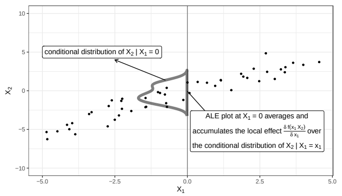
Interpretable Machine Learning

Regional Effects

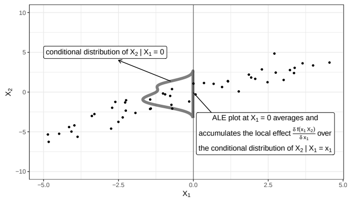


Learning goals

- Difference between feature effects and feature interactions
- REPID

Interpretable Machine Learning

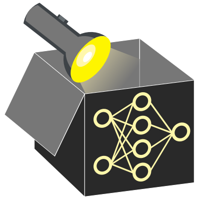
Regional Effects
REPID



Learning goals

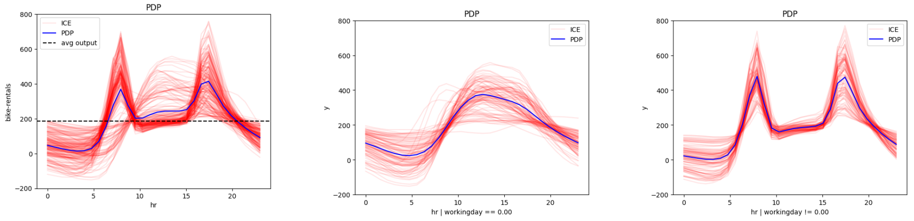
- Difference between feature effects and feature interactions
- REPID

WHY REGIONAL EXPLANATIONS?



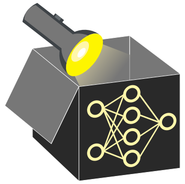
Problem: PD & ICE plots can be confounded by feature interactions.
Solution: Group homogeneous ICE curves in such a way that reduces the presence of individual interaction effects within a group $\rightsquigarrow$ Regional effect plots (REPs).

Global Effect	Regional Effect (1)	Regional Effect (2)



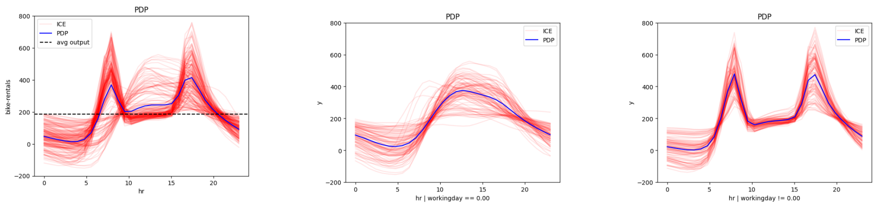
- Splitting by the "workingday" revealed 2 different patterns that we're clashed together in the initial plot
- $\cup_i(\texttt{regional\_explanation}_i) = \texttt{global\_explanation}$
- $\textit{Fidelity}(\texttt{regional\_explanation}_i) > \textit{Fidelity}(\texttt{global\_explanation})$

WHY REGIONAL EXPLANATIONS?



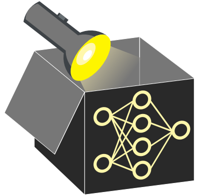
Problem: PD & ICE plots can be confounded by feature interactions.
Solution: Group homogeneous ICE curves in such a way that reduces the presence of individual interaction effects within a group
$\rightsquigarrow$ Regional effect plots (REPs).

Global Effect	Regional Effect (1)	Regional Effect (2)



- Splitting by the "workingday" revealed 2 different patterns that we're clashed together in the initial plot
- $\cup_i(\texttt{regional\_explanation}_i) = \texttt{global\_explanation}$
- $\textit{Fidelity}(\texttt{regional\_explanation}_i) > \textit{Fidelity}(\texttt{global\_explanation})$

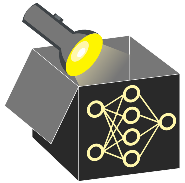
ICE CURVE: LOCAL FEATURE EFFECTS



Question: How do feature changes affect the prediction for **one observation**?
Idea: Split $\mathbf{x} = (x_j, \mathbf{x}_{-j})$ into x_j (feature of interest) and $\mathbf{x}_{-j}$ (remaining features)

- Replace observed value x_j with grid values $\tilde{x}_j$ while keeping values $\mathbf{x}_{-j}$ fixed
- Visualize function $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$ for varying $\tilde{x}_j$ (ICE)

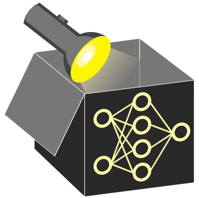
ICE CURVE: LOCAL FEATURE EFFECTS



Question: How do feature changes affect the prediction for **one obs.**?
Idea: Split $\mathbf{x} = (x_j, \mathbf{x}_{-j})$ into x_j (feat of interest) and $\mathbf{x}_{-j}$ (remaining feats)

- Replace observed values x_j with grid values $\tilde{x}_j$ while keeping $\mathbf{x}_{-j}$ fixed
- Visualize function $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$ for varying $\tilde{x}_j$ (ICE)

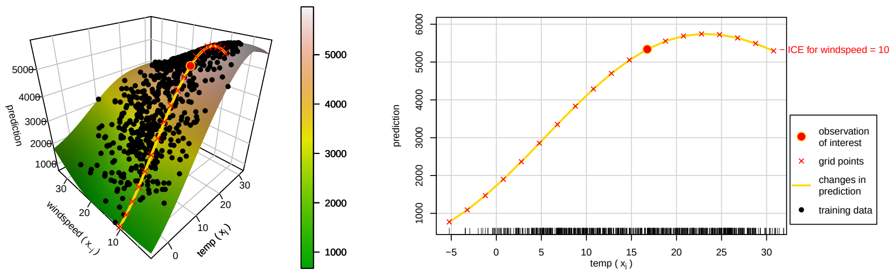
ICE CURVE: LOCAL FEATURE EFFECTS



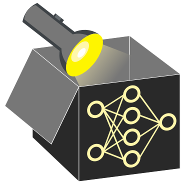
Question: How do feature changes affect the prediction for **one observation**?
Idea: Split $\mathbf{x} = (x_j, \mathbf{x}_{-j})$ into x_j (feature of interest) and $\mathbf{x}_{-j}$ (remaining features)

- Replace observed value x_j with grid values $\tilde{x}_j$ while keeping values $\mathbf{x}_{-j}$ fixed
- Visualize function $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$ for varying $\tilde{x}_j$ (ICE)

Example: SVM prediction surface (left), select observation and visualize changes in prediction for varying x_2 while keeping x_1 fixed $\Rightarrow$ **local interpretation**



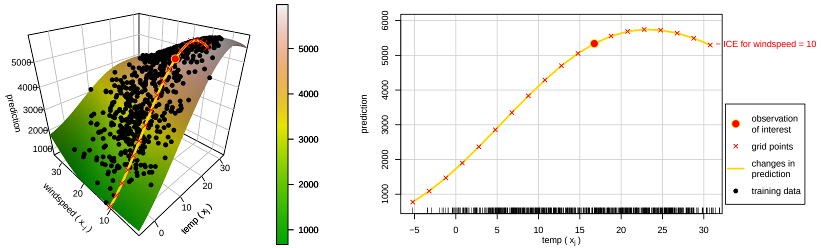
ICE CURVE: LOCAL FEATURE EFFECTS



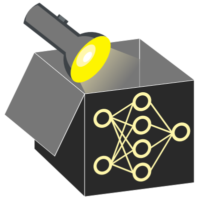
Question: How do feature changes affect the prediction for **one obs.**?
Idea: Split $\mathbf{x} = (x_j, \mathbf{x}_{-j})$ into x_j (feat of interest) and $\mathbf{x}_{-j}$ (remaining feats)

- Replace observed values x_j with grid values $\tilde{x}_j$ while keeping $\mathbf{x}_{-j}$ fixed
- Visualize function $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$ for varying $\tilde{x}_j$ (ICE)

Example: SVM prediction surface (left), select obs. and visualize changes in prediction for varying x_2 while keeping x_1 fixed $\Rightarrow$ **local interpretation**



PD PLOT - GLOBAL FEATURE EFFECTS



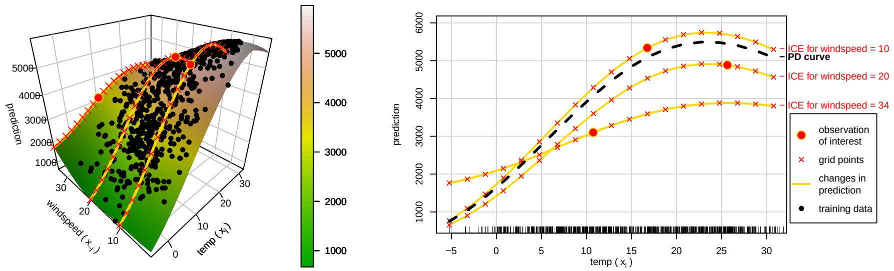
Question: How do changes of feature values affect model prediction **on average**?

