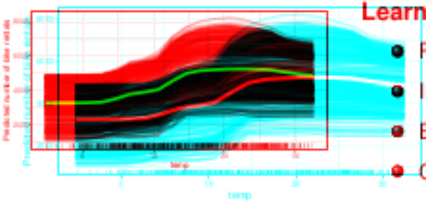
Interpretable Machine Learning

Partial Dependence (PD) plot



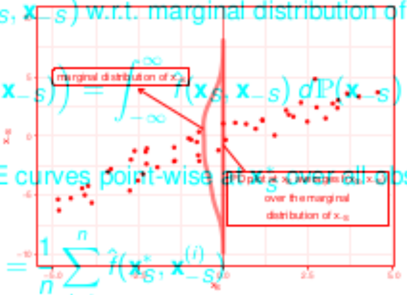
Learning goals

- PD plots and relation to ICE plots
- Interpretation of PDP
- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

PARTIAL DEPENDENCE (PD) ▸ Friedman (2001)

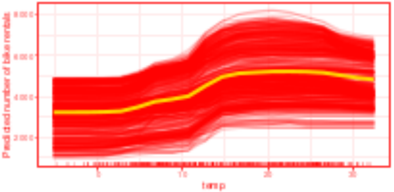
Definition: PD function is expectation of $\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})$ w.r.t. marginal distribution of features $\mathbf{x}_{-S}$:

$$f_{S,PD}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\right) = \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\, d\mathbb{P}(\mathbf{x}_{-S})$$



Estimation: For a grid value $\mathbf{x}_S^*$, average ICE curves point-wise at $\mathbf{x}_S^*$ over all observed $\mathbf{x}_{-S}^{(i)}$:

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{S,ICE}^{(i)}(\mathbf{x}_S^*)$$



Definition: PD function is expectation of $\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})$ w.r.t. marginal distribution of features $\mathbf{x}_{-S}$:

$$f_{S,PD}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\right) = \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\, d\mathbb{P}(\mathbf{x}_{-S})$$

Estimation: For a grid value $\mathbf{x}_S^*$, average ICE curves point-wise at $\mathbf{x}_S^*$ over all observed $\mathbf{x}_{-S}^{(i)}$:

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$$

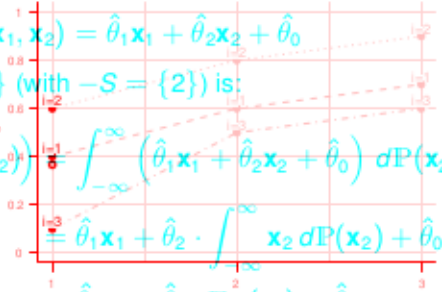
PARTIAL DEPENDENCE

i	x_S	x_2	x_3	$\hat f$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	x_S	x_2	x_3	$\hat f$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

i	x_S	x_2	x_3	$\hat f$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$\frac{1}{3}\sum_{i=1}^{3} \hat f$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



Estimate PD function by **point-wise** average of ICE curves at grid value $\boxed{\mathbf{x}_S^* = x_1^* = 1}$:

$$\hat f_{1,PD}(x_1^*) = \frac{1}{n}\sum_{i=1}^{n} \hat f(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

PARTIAL DEPENDENCE FOR LINEAR MODEL

Assume a linear regression model with two features:

$$\hat f(\mathbf{x}) = \hat f(\mathbf{x}_1, \mathbf{x}_2) = \hat\theta_1 \mathbf{x}_1 + \hat\theta_2 \mathbf{x}_2 + \hat\theta_0$$

PD function for feature of interest $S = \{1\}$ (with $-S = \{2\}$) is:

$$\begin{aligned} f_{1,PD}(\mathbf{x}_1) &= \mathbb{E}_{\mathbf{x}_2}\left(\hat f(\mathbf{x}_1, \mathbf{x}_2)\right) = \int_{-\infty}^{\infty} \left(\hat\theta_1 \mathbf{x}_1 + \hat\theta_2 \mathbf{x}_2 + \hat\theta_0\right) d\mathbb{P}(\mathbf{x}_2) \\ &= \hat\theta_1 \mathbf{x}_1 + \hat\theta_2 \cdot \int_{-\infty}^{\infty} \mathbf{x}_2 \, d\mathbb{P}(\mathbf{x}_2) + \hat\theta_0 \\ &= \hat\theta_1 \mathbf{x}_1 + \underbrace{\hat\theta_2 \cdot \mathbb{E}_{\mathbf{x}_2}(\mathbf{x}_2) + \hat\theta_0}_{:=const} \end{aligned}$$

