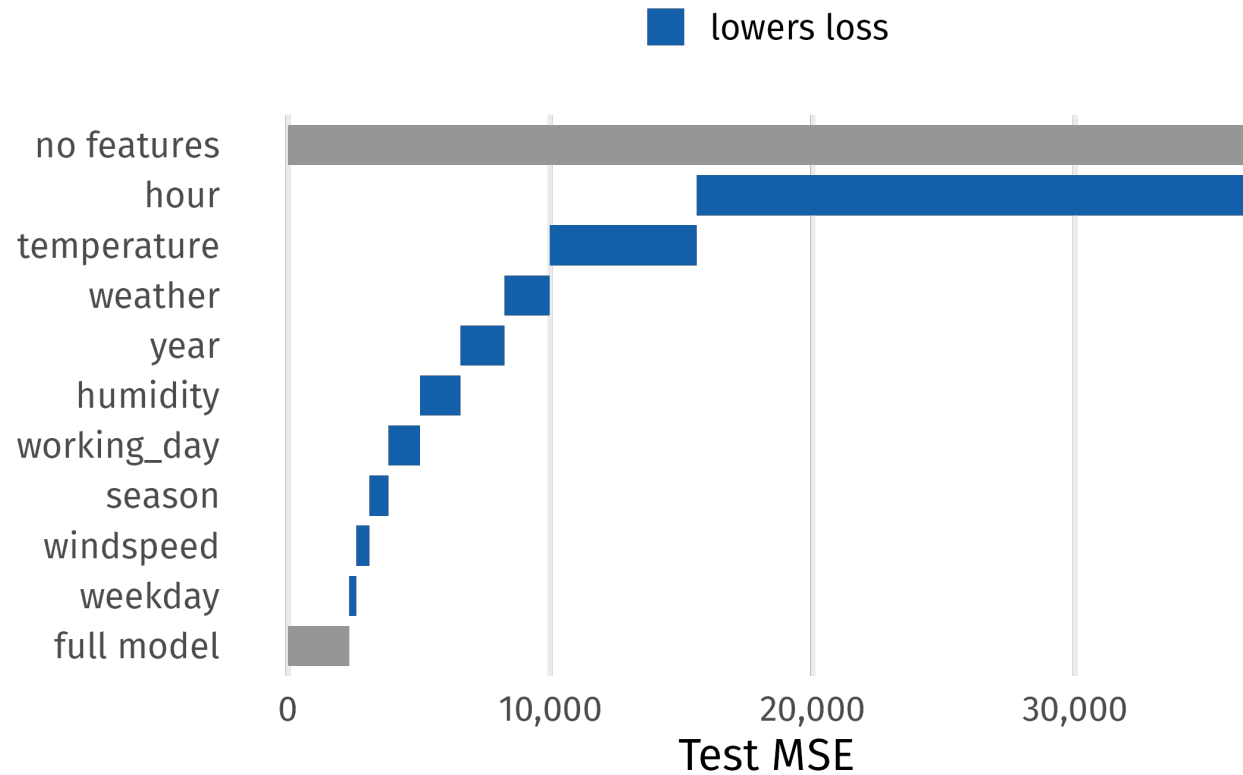
Feature	LOCO
hour	20460
year	3285
temperature	1253
working_day	646
season	542
weather	331
weekday	239
humidity	150
windspeed	-32

```
sage <- MarginalSAGE$new(  
  task, learner, measure,  
  resampling = resampling,  
  estimator = "kernel",  
  n_coalitions = 16384  
)  
sage$compute()  
sage$importance()
```