- **PD function**: Integrate out effect of X_{-j} to obtain marginal effect of x_j

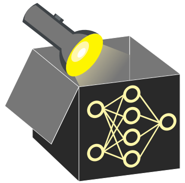
$$f_j^{PD}(\tilde{x}_j) = \mathbb{E}_{X_{-j}}[\hat{f}(\tilde{x}_j, X_{-j})] = \int \hat{f}(\tilde{x}_j, X_{-j}) d\mathbb{P}(X_{-j})$$

- **Estimate (MC integration):** Average ICE curves at grid points $\tilde{x}_j$

$$\hat{f}_j^{PD}(\tilde{x}_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}^{(i)})$$



PD PLOT - GLOBAL FEATURE EFFECTS



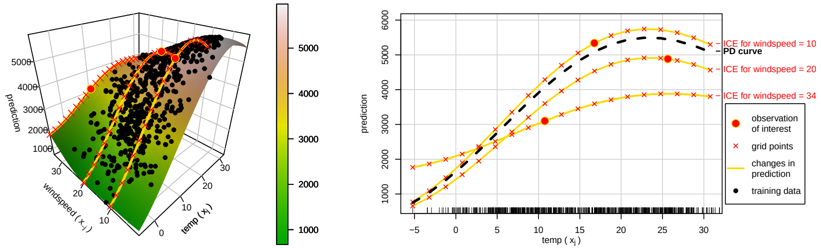
Question: How do changes of feat values affect model prediction **on avg.**?

- **PD function**: Integrate out effect of X_{-j} to obtain marginal effect of x_j

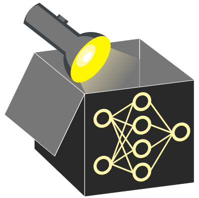
$$f_j^{PD}(\tilde{x}_j) = \mathbb{E}_{X_{-j}}[\hat{f}(\tilde{x}_j, X_{-j})] = \int \hat{f}(\tilde{x}_j, X_{-j}) d\mathbb{P}(X_{-j})$$

- **Estimate (MC integration):** Avgerage ICE curves at grid points $\tilde{x}_j$

$$\hat{f}_j^{PD}(\tilde{x}_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}^{(i)})$$



FEATURE INTERACTIONS



Hooker (2004, 2007): Functional ANOVA decomposition of a function

$$\hat{f}(\mathbf{x}) = g_0 + \underbrace{\textstyle\sum_{j=1}^{p} g_j(x_j)}_{\text{main effect}} + \underbrace{\textstyle\sum_{j \neq k} g_{j,k}(x_j, x_k)}_{\text{two-way interaction effect}} + \cdots + \underbrace{g_{1,2,\dots,p}(\mathbf{x})}_{\text{p-way interaction effect}}$$

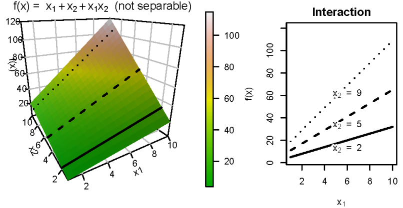
Friedman and Popescu (2008):

$\Rightarrow$ If x_j and $\mathbf{x}_{-j}$ do not interact, we can decompose $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

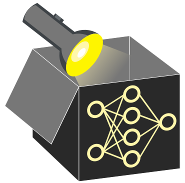
$\Rightarrow$ If x_j and x_k do not interact, we can decompose $f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

Example: Not additively separable:

$f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2 \neq g(x_1) + g(x_2)$



FEATURE INTERACTIONS



Hooker (2004, 2007): Functional ANOVA decomp. of a function

$$\hat{f}(\mathbf{x}) = g_0 + \underbrace{\textstyle\sum_{j=1}^{p} g_j(x_j)}_{\text{main effect}} + \underbrace{\textstyle\sum_{j \neq k} g_{j,k}(x_j, x_k)}_{\text{two-way interaction effect}} + \cdots + \underbrace{g_{1,2,\dots,p}(\mathbf{x})}_{\text{p-way interaction effect}}$$

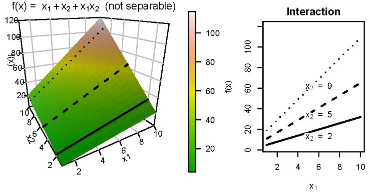
Friedman and Popescu (2008):

$\Rightarrow$ If x_j and $\mathbf{x}_{-j}$ don't interact, we can decomp. $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

$\Rightarrow$ If x_j and x_k don't interact, decomposition:
$f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

Example: Not additively separable:

$f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2 \neq g(x_1) + g(x_2)$



FEATURE INTERACTIONS

Hooker (2004, 2007): Functional ANOVA decomposition of a function

$$\hat{f}(\mathbf{x}) = g_0 + \underbrace{\sum_{j=1}^{p} g_j(x_j)}_{\text{main effect}} + \underbrace{\sum_{j \neq k} g_{j,k}(x_j, x_k)}_{\text{two-way interaction effect}} + \cdots + \underbrace{g_{1,2,\ldots,p}(\mathbf{x})}_{\text{p-way interaction effect}}$$

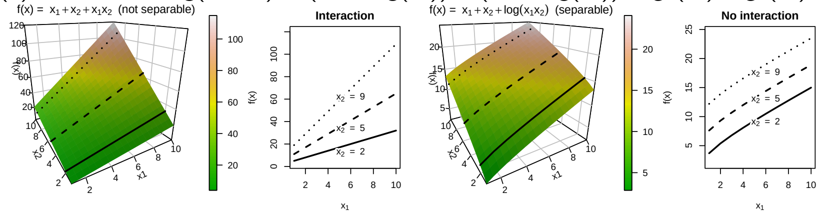
Friedman and Popescu (2008):

$\Rightarrow$ If x_j and $\mathbf{x}_{-j}$ do not interact, we can decompose $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

$\Rightarrow$ If x_j and x_k do not interact, we can decompose $f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

Example: Separable:

$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = (x_1 + \log(x_1)) + (x_2 + \log(x_2)) = g_1(x_1) + g_2(x_2)$



FEATURE INTERACTIONS

Hooker (2004, 2007): Functional ANOVA decomp. of a function

$$\hat{f}(\mathbf{x}) = g_0 + \underbrace{\sum_{j=1}^{p} g_j(x_j)}_{\text{main effect}} + \underbrace{\sum_{j \neq k} g_{j,k}(x_j, x_k)}_{\text{two-way interaction effect}} + \cdots + \underbrace{g_{1,2,\ldots,p}(\mathbf{x})}_{\text{p-way interaction effect}}$$