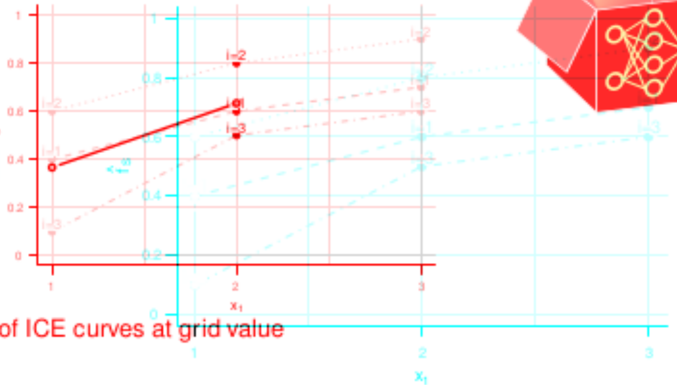
$\Rightarrow$ PD plot visualizes the function $f_{1,PD}(\mathbf{x}_1) = \hat\theta_1 \mathbf{x}_1 + const$ ($\widehat{=}$ feature effect of $\mathbf{x}_1$).

i	x_S	x_2	x_3	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	x_S	x_2	x_3	$\hat{f}$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

i	x_S	x_2	x_3	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



Estimate PD function by **point-wise** average of ICE curves at grid value $\boxed{\mathbf{x}_S^* = x_1^* = 2}$:

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

Estimate PD function by point-wise average of ICE curves at grid value $\mathbf{x}_S^* = x_1^* = 1$:

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

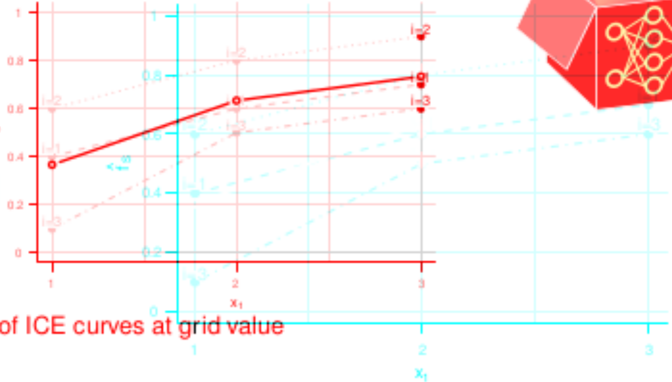
PARTIAL DEPENDENCE

i	x_S	x_2	x_3	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	x_S	x_2	x_3	$\hat{f}$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

i	x_S	x_2	x_3	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



Estimate PD function by **point-wise** average of ICE curves at grid value $\mathbf{x}_S^* = x_1^* = 3$:

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

EXAMPLE: PD FOR LINEAR MODEL

Assume a linear regression model with two features:

$$\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}_1, \mathbf{x}_2) = \hat{\theta}_1 \mathbf{x}_1 + \hat{\theta}_2 \mathbf{x}_2 + \hat{\theta}_0$$

PD function for feature of interest $S = \{1\}$ (with $-S = \{2\}$) is:

$$\begin{aligned} f_{1,PD}(\mathbf{x}_1) = \mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(\mathbf{x}_1, \mathbf{x}_2)\right) &= \int_{-\infty}^{\infty} \left(\hat{\theta}_1 \mathbf{x}_1 + \hat{\theta}_2 \mathbf{x}_2 + \hat{\theta}_0\right) d\mathbb{P}(\mathbf{x}_2) \\ &= \hat{\theta}_1 \mathbf{x}_1 + \hat{\theta}_2 \cdot \int_{-\infty}^{\infty} \mathbf{x}_2 \, d\mathbb{P}(\mathbf{x}_2) + \hat{\theta}_0 \\ &= \hat{\theta}_1 \mathbf{x}_1 + \underbrace{\hat{\theta}_2 \cdot \mathbb{E}_{\mathbf{x}_2}(\mathbf{x}_2) + \hat{\theta}_0}_{:=const} \end{aligned}$$

