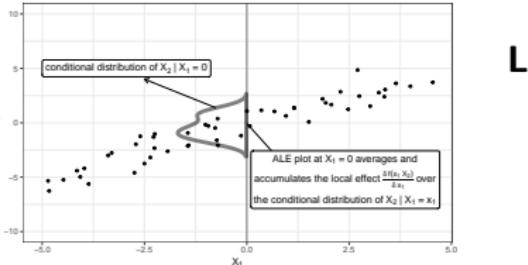


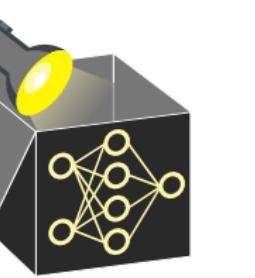
Interpretable Machine Learning

Regional Effects



Learning goals

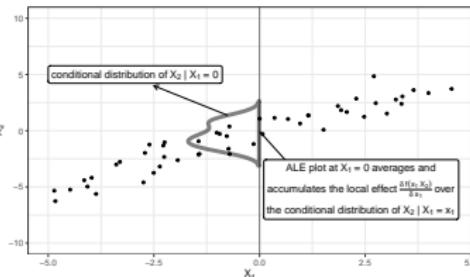
- Difference between feature effects and feature interactions
- REPID



Interpretable Machine Learning

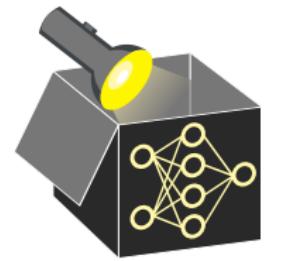
Regional Effects

REPID



Learning goals

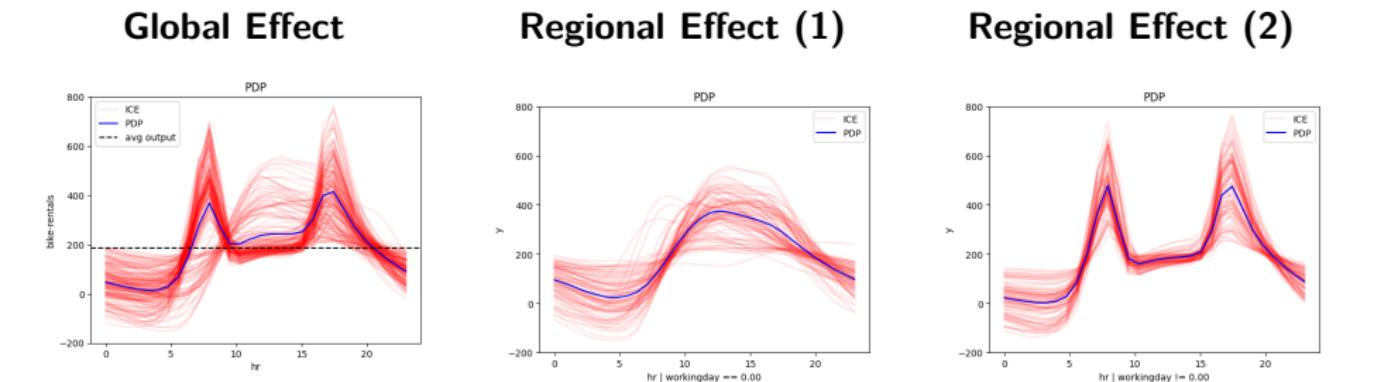
- Difference between feature effects and feature interactions
- REPID



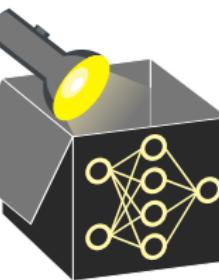
WHY REGIONAL EXPLANATIONS?

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).



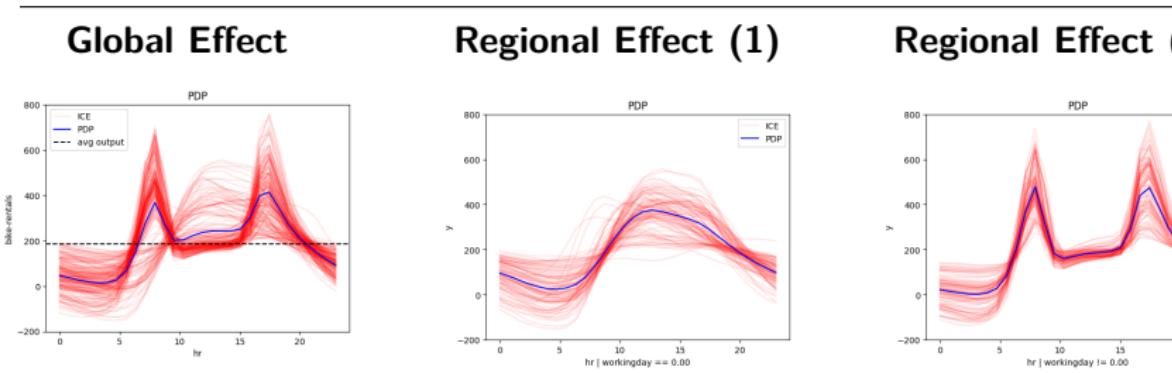
- Splitting by the "workingday" revealed 2 different patterns that we're clashed together in the initial plot
- $\cup_i(\text{regional_explanation}_i) = \text{global_explanation}$
- $Fidelity(\text{regional_explanation}_i) > Fidelity(\text{global_explanation})$



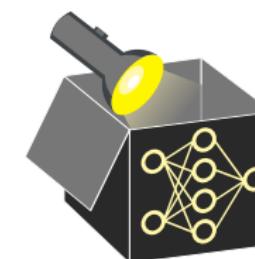
WHY REGIONAL EXPLANATIONS?