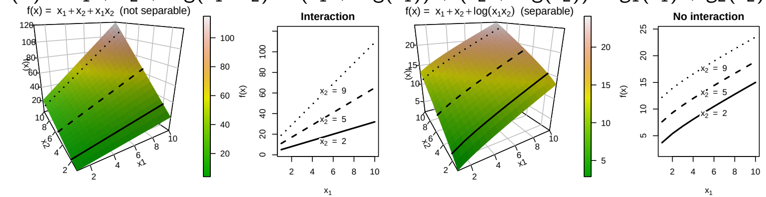
Friedman and Popescu (2008):

$\Rightarrow$ If x_j and $\mathbf{x}_{-j}$ don't interact, we can decomp. $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

$\Rightarrow$ If x_j and x_k don't interact, decomposition:
$f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

Example: Separable:

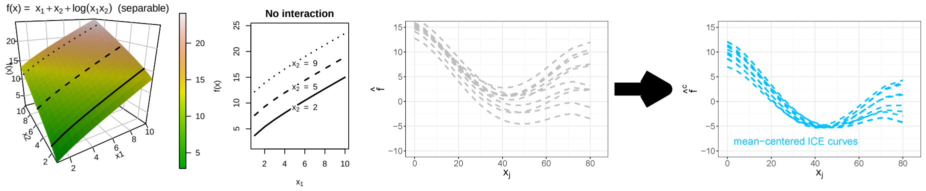
$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = (x_1 + \log(x_1)) + (x_2 + \log(x_2)) = g_1(x_1) + g_2(x_2)$



REPID: REGIONAL EFFECT PLOTS ▸ Herbinger et al. (2022)

Recall: Different shapes of ICE curves indicate interactions (ignore vertical shifts)
$\Rightarrow$ Focus on shape differences of **mean-centered ICE curves**.
Mean-centered ICE curve for obs. $\mathbf{x}$ evaluated at m grid points $\tilde{x}_j^{(1)}, \dots, \tilde{x}_j^{(m)}$ is:

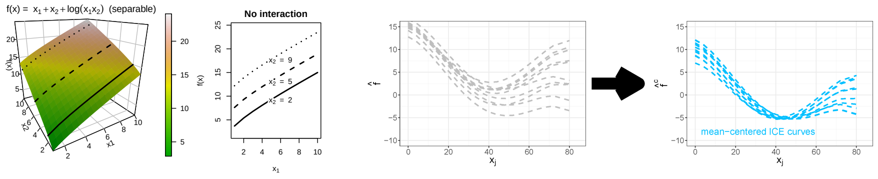
$$\hat{f}^c(\tilde{x}_j, \mathbf{x}_{-j}) = \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}) - \frac{1}{m}\sum_{k=1}^{m} \hat{f}(\tilde{x}_j^{(k)}, \mathbf{x}_{-j})$$



REPID: REGIONAL EFFECT PLOTS ▸ HERBINGER_2022

Recall: Different shapes of ICE curves $\Rightarrow$ interactions (ignore vertical shifts)
$\Rightarrow$ Focus on shape differences of **mean-centered ICE curves**.
Mean-centered ICE curve for obs. $\mathbf{x}$ evaluated at m grid points $\tilde{x}_j^{(1)}, \dots, \tilde{x}_j^{(m)}$ is:

$$\hat{f}^c(\tilde{x}_j, \mathbf{x}_{-j}) = \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}) - \frac{1}{m}\sum_{k=1}^{m} \hat{f}(\tilde{x}_j^{(k)}, \mathbf{x}_{-j})$$



REGIONAL EFFECTS - SYNTHETIC EXAMPLE

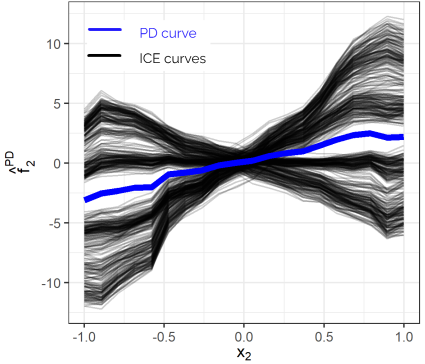
Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1,1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 \boxed{-8X_2 + 8X_2\mathbb{1}_{(X_1>0)} + 16X_2\mathbb{1}_{(X_3=0)}} + \epsilon$

$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of $\boxed{X_2}$ is misleading due to interactions $\rightsquigarrow$ ICE
- ICE curves do not identify the interacting features



REGIONAL EFFECTS - SYNTHETIC EXAMPLE

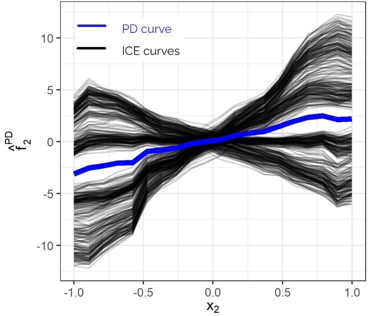
Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1,1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 \boxed{-8X_2 + 8X_2 1_{(X_1>0)} + 16X_2 1_{(X_3=0)}} + \epsilon$

$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of $\boxed{X_2}$ is misleading due to interactions $\rightsquigarrow$ ICE
- ICE curves do not identify the interacting features



REGIONAL EFFECTS - SYNTHETIC EXAMPLE

Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

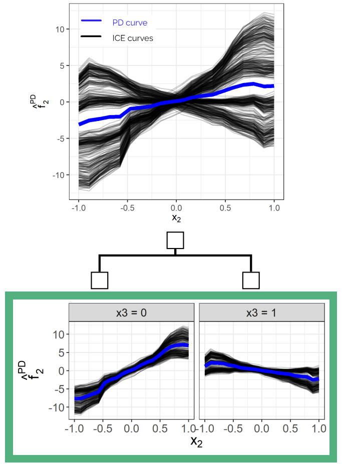
$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(X_1>0)} + 16X_2\mathbb{1}_{(X_3=0)} + \epsilon$

$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions $\rightsquigarrow$ ICE
- ICE curves do not identify the interacting features

Idea: Find regions with similar ICE curves and aggregate them to regional effects



REGIONAL EFFECTS - SYNTHETIC EXAMPLE

Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

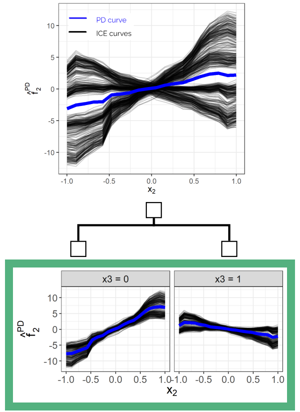
$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2 1_{(X_1>0)} + 16X_2 1_{(X_3=0)} + \epsilon$

$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions $\rightsquigarrow$ ICE
- ICE curves do not identify the interacting features

Idea: Find regions with similar ICE curves and aggregate them to regional effects



Interpretable Machine Learning – 6 / 9

REGIONAL EFFECTS - SYNTHETIC EXAMPLE

Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1,1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(X_1>0)} + 16X_2\mathbb{1}_{(X_3=0)} + \epsilon$

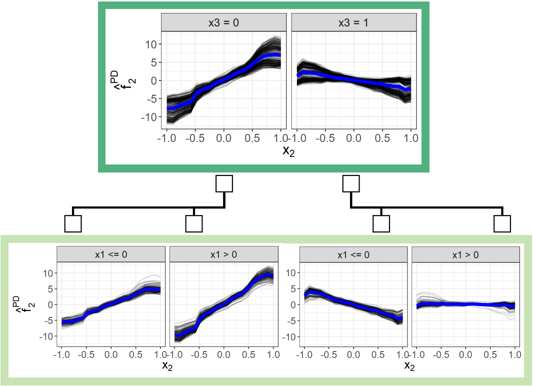
$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions $\rightsquigarrow$ ICE
- ICE curves do not identify the interacting features

Idea: Find regions with similar ICE curves and aggregate them to regional effects

Regional effect (blue curves) $\hat{=}$ Estimate PD curve in each region



$\hat{f}_2^{PD}(X_2) \approx 8X_2$ $\approx 16X_2$ $\approx -8X_2$ ≈ 0

$\Rightarrow$ Additive decomposition of global feature effect

REGIONAL EFFECTS - SYNTHETIC EXAMPLE

Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1,1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2 1_{(X_1>0)} + 16X_2 1_{(X_3=0)} + \epsilon$