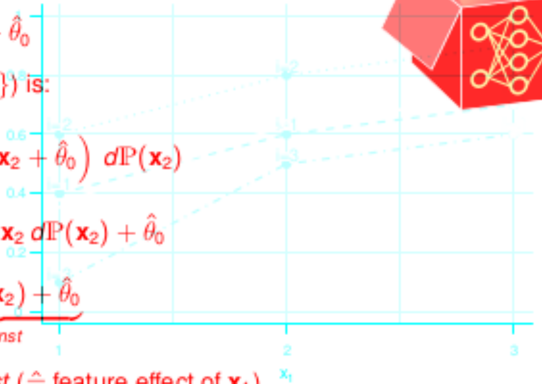
$\Rightarrow$ PD plot visualizes the function $f_{1,PD}(\mathbf{x}_1) = \hat{\theta}_1 \mathbf{x}_1 + const$ ($\hat{=}$ feature effect of $\mathbf{x}_1$).

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$\frac{1}{3}\sum_{i=1}^{3} \hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



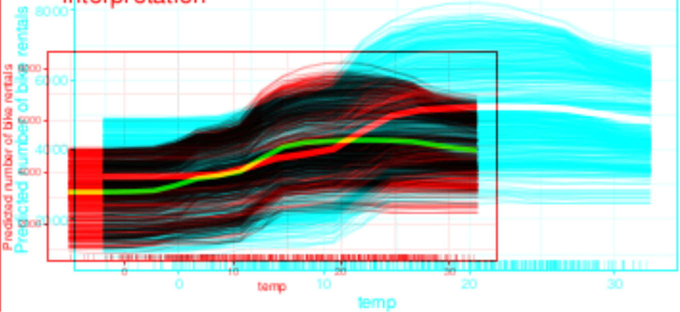
Estimate PD function by **point-wise** average of ICE curves at grid value $\mathbf{x}_S^* = x_1^* = 3$:

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

INTERPRETATION: PD AND ICE

If feature varies:

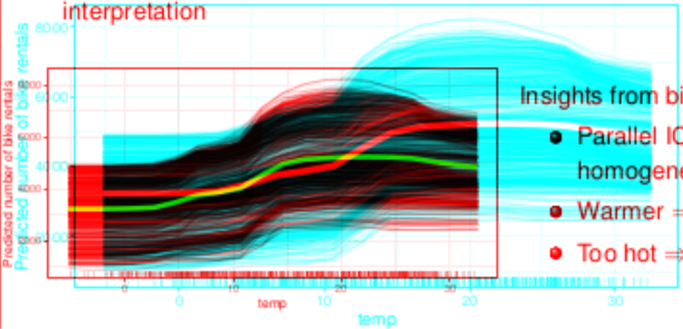
- **ICE:** How does **prediction of individual observation** change? ⇒ **local** interpretation
- **PD:** How does **average effect / expected prediction** change? ⇒ **global** interpretation



INTERPRETATION: PD AND ICE

If feature varies:

- **ICE:** How does **prediction of individual observation** change? ⇒ **local** interpretation
- **PD:** How does **average effect / expected prediction** change? ⇒ **global** interpretation

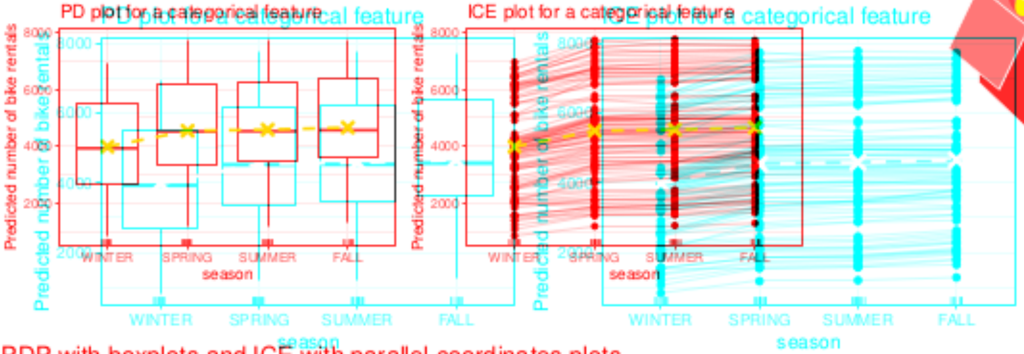


Insights from bike sharing data:

- Parallel ICE curves = homogeneous effect
- Warmer ⇒ more rented bikes
- Too hot ⇒ slightly less bikes

Interpretable Machine Learning – 4 / 10

INTERPRETATION: CATEGORICAL FEATURES

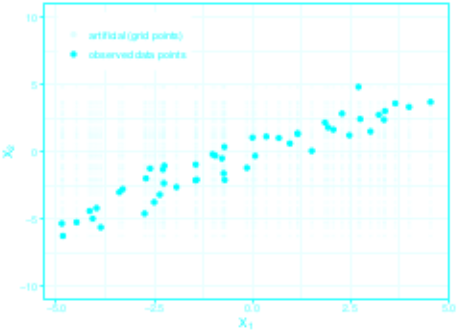


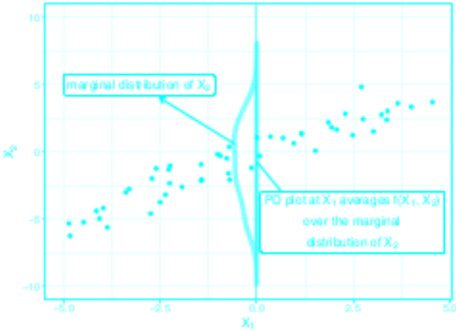
- PDP with boxplots and ICE with parallel coordinates plots
- NB: Categories can be unordered, if so, rather compare pairwise

- PDP with boxplots and ICE with parallel coordinates plots
- NB: Categories can be unordered, if so, rather compare pairwise

COMMENTS ON EXTRAPOLATION

Extrapolation can cause issues in regions with few observations or if features are correlated





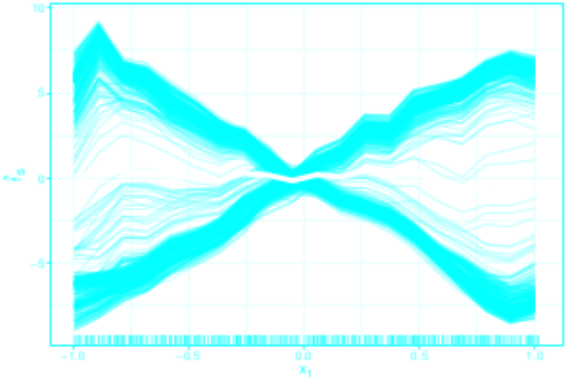
- **Example:** Features x_1 and x_2 are strongly correlated
- **Black points:** Observed points of the original data
- Grid points used to calculate the ICE and PD curves (several unrealistic values)

$\Rightarrow$ PD plot at $x_1 = 0$ averages predictions over the whole marginal distribution of feature x_2

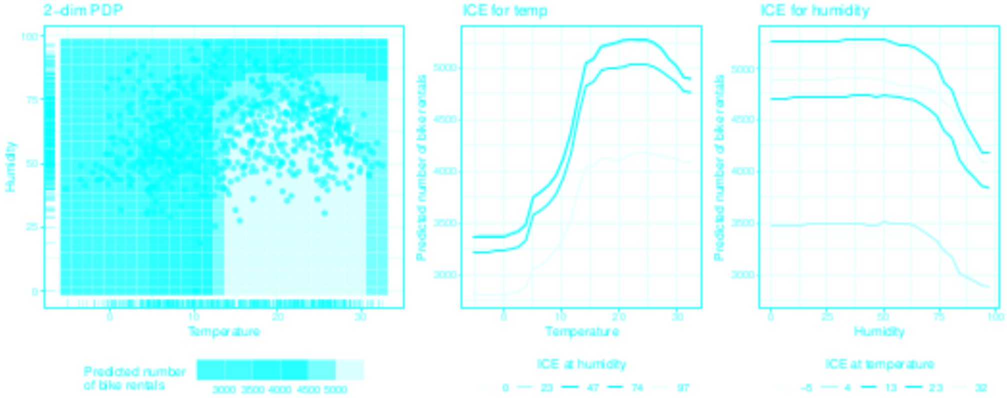
$\Rightarrow$ May be problematic if model behaves strange outside training distribution

COMMENTS ON INTERACTIONS

- PD plots: averaging of ICE curves might **obfuscate** heterogeneous effects and interactions
  $\Rightarrow$ Ideally plot ICE curves and PD plots together to uncover this fact
  $\Rightarrow$ Different shapes of ICE curves suggest interaction (but does not tell with which feature)