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).



- Splitting by the "workingday" revealed 2 different patterns that we're clashed together in the initial plot
- $\cup_i(\text{regional_explanation}_i) = \text{global_explanation}$
- $Fidelity(\text{regional_explanation}_i) > Fidelity(\text{global_explanation})$

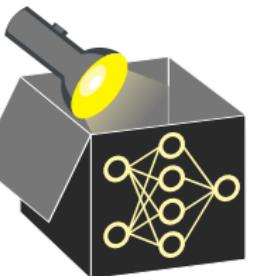


ICE CURVE: LOCAL FEATURE EFFECTS

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)

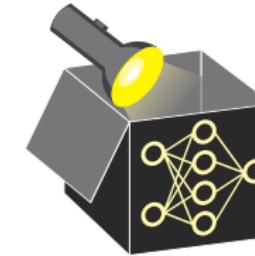


ICE CURVE: LOCAL FEATURE EFFECTS

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)



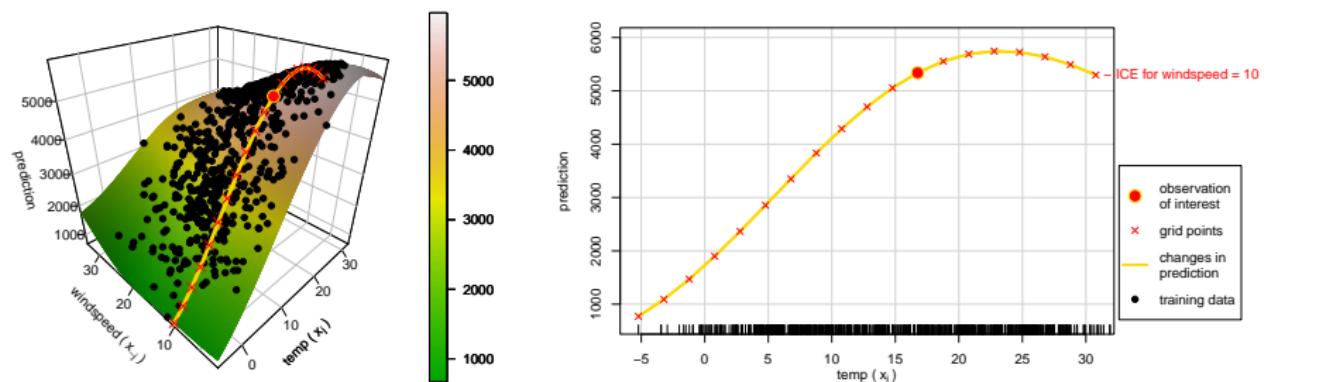
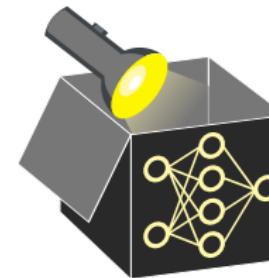
ICE CURVE: LOCAL FEATURE EFFECTS

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**



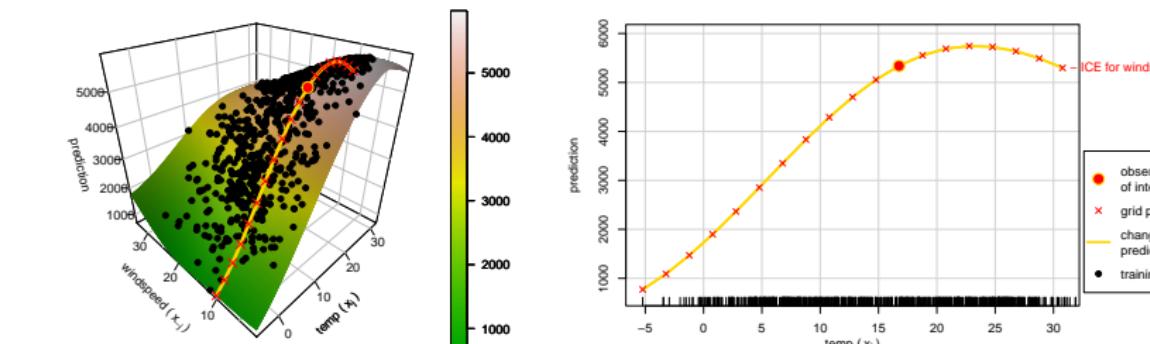
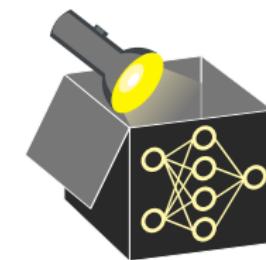
ICE CURVE: LOCAL FEATURE EFFECTS