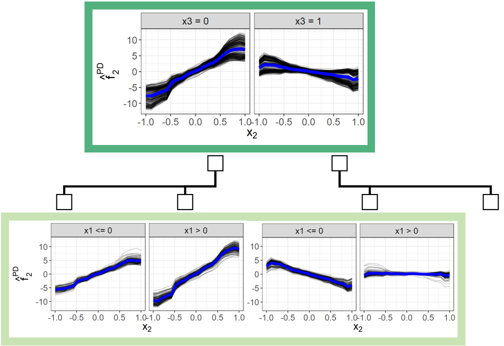
$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions $\rightsquigarrow$ ICE
- ICE curves do not identify the interacting features

Idea: Find regions with similar ICE curves and aggregate them to regional effects

Regional effect (blue curves) $\hat{=}$ Estimate PD curve in each region



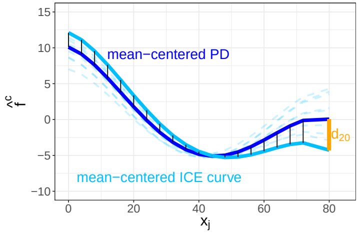
$\hat{f}_2^{PD}(X_2) \approx 8X_2$ $\approx 16X_2$ $\approx -8X_2$ ≈ 0

$\Rightarrow$ Additive decomposition of global feat effect

REGIONAL EFFECTS - DETAILS

Question: How to split the curves into regions?
Define risk as L2 loss of mean-centered ICE curves:

$$\mathcal{R}_j(\mathcal{N}) = \sum_{\mathbf{x} \in \mathcal{N}} \sum_{k=1}^{m} \underbrace{\left(\hat{f}^c(\tilde{x}_j^{(k)}, \mathbf{x}_{-j}) - \hat{f}^{PD,c}_{j|\mathcal{N}}(\tilde{x}_j^{(k)})\right)^2}_{d_k}$$



with the average feature effect in region $\mathcal{N} \subseteq \mathcal{X}$:

$$\hat{f}^{PD,c}_{j|\mathcal{N}}(\tilde{x}_j) = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{x} \in \mathcal{N}} \hat{f}^c(\tilde{x}_j, \mathbf{x}_{-j})$$

$\Rightarrow$ Measures interaction-related heterogeneity (variance) of ICE curves in $\mathcal{N}$

$\Rightarrow$ Recursive partitioning (CART): Find best feature-split combination that solves

$$\arg\min_{z,t} \; \mathcal{R}_j(\mathcal{N}_{left}) + \mathcal{R}_j(\mathcal{N}_{right})$$

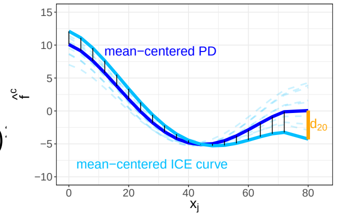
- $\mathcal{N}_{left} = \{\mathbf{x} \in \mathcal{N} | x_z \leq t\}$
- $\mathcal{N}_{right} = \{\mathbf{x} \in \mathcal{N} | x_z > t\}$
- Split point t for feature $x_z, z \in -j$

Intuition: Is another feature x_z responsible for the heterogeneity (measured by $\mathcal{R}_j$)?

REGIONAL EFFECTS - DETAILS

Question: How to split curves into regions?
Define risk as L2 loss of mean-centered ICE curves:

$$\mathcal{R}_j(\mathcal{N}) = \sum_{\mathbf{x} \in \mathcal{N}} \sum_{k=1}^{m} \underbrace{\left(\hat{f}^c(\tilde{x}_j^{(k)}, \mathbf{x}_{-j}) - \hat{f}^{PD,c}_{j|\mathcal{N}}(\tilde{x}_j^{(k)})\right)^2}_{d_k}$$



with the avg. feature effect in region $\mathcal{N} \subseteq \mathcal{X}$:

$$\hat{f}^{PD,c}_{j|\mathcal{N}}(\tilde{x}_j) = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{x} \in \mathcal{N}} \hat{f}^c(\tilde{x}_j, \mathbf{x}_{-j})$$

$\leadsto$ Measures interaction-related heterogeneity (variance) of ICE curves in $\mathcal{N}$

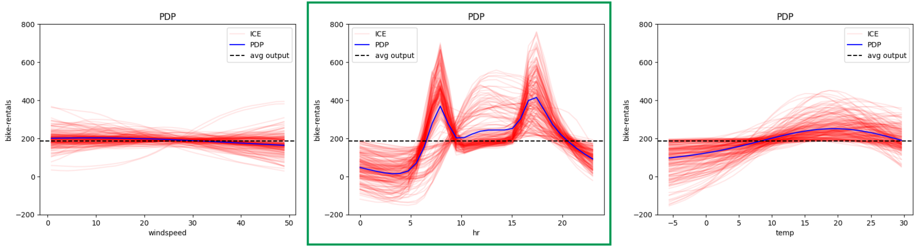
$\leadsto$ Recursive partitioning (CART): Find best feat-split combo that solves

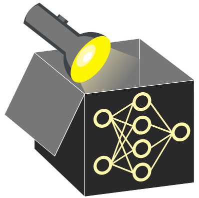
$$\arg\min_{z,t} \; \mathcal{R}_j(\mathcal{N}_{left}) + \mathcal{R}_j(\mathcal{N}_{right})$$

- $\mathcal{N}_{left} = \{\mathbf{x} \in \mathcal{N} | x_z \leq t\}$
- $\mathcal{N}_{right} = \{\mathbf{x} \in \mathcal{N} | x_z > t\}$
- Split point t for feature $x_z, z \in -j$

Intuition: Is another feature x_z responsible for the heterogeneity (measured by $\mathcal{R}_j$)?

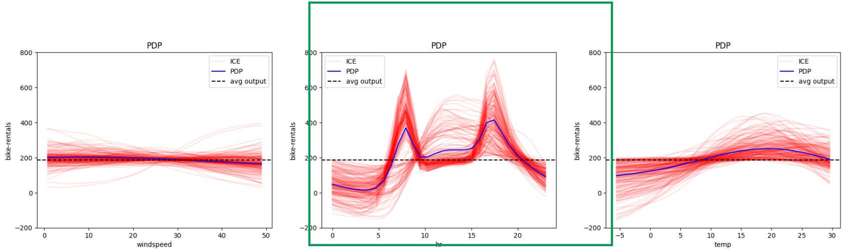
REGIONAL EFFECT PLOTS - REAL EXAMPLE

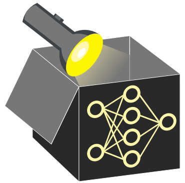




- Identify feature with highly heterogeneous local effects
  ⇝ `hour`: Most important and highly heterogeneous feature (highest variance)
- Find regions in feature space where this heterogeneity is minimal
  ⇝ Partition feature space using CART to minimize variance of mean-centered ICE curves within each region

