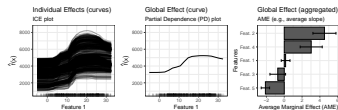
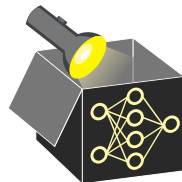


Interpretable Machine Learning

Feature Effects

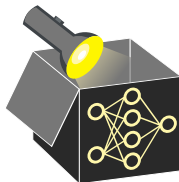
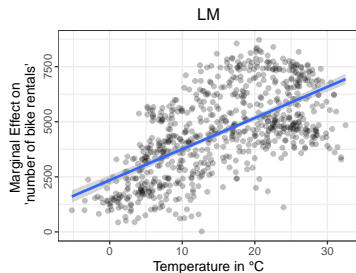
Introduction to Feature Effects



Learning goals

- Global Feature Effects
- Local Feature Effects

FEATURE EFFECTS - GLOBAL VIEW



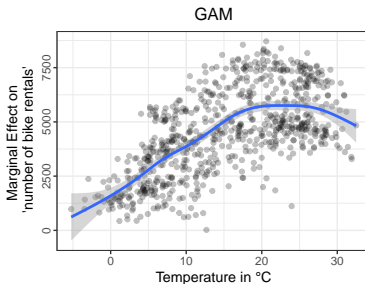
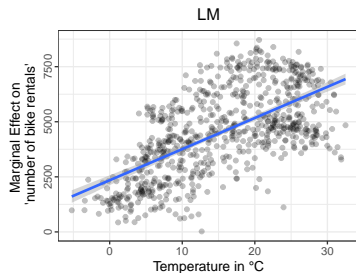
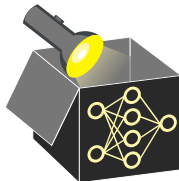
LM without interaction:

$\hat{\theta}_j$ is linear effect of feature x_j

(applies globally to all observations):

- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1 \hat{\theta}_1$
- Scalar $\hat{\theta}_1$ describes global effect

FEATURE EFFECTS - GLOBAL VIEW



LM without interaction:

$\hat{\theta}_j$ is linear effect of feature x_j

(applies globally to all observations):

- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1 \hat{\theta}_1$
- Scalar $\hat{\theta}_1$ describes global effect

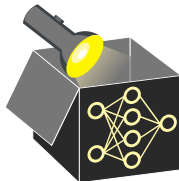
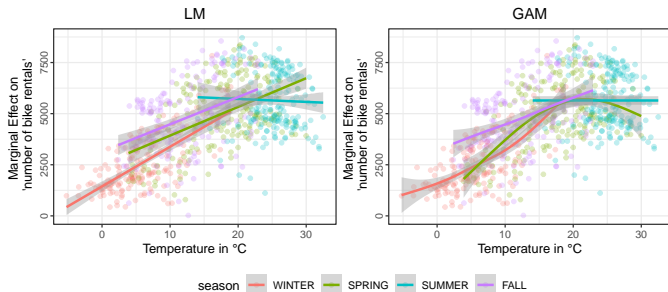
GAM without interaction:

$\hat{f}_j(x_j)$ is non-lin. effect of feature x_j

(applies globally to all observations):

- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + \hat{f}_1(x_1)$
- Curve \hat{f}_1 describes global effect

FEATURE EFFECTS - LOCALIZED VIEW

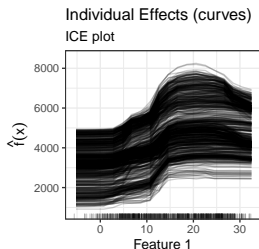
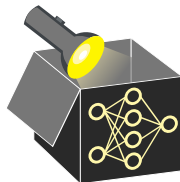


- **Interactions:** Feature effect depends on other features and varies across obs.
 - ⇒ E.g., effect of **temperature** varies across **season**
 - ⇒ Multiple values / curves needed to describe effect
- ML models capture non-linear effects and high-order interactions
 - ⇒ Global view may mislead (single curve may fail to capture complexity)
 - ⇒ Local feat. effect methods needed to estimate effects for individ. obs.
 - ⇒ Global view can be reconstructed by aggregating local effects

FEATURE EFFECTS

Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

- Analogous to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels exist (simplification but information loss)
- Methods: ICE curves (local curves)

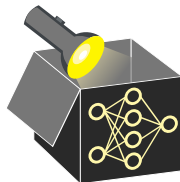


Individual (curves)

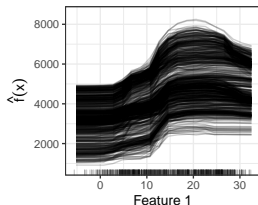
FEATURE EFFECTS

Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

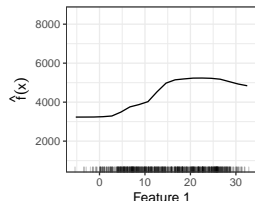
- Analogous to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels exist (simplification but information loss)
- Methods: ICE curves (local curves), PD and ALE plots (global curves)



Individual Effects (curves)
ICE plot



Global Effect (curve)
Partial Dependence (PD) plot

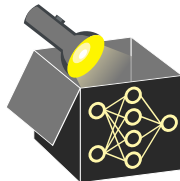


Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$ Global (single curve)

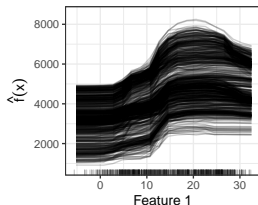
FEATURE EFFECTS

Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

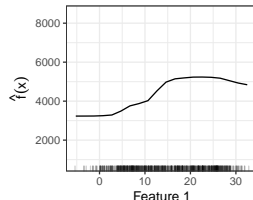
- Analogous to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels exist (simplification but information loss)
- Methods: ICE curves (local curves), PD and ALE plots (global curves), AME (global value)



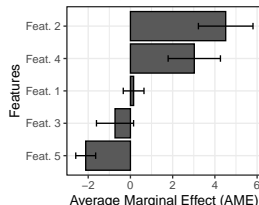
Individual Effects (curves)
ICE plot



Global Effect (curve)
Partial Dependence (PD) plot



Global Effect (aggregated)
AME (e.g., average slope)



Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$

Global (single curve) $\xrightarrow[\text{slopes}]{\text{aggregate}}$

Global (single value)