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

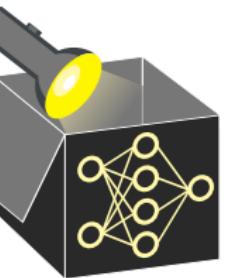
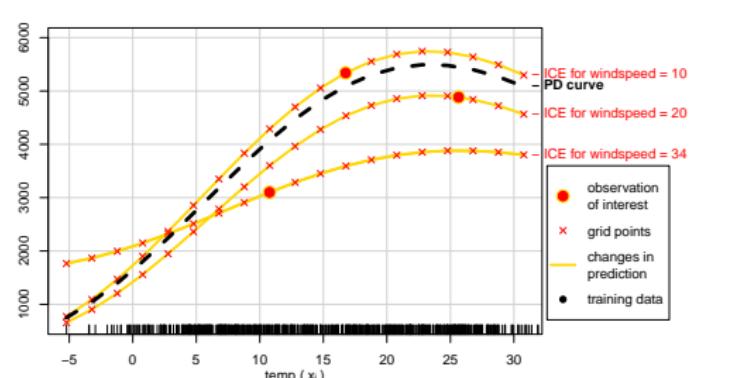
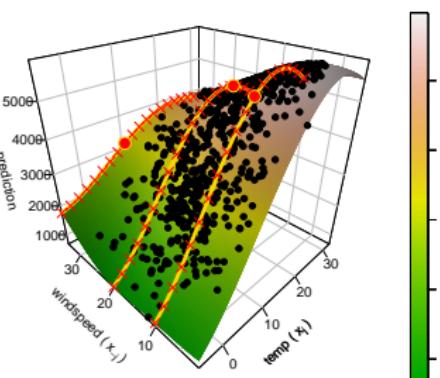
Question: How do changes of feature values affect model prediction **on average**?

- **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

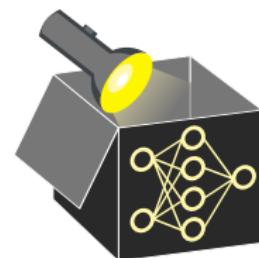
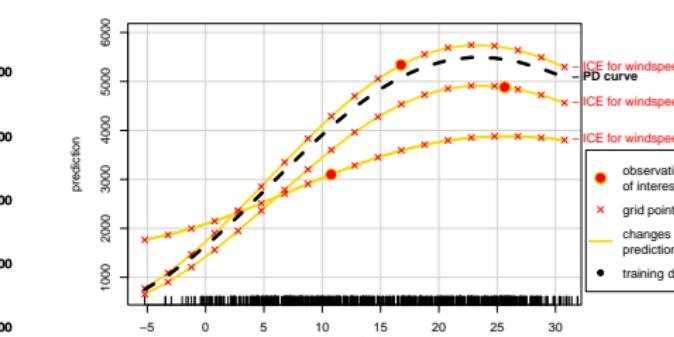
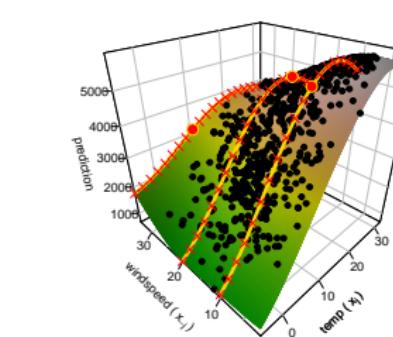
Question: How do changes of feat values affect model prediction **on avg.**?

- **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

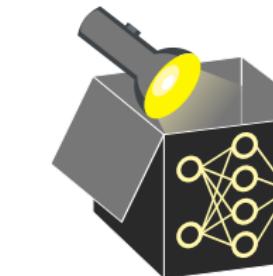
$$\hat{f}_j^{PD}(\tilde{x}_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}^{(i)})$$



FEATURE INTERACTIONS

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

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

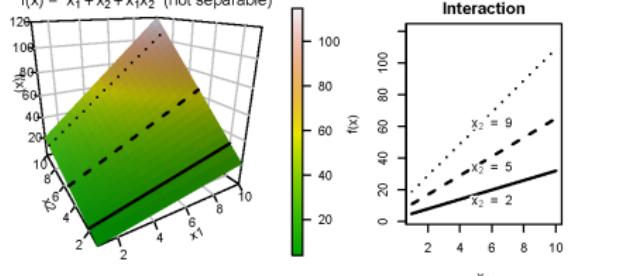


Friedman and Popescu (2008):

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

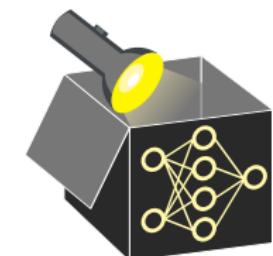
$$f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2 \neq g(x_1) + g(x_2)$$