REGIONAL EFFECT PLOTS - REAL EXAMPLE

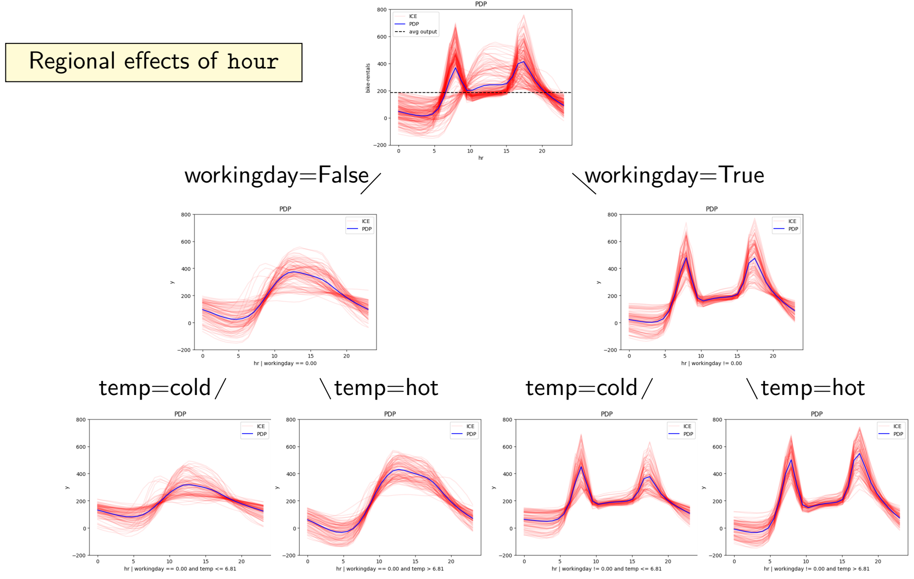




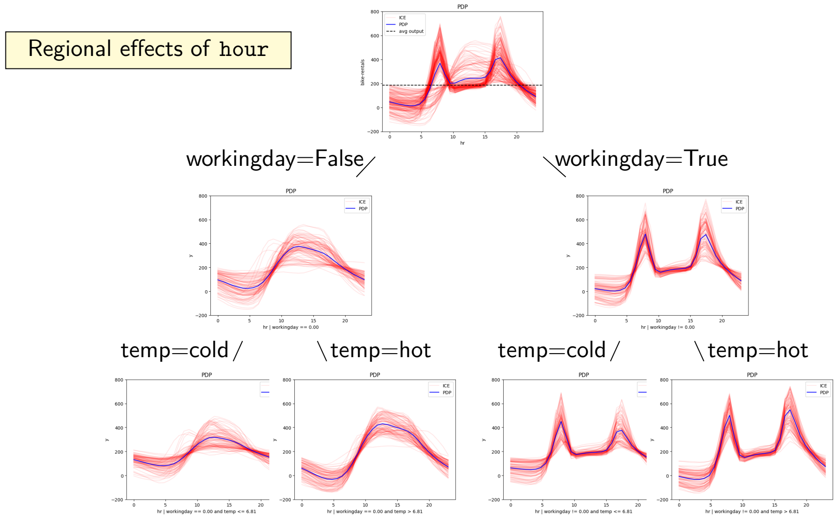
- Identify feature with highly heterogeneous local effects
  ⇝ `hour`: Most important; highly heterogeneous feat (highest variance)
- Find regions in feature space where this heterogeneity is minimal
  ⇝ Partition feature space using CART to minimize variance of mean-centered ICE curves within each region

Interpretable Machine Learning – 8 / 9

REGIONAL EFFECT PLOTS - REAL EXAMPLE

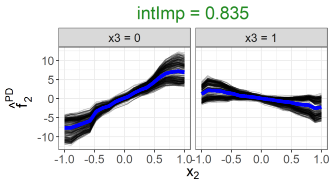


REGIONAL EFFECT PLOTS - REAL EXAMPLE



Interpretable Machine Learning

Regional Effects



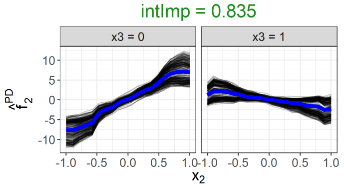
Learning goals

- Interaction quantification
- REPID interaction importance

Interpretable Machine Learning

Regional Effects

Interaction importance



Learning goals

- Interaction quantification
- REPID interaction importance

INTERACTION QUANTIFICATION

It's helpful to know not just how another feature changes the marginal effect of x_S but how strong that interaction is and want to rank features by it.

Approaches:

- **H-Statistics:** Variance of the deviation between the joint PDP and the sum of marginal PDPs (larger variance $\Rightarrow$ stronger interaction).
- **Greenwell's Interaction Index:** Difference between the variance of the PDP and the mean variance of centered ICE curves for the same feature pair.
- **SHAP interaction index (▸ Herbinger et al. (2022), ▸ Lundberg et al. 2018):** Proportion of all two-way interactions with x_i to which the j-th feature contributes.

INTERACTION QUANTIFICATION

It's helpful to know not just how another feature changes the marginal effect of x_S but how strong that interaction is and want to rank features by it.

Approaches:

- **H-Statistics:** Variance of the deviation between the joint PDP and the sum of marginal PDPs (larger variance $\Rightarrow$ stronger interaction).
- **Greenwell's Interaction Index:** Difference between variance of the PDP and the mean variance of centered ICE curves for the same feature pair.
- **SHAP interaction index (▸ Herbinger 2022, ▸ Lundberg 2018):** Proportion of all two-way interactions with x_i to which the j-th feature contributes.

INTERACTION QUANTIFICATION

It's helpful to know not just how another feature changes the marginal effect of x_S but how strong that interaction is and want to rank features by it.

Approaches:

- **H-Statistics:** Variance of the deviation between the joint PDP and the sum of marginal PDPs (larger variance $\Rightarrow$ stronger interaction).
- **Greenwell's Interaction Index:** Difference between the variance of the PDP and the mean variance of centered ICE curves for the same feature pair.
- **SHAP interaction index (** ▸ Herbinger et al. (2022) **,** ▸ Lundberg et al. 2018 **):** Proportion of all two-way interactions with x_i to which the j-th feature contributes.

Pitfalls:

- The values of H-Statistic and the Greenwell's Interaction Index are influenced by the main effects of the two regarded features.
- SHAP interaction index does not suffer from main effect problem. However, correlation between the two features can bias the interaction value. Same applies to the H-Statistic.

INTERACTION QUANTIFICATION

It's helpful to know not just how another feature changes the marginal effect of x_S but how strong that interaction is and want to rank features by it.

Approaches:

- **H-Statistics:** Variance of the deviation between the joint PDP and the sum of marginal PDPs (larger variance $\Rightarrow$ stronger interaction).
- **Greenwell's Interaction Index:** Difference between variance of the PDP and the mean variance of centered ICE curves for the same feature pair.
- **SHAP interaction index (** ▸ Herbinger 2022 **,** ▸ Lundberg 2018 **):** Proportion of all two-way interactions with x_i to which the j-th feature contributes.

Pitfalls:

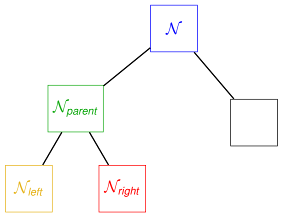
- The values of H-Statistic and the Greenwell's Interaction Index are influenced by the main effects of the two regarded features.
- SHAP interaction index does not suffer from main effect problem. However, correlation between the two features can bias the interaction value. Same applies to the H-Statistic.

REPID INTERACTION IMPORTANCE

On parent node level (for $\mathcal{N}_{parent}$):

$$intImp(\mathcal{N}_{parent}) = \frac{\mathcal{R}(\mathcal{N}_{parent}) - (\mathcal{R}(\mathcal{N}_{left}) + \mathcal{R}(\mathcal{N}_{right}))}{\mathcal{R}(\mathcal{N})}$$

Interpretation: Reduction of ICE curve variance after one split of $\mathcal{N}_{parent}$ into $\mathcal{N}_{left}$ and $\mathcal{N}_{right}$ relative to the ICE curve variance in the root node $\mathcal{N}$.

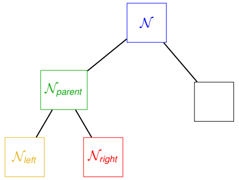


REPID INTERACTION IMPORTANCE

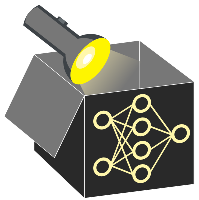
On parent node level (for $\mathcal{N}_{parent}$):

$$intImp(\mathcal{N}_{parent}) = \frac{\mathcal{R}(\mathcal{N}_{parent}) - (\mathcal{R}(\mathcal{N}_{left}) + \mathcal{R}(\mathcal{N}_{right}))}{\mathcal{R}(\mathcal{N})}$$

Interpretation: Reduction of ICE curve variance after one split of $\mathcal{N}_{parent}$ into $\mathcal{N}_{left}$ and $\mathcal{N}_{right}$ relative to the ICE curve variance in the root node $\mathcal{N}$.