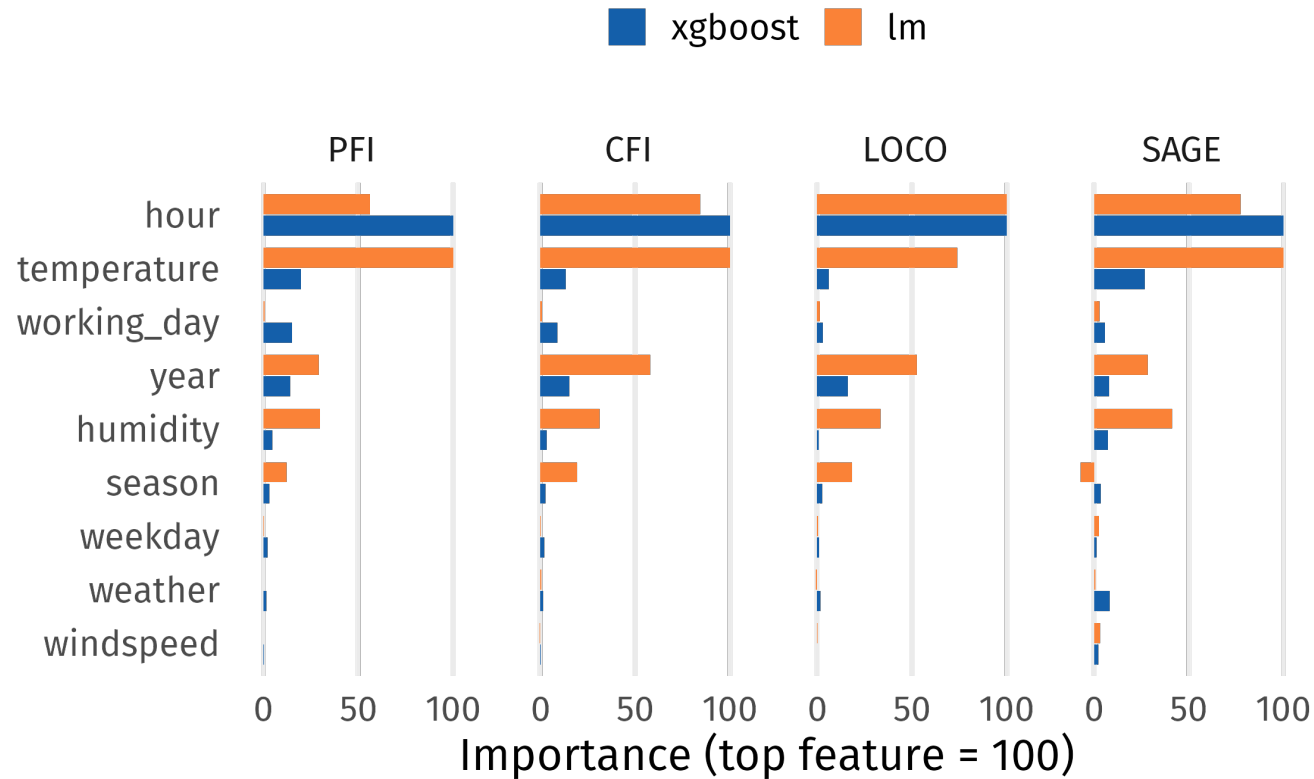
Feature	SAGE
hour	20923
temperature	5593
weather	1731
year	1687
humidity	1533
working_day	1197
season	725
windspeed	499
weekday	274

SAGE adds up

Loss without features → loss of the model, one step per feature



Comparing methods and learners



Inference

CFI with CPI, 600 held-out hours

```
cfi$importance(ci_method = "cpi", p_adjust = "BH")
```

Feature	CFI	SE	Lower 95%	p (BH)
hour	36061	2035	32064	< 0.001
year	5537	491	4572	< 0.001
temperature	4817	541	3754	< 0.001
working_day	3268	349	2582	< 0.001
humidity	1180	239	710	< 0.001
season	1020	146	733	< 0.001
weekday	735	99	540	< 0.001
weather	567	122	328	< 0.001
windspeed	89	66	-40	0.175

- hour, year, temperature clearly important
- windspeed: no evidence after BH adjustment

Summary

Question	Method	Watch out for
Does the model use X_j ?	PFI	extrapolation
Information beyond the others?	CFI: perturb LOCO: refit	conditional sampler refits, learner tuning
Share of the explained loss?	SAGE	cost, estimator noise

- CFI and LOCO: near-identical estimands, different estimation (Ewald et al. 2024)
- Generalize to learner/DGP importance: Combine with resampling procedure
→ inference becomes more complicated (e.g., Nadeau-Bengio correction)

Where to go next

1. `install.packages("xplainfi")`
2. Documentation: <https://mlr-org.github.io/xplainfi>
3. {xplainfi} and practical guidance:
[*xplainfi: Feature Importance and Statistical Inference for Machine Learning in R*](#)
 Burk et al. (2026), The R Journal
4. Methodological background:
[*A Guide to Feature Importance Methods for Scientific Inference*](#)
 Ewald et al. (2024), Explainable Artificial Intelligence (xAI 2024)

Thank you for your attention!