FEATURE INTERACTIONS

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

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

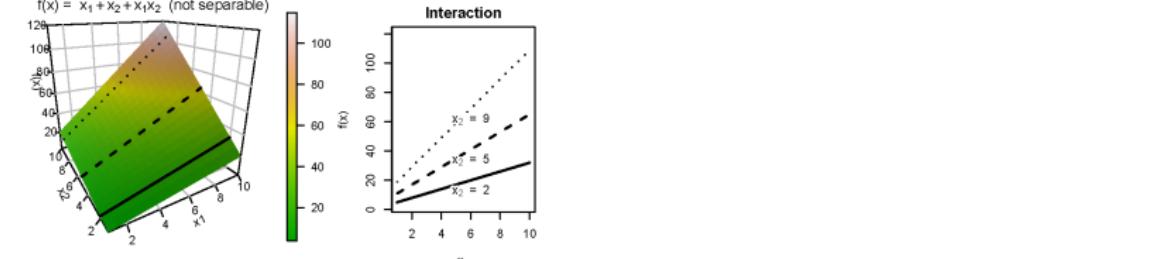


Friedman and Popescu (2008):

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

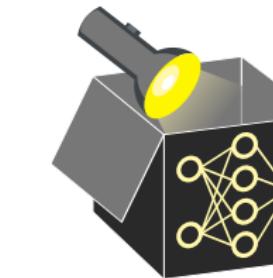
$$f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2 \neq g(x_1) + g(x_2)$$



FEATURE INTERACTIONS

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

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

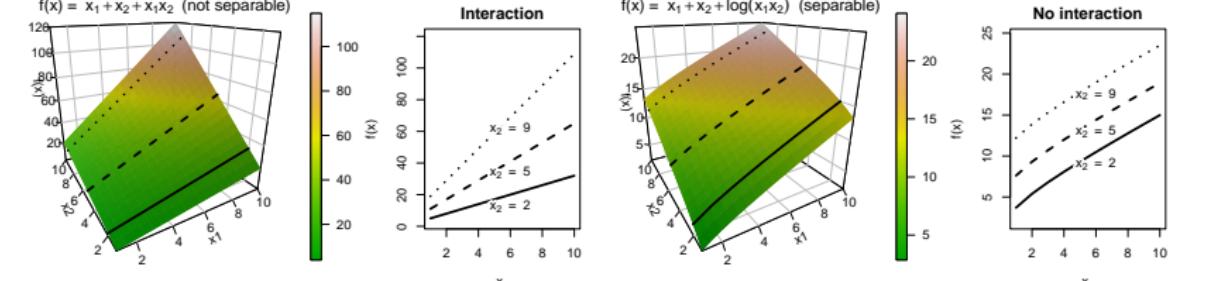


Friedman and Popescu (2008):

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

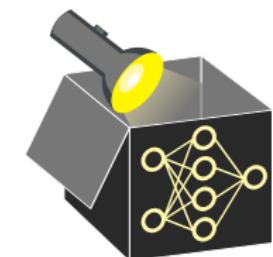
$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = (x_1 + \log(x_1)) + (x_2 + \log(x_2)) = g_1(x_1) + g_2(x_2)$$



FEATURE INTERACTIONS

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

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

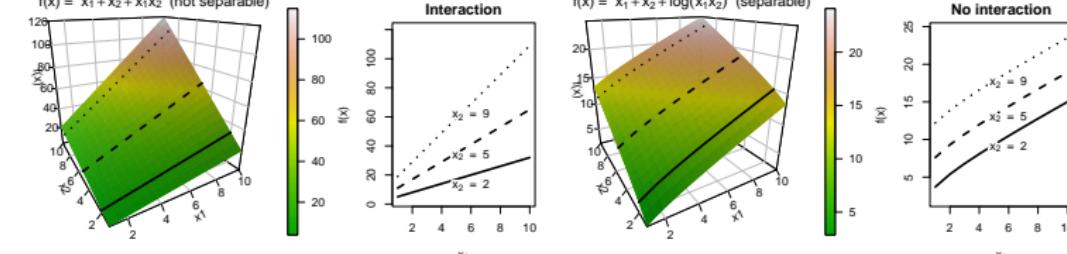


Friedman and Popescu (2008):

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

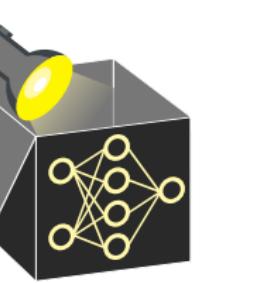
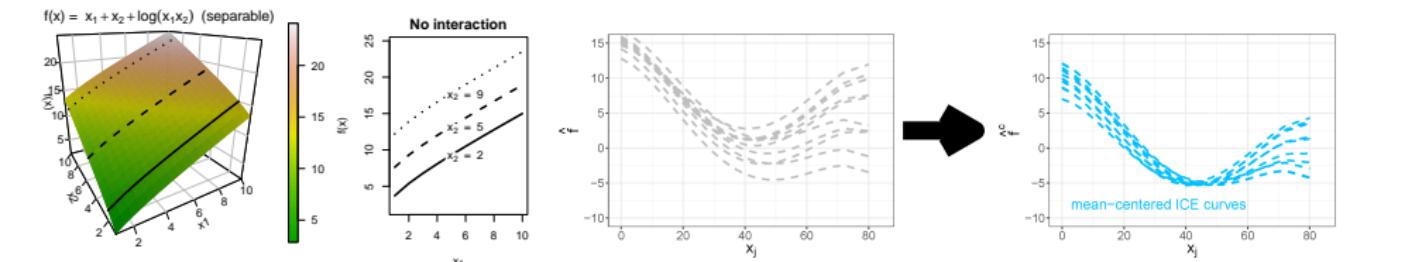
$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = (x_1 + \log(x_1)) + (x_2 + \log(x_2)) = g_1(x_1) + g_2(x_2)$$



Recall: Different shapes of ICE curves indicate interactions (ignore vertical shifts)
 ⇒ Focus on shape differences of **mean-centered ICE curves**.