COMMENTS ON INTERACTIONS - 2D PARTIAL DEPENDENCE



- Humidity and temperature interact with each other at high values (see shape difference)
  $\rightsquigarrow$ Shape of ICE curves at different horizontal and vertical slices varies (for high values)
- Low to medium humidity and high temperature $\Rightarrow$ many rented bikes

CENTERED ICE PLOT (C-ICE)

Issue: Difficult to identify heterogenous ICE curves if curves have different intercepts (are stacked)
Solution: Center ICE curves at fixed reference value $x' \sim \mathbb{P}(\mathbf{x}_S)$, often $x' = \min(\mathbf{x}_S)$
$\Rightarrow$ Easier to identify heterogenous shapes with c-ICE curves

$$\begin{aligned}\hat{f}^{(i)}_{S,cICE}(\mathbf{x}_S) &= \hat{f}(\mathbf{x}_S, \mathbf{x}^{(i)}_{-S}) - \hat{f}(x', \mathbf{x}^{(i)}_{-S}) \\ &= \hat{f}^{(i)}_S(\mathbf{x}_S) - \hat{f}^{(i)}_S(x')\end{aligned}$$

$\Rightarrow$ Visualize $\hat{f}^{(i)}_{S,cICE}(\mathbf{x}^*_S)$ vs. grid point $\mathbf{x}^*_S$

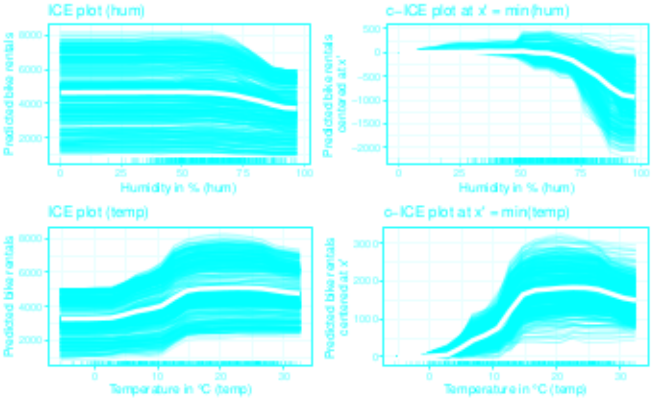
CENTERED ICE PLOT (C-ICE)

Issue: Difficult to identify heterogenous ICE curves if curves have different intercepts (are stacked)
Solution: Center ICE curves at fixed reference value $x' \sim \mathbb{P}(\mathbf{x}_S)$, often $x' = \min(\mathbf{x}_S)$
$\Rightarrow$ Easier to identify heterogenous shapes with c-ICE curves

$$\begin{aligned}\hat{f}^{(i)}_{S,cICE}(\mathbf{x}_S) &= \hat{f}(\mathbf{x}_S, \mathbf{x}^{(i)}_{-S}) - \hat{f}(x', \mathbf{x}^{(i)}_{-S}) \\ &= \hat{f}^{(i)}_S(\mathbf{x}_S) - \hat{f}^{(i)}_S(x')\end{aligned}$$

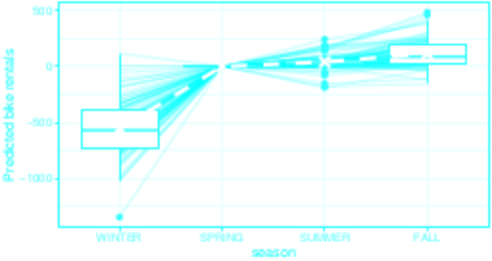
$\Rightarrow$ Visualize $\hat{f}^{(i)}_{S,cICE}(\mathbf{x}^*_S)$ vs. grid point $\mathbf{x}^*_S$

Interpretation (yellow curve in c-ICE):
On average, the number of bike rentals at $\sim$ 97 % humidity decreased by 1000 bikes compared to a humidity of 0 %



CENTERED ICE PLOT (C-ICE)

For categorical features, c-ICE plots can be interpreted as in LMs due to reference value



Interpretation:

- The reference category is $x' = \text{SPRING}$
- Golden crosses: Average number of bike rentals if we jump from SPRING to any other season $\Rightarrow$ Number of bike rentals drops by ~ 560 in `WINTER` and is slightly higher in `SUMMER` and `FALL` compared to `SPRING`