Contact

Lukas Burk

Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
28359 Bremen
Germany

burk@leibniz-bips.de



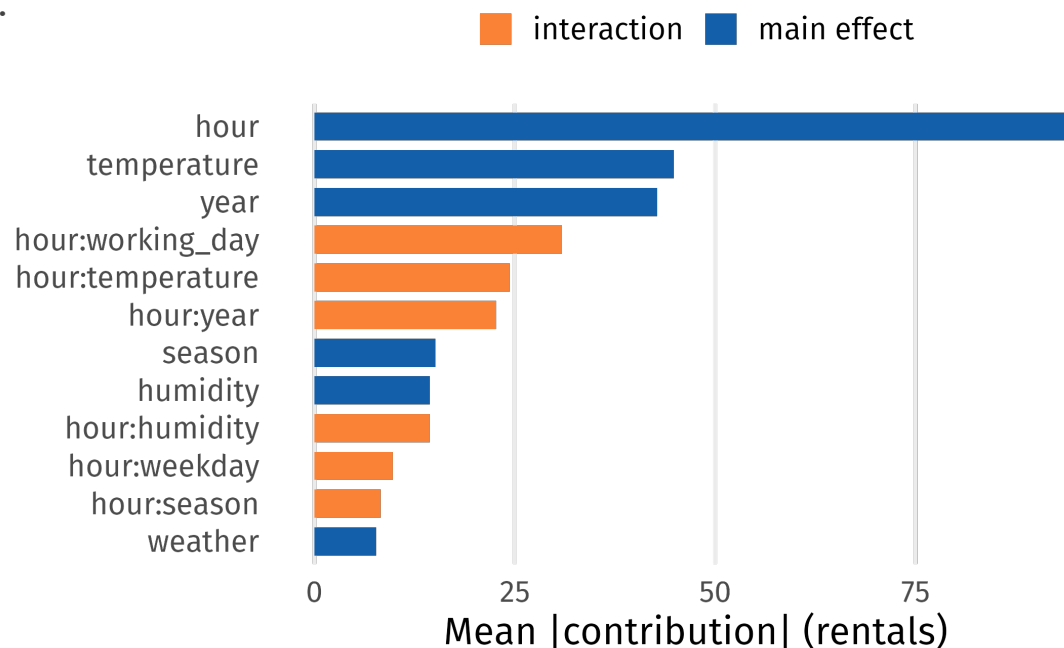
Backup

Prediction-based: decomposing $\hat{f}$

Functional decomposition of the xgboost model with `glex` (Hiabu et al. 2023)

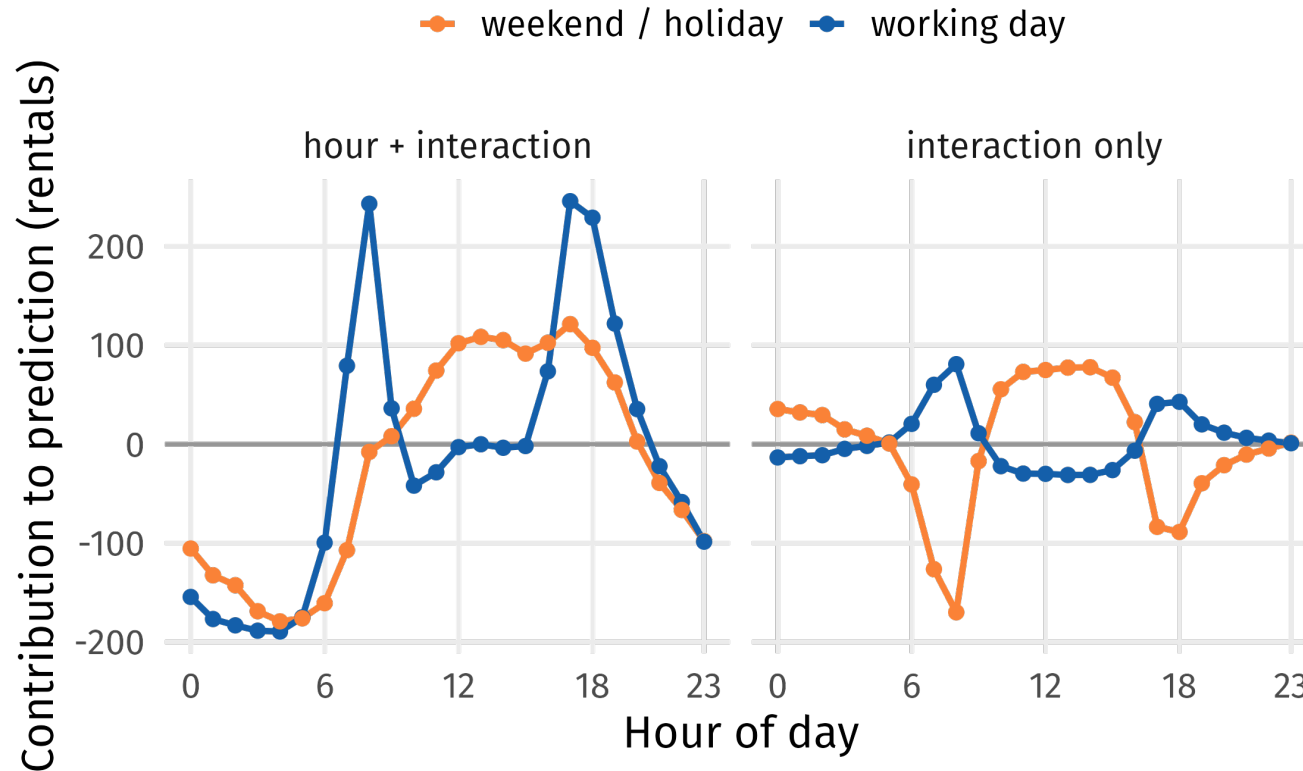
$$\hat{f}(x) = m_0 + \sum_j m_j(x_j) + \sum_{j < k} m_{jk}(x_j, x_k) + \dots$$