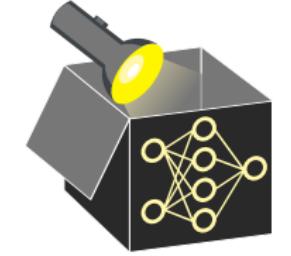
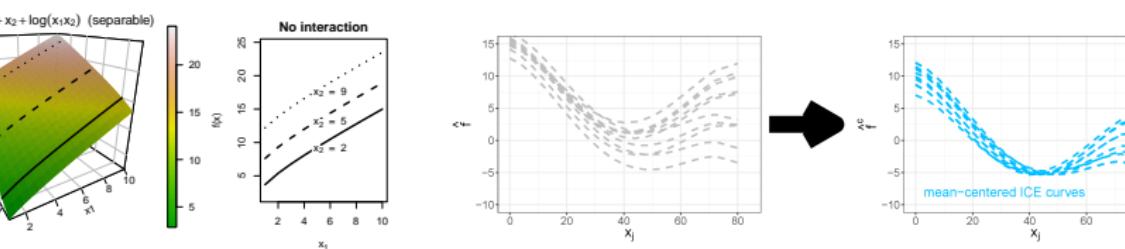
Mean-centered ICE curve for obs. \mathbf{x} evaluated at m grid points $\tilde{\mathbf{x}}_j^{(1)}, \dots, \tilde{\mathbf{x}}_j^{(m)}$ is:

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



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

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



REGIONAL EFFECTS - SYNTHETIC EXAMPLE

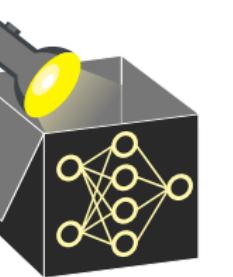
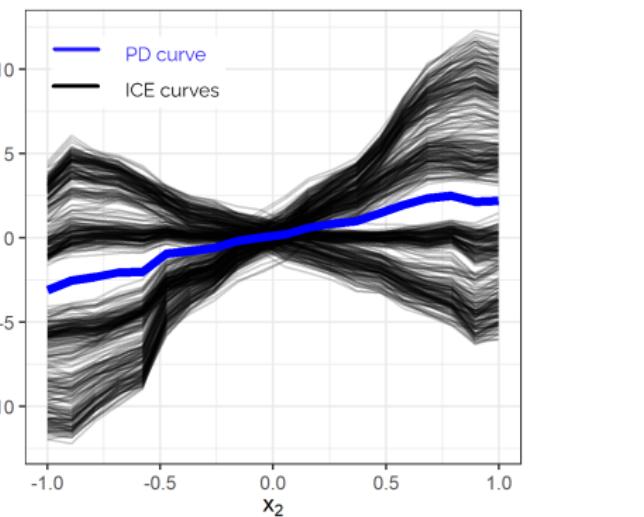
Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

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

~~ Model: Random forest

Problem:

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



REGIONAL EFFECTS - SYNTHETIC EXAMPLE

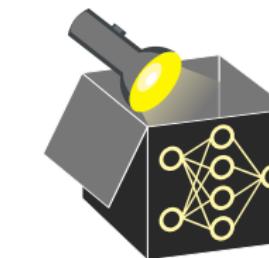
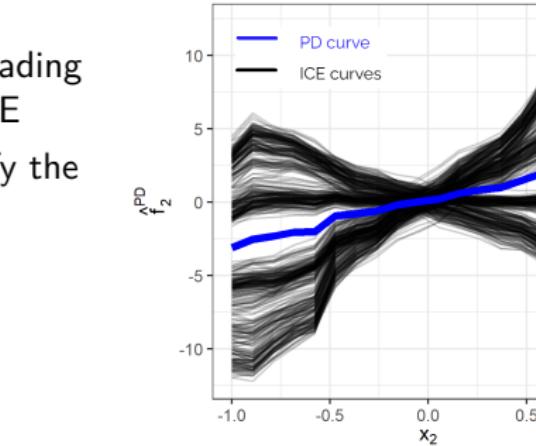
Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

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

~~ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions ~~ 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)

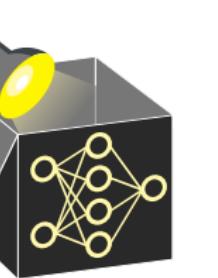
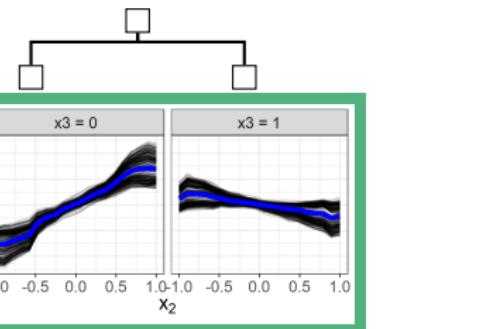
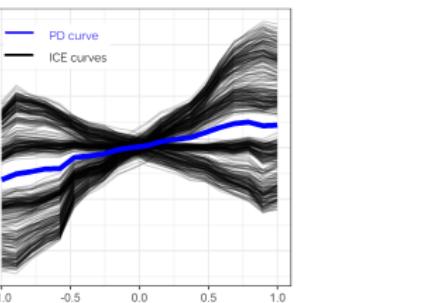
~~ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(X_1>0)} + 16X_2\mathbb{1}_{(X_3=0)} + \epsilon$