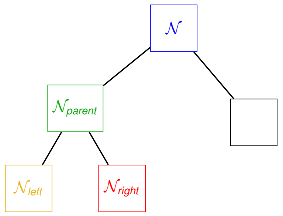
REPID INTERACTION IMPORTANCE

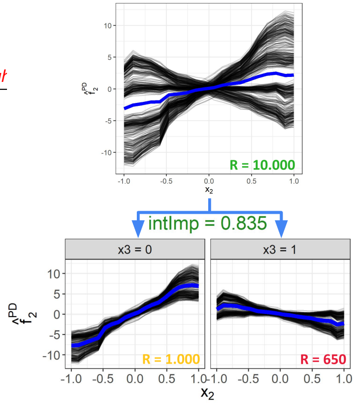


On parent node level (for $\mathcal{N}_{parent}$):

$$intImp(\mathcal{N}_{parent}) = \frac{\mathcal{R}(\mathcal{N}_{parent}) - (\mathcal{R}(\mathcal{N}_{left}) + \mathcal{R}(\mathcal{N}_{right}))}{\mathcal{R}(\mathcal{N})}$$

Interpretation: Reduction of ICE curve variance after one split of $\mathcal{N}_{parent}$ into $\mathcal{N}_{left}$ and $\mathcal{N}_{right}$ relative to the ICE curve variance in the root node $\mathcal{N}$.





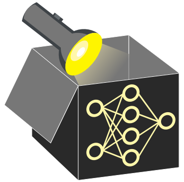
Split reduces 83.5% of variance.

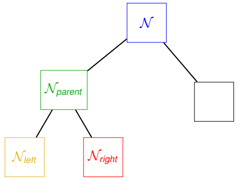
REPID INTERACTION IMPORTANCE

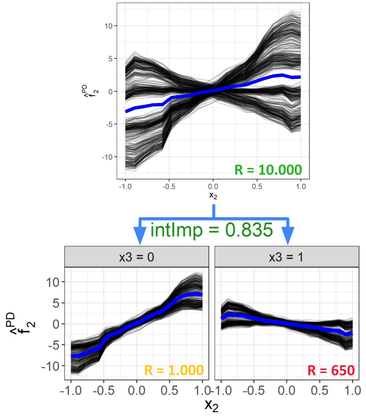
On parent node level (for $\mathcal{N}_{parent}$):

$$intImp(\mathcal{N}_{parent}) = \frac{\mathcal{R}(\mathcal{N}_{parent}) - (\mathcal{R}(\mathcal{N}_{left}) + \mathcal{R}(\mathcal{N}_{right}))}{\mathcal{R}(\mathcal{N})}$$

Interpretation: Reduction of ICE curve variance after one split of $\mathcal{N}_{parent}$ into $\mathcal{N}_{left}$ and $\mathcal{N}_{right}$ relative to the ICE curve variance in the root node $\mathcal{N}$.



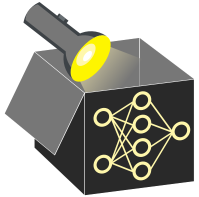




Split reduces 83.5% of variance.

Interpretable Machine Learning – 2 / 8

REPID INTERACTION IMPORTANCE



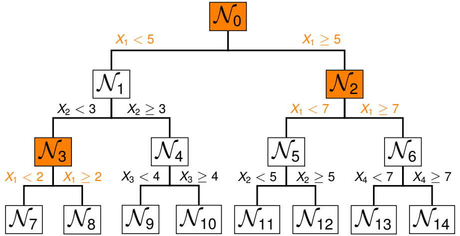
On feature level (for x_j):

$$\mathit{intImp}_j = \sum_{i \in \mathcal{B}_j} \mathit{intImp}(\mathcal{N}_i)$$

where $\mathcal{B}_j$ indexes parent nodes split by x_j.

Interpretation: Overall reduction of ICE curve variance due to splits by X_j (in %).

Example: For $X_1 \Rightarrow \mathcal{B}_1 = \{0, 2, 3\}$



REPID INTERACTION IMPORTANCE

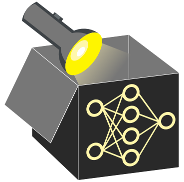
On feature level (for x_j):

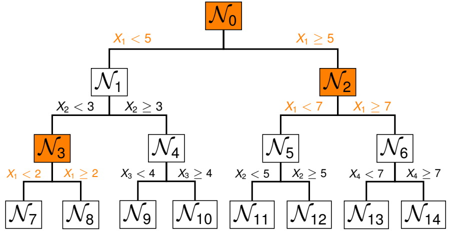
$$\mathit{intImp}_j = \sum_{i \in \mathcal{B}_j} \mathit{intImp}(\mathcal{N}_i)$$

where $\mathcal{B}_j$ indexes parent nodes split by x_j.

Interpretation: Overall reduction of ICE curve variance due to splits by X_j (in %).

Example: For $X_1 \Rightarrow \mathcal{B}_1 = \{0, 2, 3\}$





REPID INTERACTION IMPORTANCE

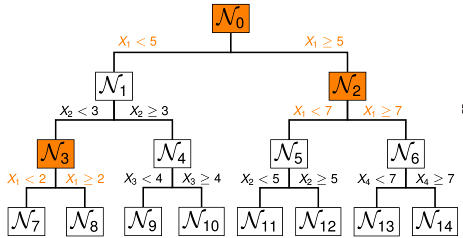
On feature level (for x_j):

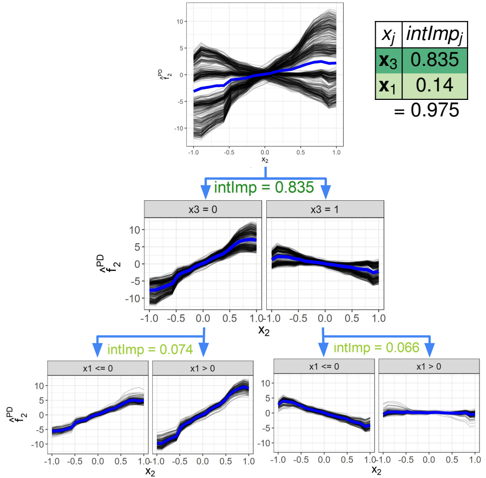
$$intImp_j = \sum_{i \in \mathcal{B}_j} intImp(\mathcal{N}_i)$$

where $\mathcal{B}_j$ indexes parent nodes split by x_j.

Interpretation: Overall reduction of ICE curve variance due to splits by X_j (in %).

Example: For $X_1 \Rightarrow \mathcal{B}_1 = \{0, 2, 3\}$





x_j	$intImp_j$
x_3	0.835
x_1	0.14
	= 0.975

Note: *intImp* can also be used as a stopping criterion.

REPID INTERACTION IMPORTANCE

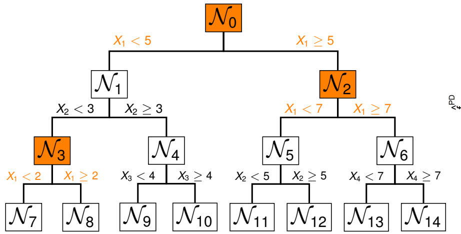
On feature level (for x_j):

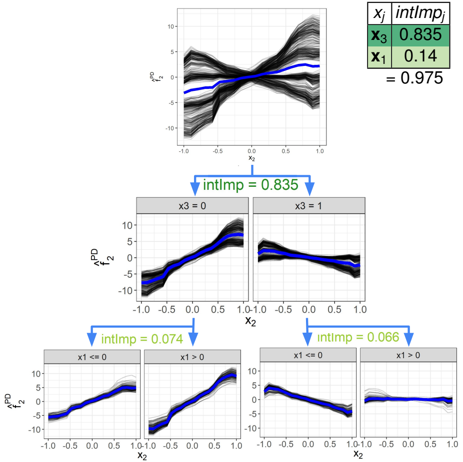
$$intImp_j = \sum_{i \in \mathcal{B}_j} intImp(\mathcal{N}_i)$$

where $\mathcal{B}_j$ indexes parent nodes split by x_j.