- Components on the scale of the *prediction* (rentals), not the loss
- Importance of a component: mean $|m_S|$
- No refits, no sampling: read off the trees



Low importance $\neq$ no effect

working_day matters through its interaction with hour



References

- Blesch K, Koenen N, Kapar J, et al (2025) Conditional Feature Importance with Generative Modeling Using Adversarial Random Forests. Proceedings of the AAAI Conference on Artificial Intelligence 39:15596–15604. <https://doi.org/10.1609/aaai.v39i15.33712>
- Breiman L (2001) Random Forests. Machine Learning 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Burk L, Ewald FK, Casalicchio G, et al (2026) Xplainfi: Feature Importance and Statistical Inference for Machine Learning in R
- Candès E, Fan Y, Janson L, Lv J (2018) Panning for Gold: ‘Model-X’ Knockoffs for High Dimensional Controlled Variable Selection. Journal of the Royal Statistical Society Series B: Statistical Methodology 80:551–577. <https://doi.org/10.1111/rssb.12265>
- Covert I, Lundberg SM, Lee S-I (2020) Understanding Global Feature Contributions With Additive Importance Measures. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 17212–17223
- Ewald FK, Bothmann L, Wright MN, et al (2024) A Guide to Feature Importance Methods for Scientific Inference. In: Longo L, Lapuschkin S, Seifert C (eds) Explainable Artificial Intelligence. Springer Nature Switzerland, Cham, pp 440–464
- Hiabu M, Meyer JT, Wright MN (2023) Unifying local and global model explanations by functional decomposition of low dimensional structures. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. PMLR, pp 7040–7060
- Hooker G, Mentch L, Zhou S (2021) Unrestricted Permutation Forces Extrapolation: Variable Importance Requires at Least One More Model, or There Is No Free Variable Importance. Statistics and Computing 31:82. <https://doi.org/10.1007/s11222-021-10057-z>
- König G, Molnar C, Bischl B, Grosse-Wentrup M (2021) Relative Feature Importance. pp 9318–9325
- Lei J, Max G, Alessandro R, et al (2018) Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association 113:1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>

References

- Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.
- Molnar C, Freiesleben T, König G, et al (2023) Relating the~Partial Dependence Plot and~Permutation Feature Importance to~the~Data Generating Process. In: Longo L (ed) Explainable Artificial Intelligence. Springer Nature Switzerland, Cham, pp 456–479
- Nadeau C, Bengio Y (2003) Inference for the Generalization Error. Machine Learning 52:239–281. <https://doi.org/10.1023/A:1024068626366>
- Shapley LS (1953) A Value for N-Person Games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, pp 307–318
- Strobl C, Boulesteix A-L, Kneib T, et al (2008) Conditional Variable Importance for Random Forests. BMC Bioinformatics 9:307. <https://doi.org/10.1186/1471-2105-9-307>
- Watson DS, Blesch K, Kapar J, Wright MN (2023) Adversarial Random Forests for Density Estimation and Generative Modeling. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. PMLR, pp 5357–5375
- Watson DS, Wright MN (2021) Testing Conditional Independence in Supervised Learning Algorithms. Machine Learning 110:2107–2129. <https://doi.org/10.1007/s10994-021-06030-6>
- Williamson BD, Gilbert PB, Simon NR, Carone M (2023) A General Framework for Inference on Algorithm-Agnostic Variable Importance. Journal of the American Statistical Association 118:1645–1658. <https://doi.org/10.1080/01621459.2021.2003200>