~~ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions ~~ 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)

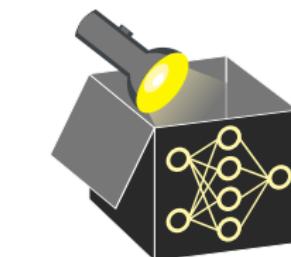
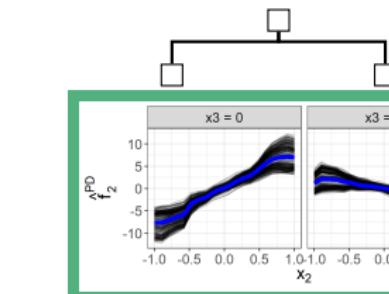
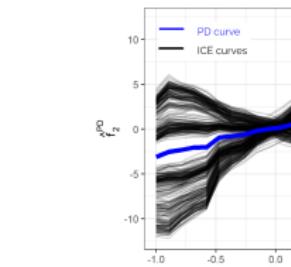
~~ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(X_1>0)} + 16X_2\mathbb{1}_{(X_3=0)} + \epsilon$

~~ Model: Random forest

Problem:

- PD curve of X_2 is misleading due to interactions ~~ 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)

~~ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(x_1>0)} + 16X_2\mathbb{1}_{(x_3=0)} + \epsilon$

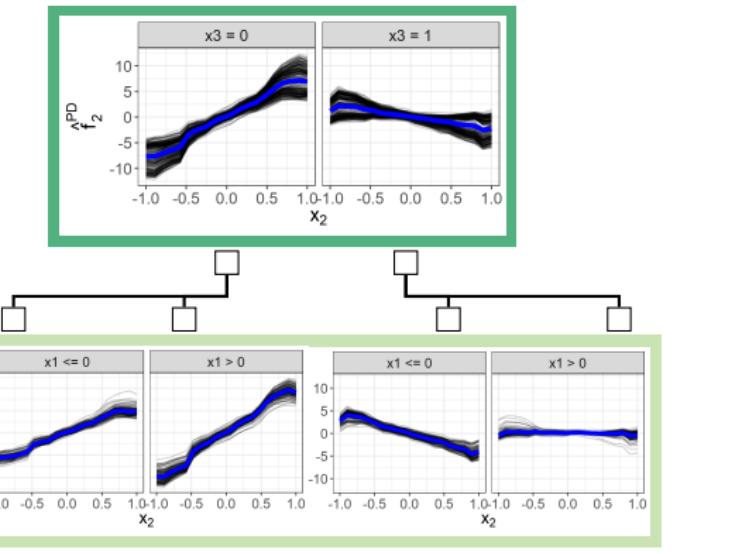
~~ Model: Random forest

Problem:

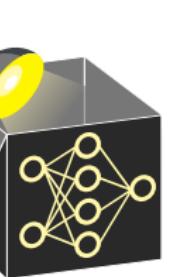
- PD curve of X_2 is misleading due to interactions ~~ ICE
- ICE curves do not identify the interacting features

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

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



\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)

~~ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(x_1>0)} + 16X_2\mathbb{1}_{(x_3=0)} + \epsilon$

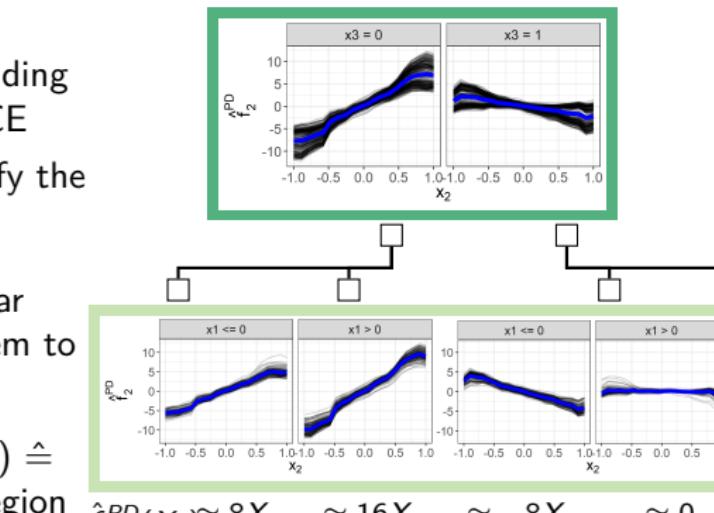
~~ Model: Random forest

Problem:

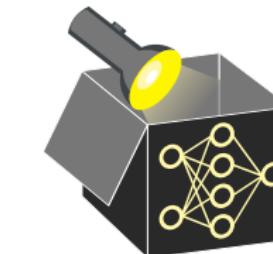
- PD curve of X_2 is misleading due to interactions ~~ ICE
- ICE curves do not identify the interacting features

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

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



\Rightarrow Additive decomposition of global feat effect

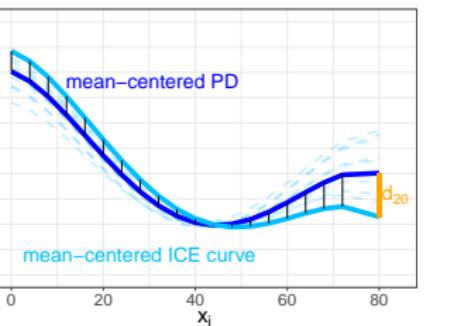


REGIONAL EFFECTS - DETAILS

Question: How to split the curves into regions?

