Imperial College London

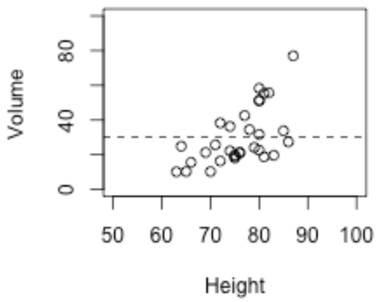
Ethics of Data Science – Part II

Measuring feature effects in ML models: ALEs

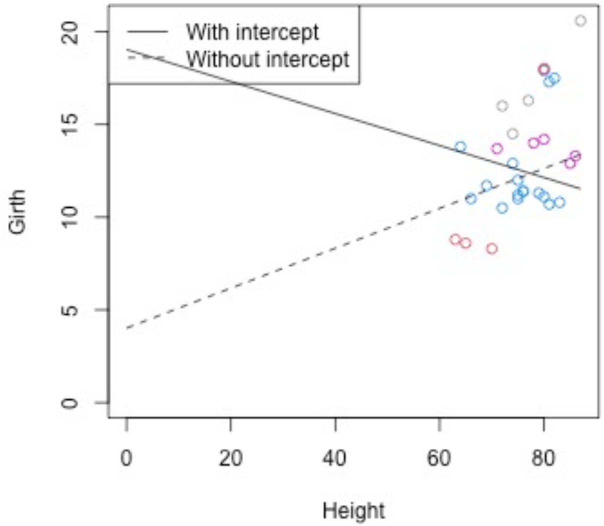
Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

How to fix the PDP

Consider the value of the PDP for linear regression at Height = 50. By virtue of linearity/additivity, it is the prediction for Height = 50 and Volume = 30 (average value of Volume across dataset).



Scatter plot revealing correlation between Volume and Height, with mean Volume indicated by dotted line.



How to fix the PDP

If we now consider a non-additive model, then each value of the PDP is the average of predictions most of which will be very unlikely (e.g., incredibly long, incredibly thin trees).

PDPs

Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4 (2020): 1059-1086.

How to fix the PDP

If we now consider a non-additive model, then each value of the PDP is the average of predictions most of which will be very unlikely (e.g., incredibly long, incredibly thin trees).

PDPs

One solution would be to rule out this "low density" datapoints. In other words, rather than averaging over the empirical distribution of the data, to average by conditioning on, say, H = 55, which would result in only considering the examples that are "near" those with H=55. This would make it impossible to tell which of two correlated features exercises the effect.

M-plots

Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4 (2020): 1059-1086.

How to fix the PDP

If we now consider a non-additive model, then each value of the PDP is the average of predictions most of which will be very unlikely (e.g., incredibly long, incredibly thin trees).

PDPs

One solution would be to rule out this “low density” datapoints. In other words, rather than averaging over the empirical distribution of the data, to average by conditioning on, say, H = 55, which would result in only considering the examples that are “near” those with H=55. This would make it impossible to tell which of two correlated features exercises the effect.

M-plots

Instead, recall our explanation from the last video: “the effect of a unit increase in ….” is the appropriate language to use in interpreting a regression coefficient. We’ll extend this.

ALEs

Apley, Daniel W., and Jingyu Zhu. “Visualizing the effects of predictor variables in black box supervised learning models.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4 (2020): 1059-1086.

Accumulated Local Effect

$$\hat{f}_{\text{PDP}}(x) = \mathbf{E}_{Z,W}[\hat{f}(x,Z,W)] \approx \frac{1}{n}\sum_{i=1:n} f(x,z_i,w_i)$$

PDPs

$$\hat{f}_{\text{M}}(x) = \mathbf{E}_{Z,W|X=x}[\hat{f}(x,Z,W)]$$

$$\approx \frac{1}{|B_x|}\sum_{i\in B_x} f(x,z_i,w_i), \text{ where } B_x = \{i : |x_i - x| < \delta\}$$

M-plots

$$\hat{f}_{\text{M}}(x) \approx \sum_{k=1}^{\lceil x \rceil}\left(\frac{1}{|B_x(k)|}\sum_{i\in B_x(k)} \underline{(f(k+1,z_i,w_i) - f(k,z_i,w_i))}\right),$$

ALEs

$$\text{where } B_x(k) = \{i : x_i \in [k, k+1)\}$$

Accumulated *over X values smaller than or equal to x*

Averaged over datapoints with x in that ***Local*** *neighborhood*

Effect *of a unit change from k to k+1*

Accumulated Local Effect

$$\hat{f}_{\mathrm{PDP}}(x) = \mathbf{E}_{Z,W}[\hat{f}(x, Z, W)] \approx \frac{1}{n} \sum_{i=1:n} f(x, z_i, w_i)$$

PDPs

Accumulated Local Effect

$$\hat{f}_{\mathrm{PDP}}(x) = \mathbf{E}_{Z,W}[\hat{f}(x, Z, W)] \approx \frac{1}{n} \sum_{i=1:n} f(x, z_i, w_i)$$

PDPs

$$\begin{aligned} \hat{f}_{\mathrm{M}}(x) &= \mathbf{E}_{Z,W|X=x}[\hat{f}(x, Z, W)] \\ &\approx \frac{1}{|B_x|} \sum_{i \in B_x} f(x, z_i, w_i), \text{ where } B_x = \{i : |x_i - x| < \delta\} \end{aligned}$$

M-plots

Accumulated Local Effect

$$\hat{f}_{\mathrm{PDP}}(x) = \mathbf{E}_{Z,W}[\hat{f}(x,Z,W)] \approx \frac{1}{n}\sum_{i=1:n} f(x,z_i,w_i)$$