Interpretation: Overall reduction of ICE curve variance due to splits by X_j (in %).

Example: For $X_1 \Rightarrow \mathcal{B}_1 = \{0, 2, 3\}$



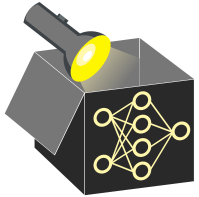


x_j	$intImp_j$
x_3	0.835
x_1	0.14
	= 0.975

Note: *intImp* can also be used as a stopping criterion.

Interpretable Machine Learning – 3 / 8

OUTPERFORMING SOTA



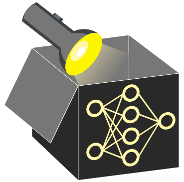
Simulation setting

- Draw 1000 i.i.d. samples from $X_1, \dots, X_4 \sim \mathcal{U}(-1, 1)$
- True underlying function: $f(\mathbf{x}) = \sum_{j=1}^{4} \mathbf{x}_j + \mathbf{x}_1\mathbf{x}_2 + \mathbf{x}_2\mathbf{x}_3 + \mathbf{x}_1\mathbf{x}_3 + \mathbf{x}_1\mathbf{x}_2\mathbf{x}_3 + \epsilon$
- Fit a correctly specified linear model (interactions with $\mathbf{x}_4$ are excluded)
- 30 repetitions, measure interaction strength between $\mathbf{x}_2$ and all other 3 features

Which methods are sensitive to changes in main effect sizes or feature correlations?

Pitfall	REPID	H-Statistic	Greenwell	SHAP
sensitive to changes of main effect	No	Yes	Yes	No
sensitive to changes of correlation between $\mathbf{x}_j$ and other features	No	Yes	No	Yes

OUTPERFORMING SOTA



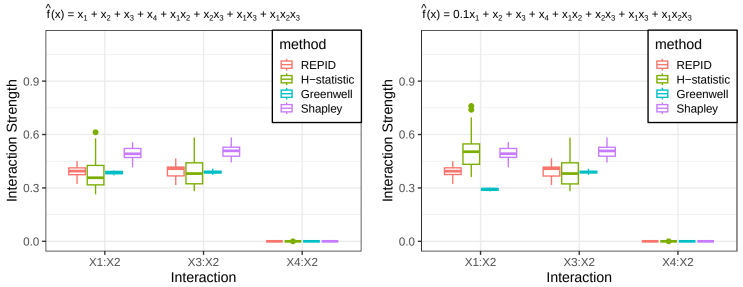
Simulation setting

- Draw 1000 i.i.d. samples from $X_1, \dots, X_4 \sim \mathcal{U}(-1, 1)$
- True underlying function:
  $f(\mathbf{x}) = \sum_{j=1}^{4} \mathbf{x}_j + \mathbf{x}_1\mathbf{x}_2 + \mathbf{x}_2\mathbf{x}_3 + \mathbf{x}_1\mathbf{x}_3 + \mathbf{x}_1\mathbf{x}_2\mathbf{x}_3 + \epsilon$
- Fit a correctly specified linear model (interactions with $\mathbf{x}_4$ are excluded)
- 30 reps, measure interaction strength between $\mathbf{x}_2$ and all other 3 features

Which methods are sensitive to changes in main effect sizes or feature correlations?

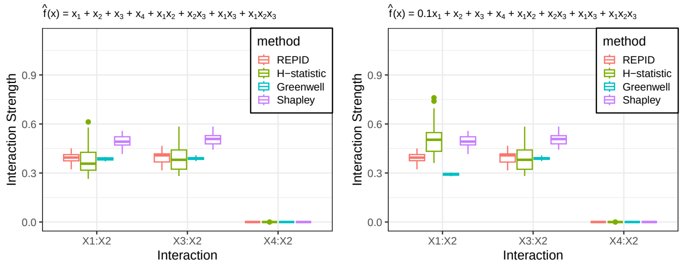
Pitfall	REPID	H-Statistic	Greenwell	SHAP
sensitive to changes of main effect	No	Yes	Yes	No
sensitive to changes of correlation between $\mathbf{x}_j$ and other features	No	Yes	No	Yes

OUTPERFORMING SOTA



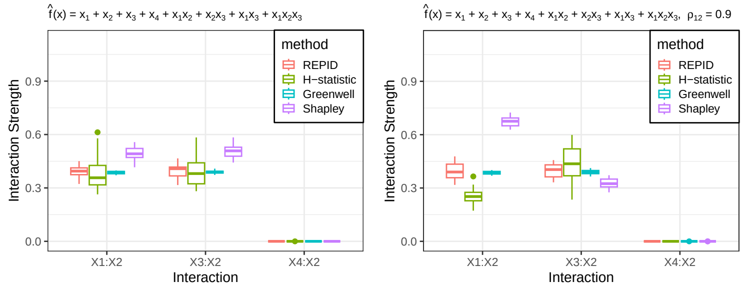
- **Left (initial setting)**: Interaction strength of $\mathbf{x}_1$:$\mathbf{x}_2$ and $\mathbf{x}_3$:$\mathbf{x}_2$ similar; $\mathbf{x}_4$:$\mathbf{x}_2$ no interaction
- **Right**: Set main effect $\beta_1 = 0.1$
  - **Expectation**: Interaction strengths should not change
  - **Fail**: H-statistic ($\mathbf{x}_1$:$\mathbf{x}_2$ > $\mathbf{x}_3$:$\mathbf{x}_2$) and Greenwell ($\mathbf{x}_1$:$\mathbf{x}_2$ < $\mathbf{x}_3$:$\mathbf{x}_2$)

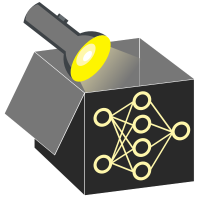
OUTPERFORMING SOTA



- **Left (initial setting)**: Interaction strength of $\mathbf{x}_1$:$\mathbf{x}_2$ and $\mathbf{x}_3$:$\mathbf{x}_2$ similar; $\mathbf{x}_4$:$\mathbf{x}_2$ no interaction
- **Right**: Set main effect $\beta_1 = 0.1$
  - **Expectation**: Interaction strengths should not change
  - **Fail**: H-statistic ($\mathbf{x}_1$:$\mathbf{x}_2$ > $\mathbf{x}_3$:$\mathbf{x}_2$) and Greenwell ($\mathbf{x}_1$:$\mathbf{x}_2$ < $\mathbf{x}_3$:$\mathbf{x}_2$)

OUTPERFORMING SOTA

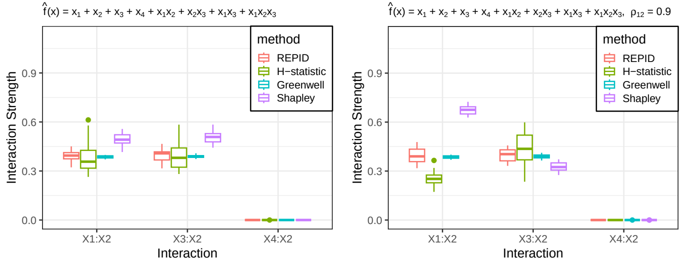


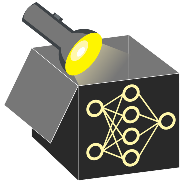


- **Left (initial setting)**: Interaction strength of $\mathbf{x}_1$:$\mathbf{x}_2$ and $\mathbf{x}_3$:$\mathbf{x}_2$ similar; $\mathbf{x}_4$:$\mathbf{x}_2$ no interaction
- **Right**: Increase correlation $\rho(\mathbf{x}_1, \mathbf{x}_2) = 0.9$
  - **Expectation**: Interaction strengths should not change
  - **Fail**: H-statistic ($\mathbf{x}_1$:$\mathbf{x}_2$ < $\mathbf{x}_3$:$\mathbf{x}_2$) and Shapley ($\mathbf{x}_1$:$\mathbf{x}_2$ > $\mathbf{x}_3$:$\mathbf{x}_2$)