Define risk as L2 loss of mean-centered ICE curves:

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



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

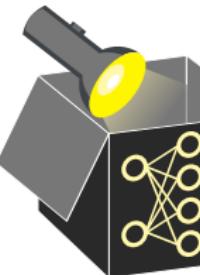
$$\hat{f}_{j|\mathcal{N}}^{PD,c}(\tilde{x}_j) = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} \hat{f}^c(\tilde{x}_j, x_{-j})$$

- ⇒ Measures interaction-related heterogeneity (variance) of ICE curves in \mathcal{N}
- ⇒ Recursive partitioning (CART): Find best feature-split combination that solves

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

- $\mathcal{N}_{left} = \{x \in \mathcal{N} | x_z \leq t\}$
- $\mathcal{N}_{right} = \{x \in \mathcal{N} | x_z > t\}$
- 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)?

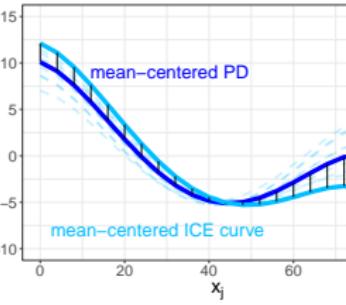


REGIONAL EFFECTS - DETAILS

Question: How to split curves into regions?

Define risk as L2 loss of mean-centered ICE curves:

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



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

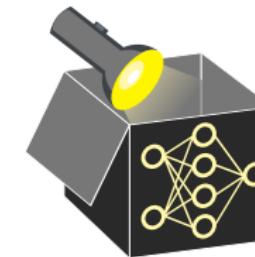
$$\hat{f}_{j|\mathcal{N}}^{PD,c}(\tilde{x}_j) = \frac{1}{|\mathcal{N}|} \sum_{x \in \mathcal{N}} \hat{f}^c(\tilde{x}_j, x_{-j})$$

- ⇒ Measures interaction-related heterogeneity (variance) of ICE curves in \mathcal{N}
- ⇒ Recursive partitioning (CART): Find best feat-split combo that solves

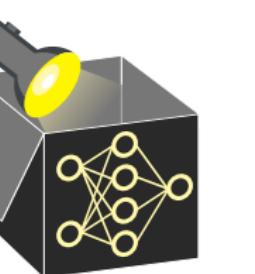
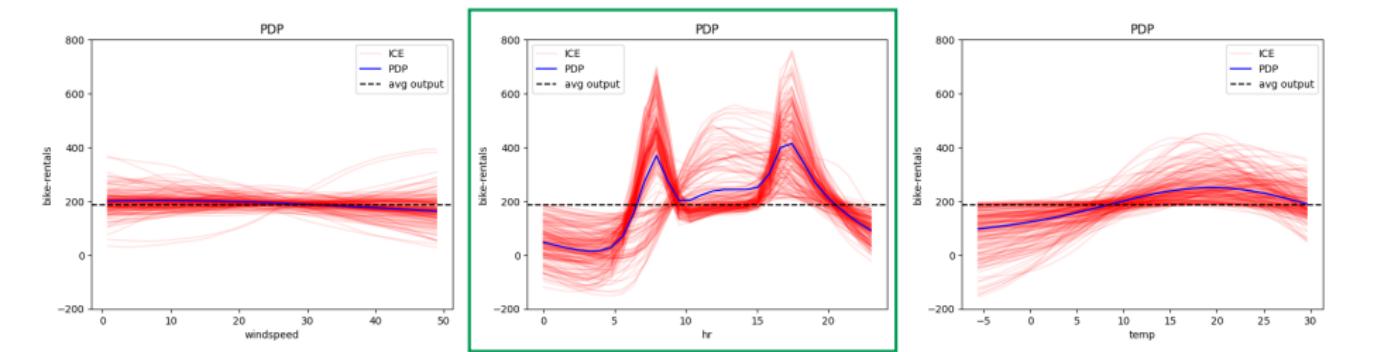
$$\arg \min_{z,t} \mathcal{R}_j(\mathcal{N}_{left}) + \mathcal{R}_j(\mathcal{N}_{right})$$

