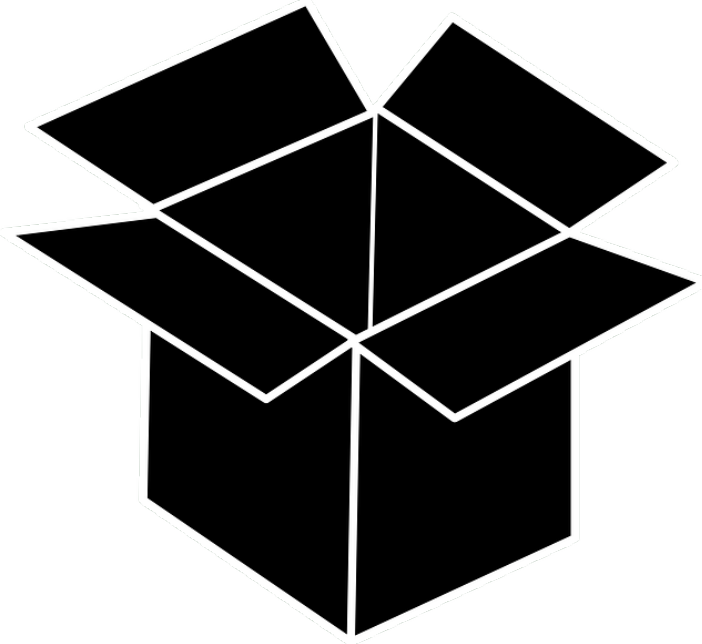
Interpretable Machine Learning

Interpretable Models



Learning goals

- Marginal Effects

MARGINAL EFFECTS

- In LMs with interactions, higher order terms, or GLMs, the feature term cannot be interpreted as a direct effect on the predicted outcome.
- The default way to interpret LMs and GLMs with non-linear feature effects is the marginal effect (ME).
- Either, we take the derivative of the prediction function w.r.t. a feature, or we evaluate changes in prediction due to an intervention in the data, e.g., by increasing a feature value by one unit:

$$ME_j(x) = \frac{\partial f(x)}{\partial x_j}$$

$$ME_j(x, h_j) = f(x_1, \ldots, x_j + h_j, \ldots, x_p) - f(x)$$

- The average marginal effect (AME) is a global feature effect estimate in form of a single metric.