PDPs

$$\hat{f}_{\mathrm{M}}(x) = \mathbf{E}_{Z,W|X=x}[\hat{f}(x,Z,W)]$$
$$\approx \frac{1}{|B_x|}\sum_{i\in B_x} f(x,z_i,w_i), \text{ where } B_x = \{i : |x_i - x| < \delta\}$$

M-plots

$$\hat{f}_{\mathrm{M}}(x) \approx \sum_{k=1}^{\lceil x \rceil}\left(\frac{1}{|B_x(k)|}\sum_{i\in B_x(k)} \underline{(f(k+1,z_i,w_i) - f(k,z_i,w_i))}\right),$$
$$\text{where } B_x(k) = \{i : x_i \in [k, k+1)\}$$

ALEs

***Effect** of a unit change*
from k to k+1

Accumulated Local Effect

$$\hat{f}_{\text{PDP}}(x) = \mathbf{E}_{Z,W}[\hat{f}(x, Z, W)] \approx \frac{1}{n} \sum_{i=1:n} f(x, z_i, w_i)$$

PDPs

$$\hat{f}_{\text{M}}(x) = \mathbf{E}_{Z,W|X=x}[\hat{f}(x, Z, W)]$$

$$\approx \frac{1}{|B_x|} \sum_{i \in B_x} f(x, z_i, w_i), \text{ where } B_x = \{i : |x_i - x| < \delta\}$$

M-plots

$$\hat{f}_{\text{M}}(x) \approx \sum_{k=1}^{\lceil x \rceil} \left(\frac{1}{|B_x(k)|} \sum_{i \in B_x(k)} (f(k+1, z_i, w_i) - f(k, z_i, w_i)) \right),$$

ALEs

$$\text{where } B_x(k) = \{i : x_i \in [k, k+1)\}$$

*Averaged over datapoints with x in that **Local** neighborhood*

***Effect** of a unit change from k to k+1*

Accumulated Local Effect

$$\hat{f}_{\text{PDP}}(x) = \mathbf{E}_{Z,W}[\hat{f}(x, Z, W)] \approx \frac{1}{n} \sum_{i=1:n} f(x, z_i, w_i)$$

PDPs

$$\hat{f}_{\text{M}}(x) = \mathbf{E}_{Z,W|X=x}[\hat{f}(x, Z, W)]$$

$$\approx \frac{1}{|B_x|} \sum_{i \in B_x} f(x, z_i, w_i), \text{ where } B_x = \{i : |x_i - x| < \delta\}$$

M-plots

$$\hat{f}_{\text{M}}(x) \approx \sum_{k=1}^{\lceil x \rceil} \left(\frac{1}{|B_x(k)|} \sum_{i \in B_x(k)} \left(f(k+1, z_i, w_i) - f(k, z_i, w_i)\right) \right),$$

$$\text{where } B_x(k) = \{i : x_i \in [k, k+1)\}$$

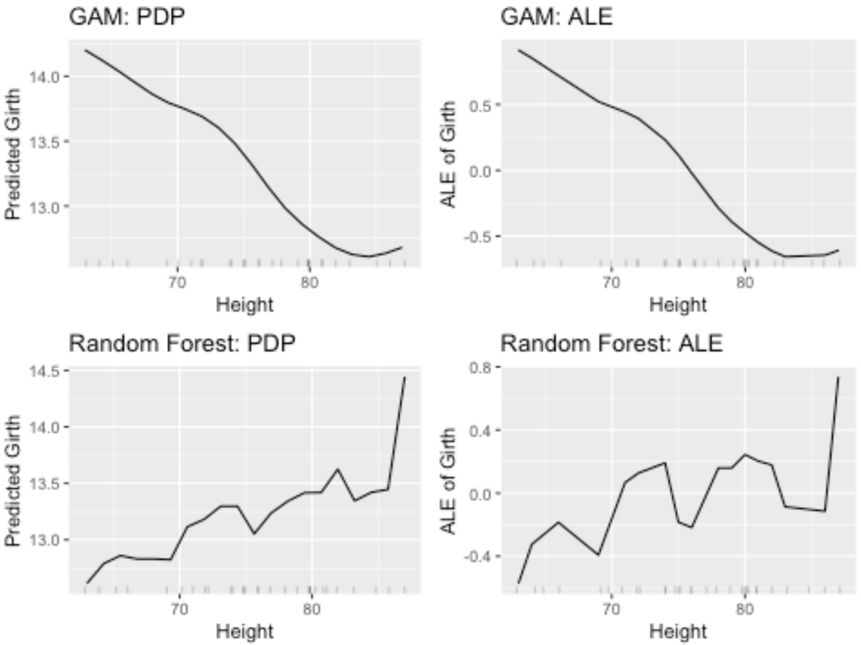
ALEs

Accumulated *over X values smaller than or equal to x*

*Averaged over datapoints with x in that **Local** neighborhood*

Effect *of a unit change from k to k+1*

Accumulated Local Effects curves




```
> library(iml)

> mod_gam <- Predictor$new(gam(Girth
~s(Height)+s(Volume), data = trees))

> FeatureEffect$new(mod_gam, feature =
"Height", method = “ale”)$plot()
```

Summary

- ALEs are the most faithful generalization of the concept of assessing the impact of the average effect of a unit change on the response.
- They generally address the problem that PDPs have of averaging over unlikely observations, though fundamentally the problem still persists if the correlation is so strong that it is also present locally in small neighborhoods.
- For GAMs, all of these concepts are identical.