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

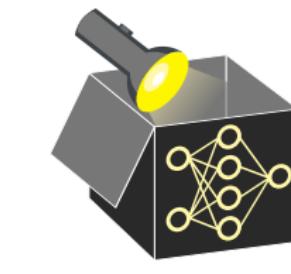
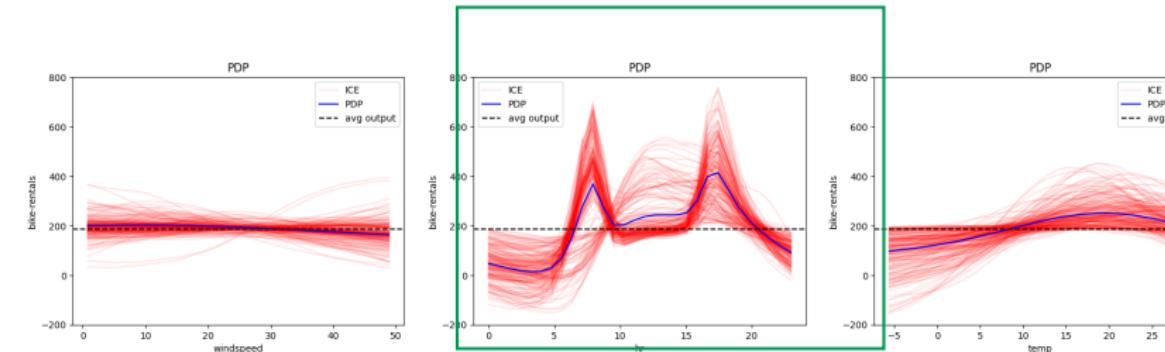
Intuition: Is another feature x_z responsible for the heterogeneity (measured by \mathcal{R}_j)?



REGIONAL EFFECT PLOTS - REAL EXAMPLE



REGIONAL EFFECT PLOTS - REAL EXAMPLE

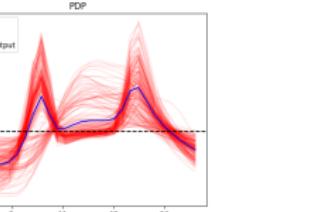


- Identify feature with highly heterogeneous local effects
 - ~~ **hour: Most important and highly heterogeneous feature (highest variance)**
- Find regions in feature space where this heterogeneity is minimal
 - ~~ Partition feature space using CART to minimize variance of mean-centered ICE curves within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

- 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

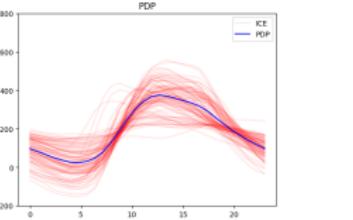
REGIONAL EFFECT PLOTS - REAL EXAMPLE

Regional effects of hour



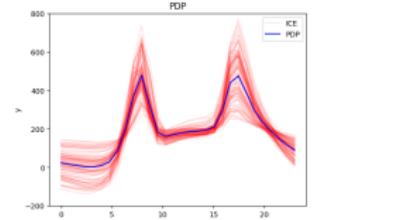
workingday=False /

\ workingday=True



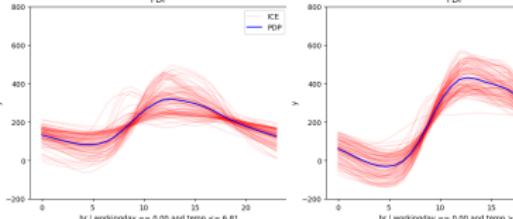
temp=cold /

\ temp=hot



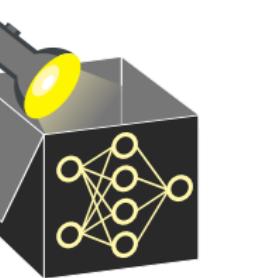
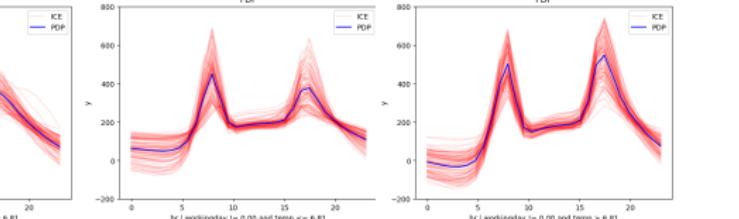
temp=cold /

\ temp=hot



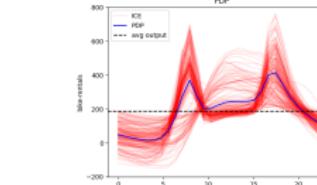
temp=cold /

\ temp=hot



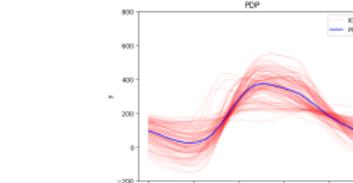
REGIONAL EFFECT PLOTS - REAL EXAMPLE

Regional effects of hour



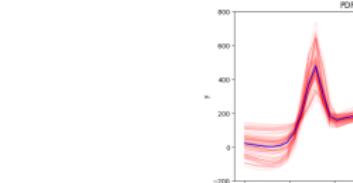
workingday=False /

\ workingday=True



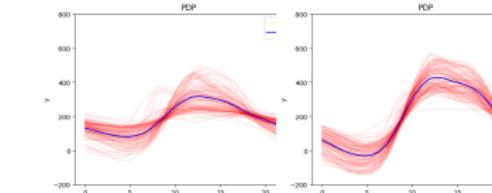
temp=cold /

\ temp=hot



temp=cold /

\ temp=hot



temp=cold /

\ temp=hot