$\rightarrow$ **REPID is the only method which always leads to correct rankings for these settings**

OUTPERFORMING SOTA



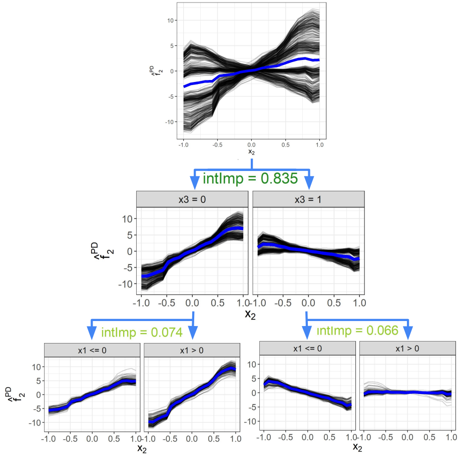


- **Left (initial setting)**: Interaction strength of $\mathbf{x}_1$:$\mathbf{x}_2$ and $\mathbf{x}_3$:$\mathbf{x}_2$ similar; $\mathbf{x}_4$:$\mathbf{x}_2$ no interaction
- **Right**: Increase correlation $\rho(\mathbf{x}_1, \mathbf{x}_2) = 0.9$
  - **Expectation**: Interaction strengths should not change
  - **Fail**: H-statistic ($\mathbf{x}_1$:$\mathbf{x}_2$ < $\mathbf{x}_3$:$\mathbf{x}_2$) and Shapley ($\mathbf{x}_1$:$\mathbf{x}_2$ > $\mathbf{x}_3$:$\mathbf{x}_2$)

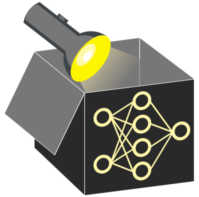
$\rightarrow$ **REPID is the only method which always leads to correct rankings for these settings**

LIMITATIONS OF REPID

1) Restricted to one feature of interest
$\Rightarrow$ Different regions for different features

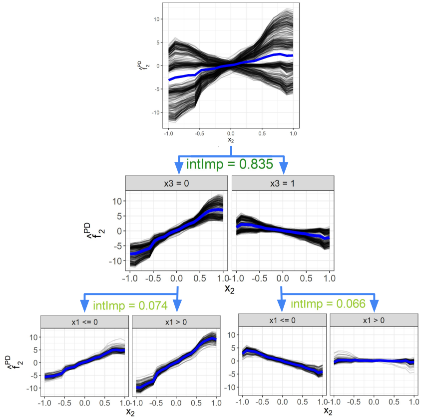


$\hat{f}_2^{PD}(X_2) \approx 8X_2 \quad \approx 16X_2 \quad \approx -8X_2 \quad \approx 0$

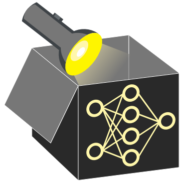


LIMITATIONS OF REPID

1) Restricted to one feature of interest
$\Rightarrow$ Different regions for different features



$\hat{f}_2^{PD}(X_2) \approx 8X_2 \quad \approx 16X_2 \quad \approx -8X_2 \quad \approx 0$



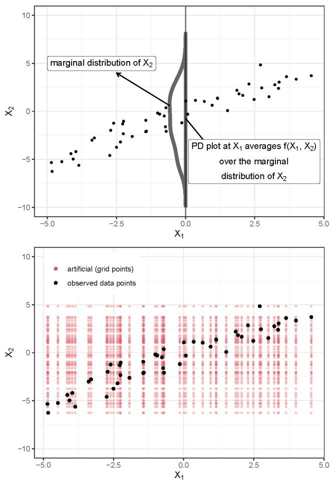
Interpretable Machine Learning – 7 / 8

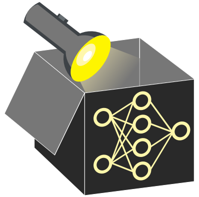
LIMITATIONS OF REPID

1) Restricted to one feature of interest
⇒ Different regions for different features

2) Restricted to PD (global) and ICE (local) as feature effect methods
⇒ Inherits extrapolation problem (unlikely combinations of feature values)

⇝ Follow-up GADGET [under review]



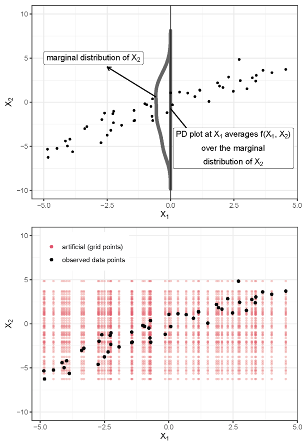


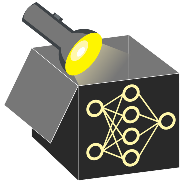
LIMITATIONS OF REPID

1) Restricted to one feature of interest
⇒ Different regions for different features

2) Restricted to PD (global) and ICE (local) as feature effect methods
⇒ Inherits extrapolation problem (unlikely combinations of feature values)

⇝ Follow-up GADGET [under review]





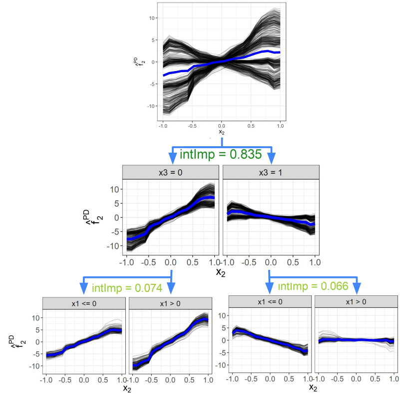
CONCLUSION

Summary of Contributions (REPID):

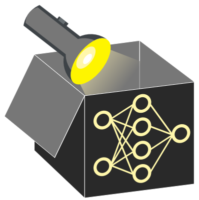
- Regional effects in interpretable regions
- Additive decomposition of feature effect
- Quantify feature interactions
- Outperforms SOTA interaction indices

Summary of Contributions (GADGET):

- Unique regions for multiple features
- Additive decomposition of prediction function
- Extension to ALE and Shapley Dependence
- Test to identify significant interactions



$\hat{f}_2^{PD}(X_2) \approx 8X_2 \quad \approx 16X_2 \quad \approx -8X_2 \quad \approx 0$



Further Directions:
Pruning, GADGET as a predictor, comparing regions across models, efficient implementation, more efficient testing and splitting approach, . . .

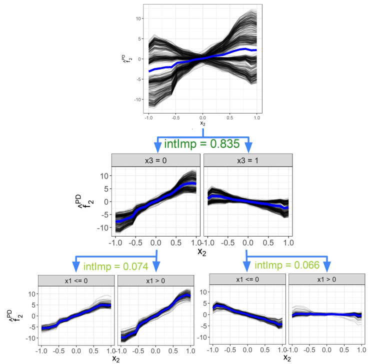
CONCLUSION

Summary of Contributions (REPID):

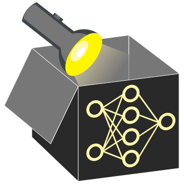
- Regional effects in interpretable regions
- Additive decomp. of feature effect
- Quantify feature interactions
- Outperforms SOTA interaction indices

Summary of Contributions (GADGET):

- Unique regions for multiple features
- Additive decomp. of prediction function
- Extension to ALE and Shapley Dependence
- Test to identify significant interactions



$\hat{f}_2^{PD}(X_2) \approx 8X_2 \quad \approx 16X_2 \quad \approx -8X_2 \quad \approx 0$



Further Directions:
Pruning, GADGET as a predictor, comparing regions across models, efficient implementation, more efficient testing and splitting approach, . . .