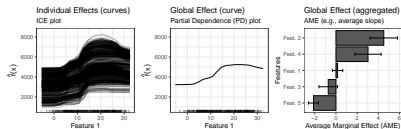


Interpretable Machine Learning

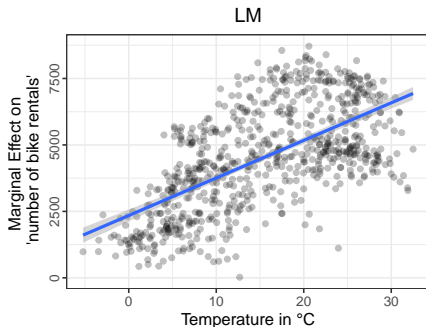
Introduction to Feature Effects



Learning goals

- Global Feature Effects
- Local Feature Effects

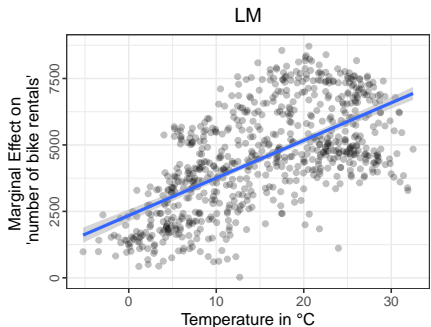
FEATURE EFFECTS - GLOBAL VIEW



LM without interaction: $\hat{\theta}_j$ is linear effect of feature x_j (applies globally to all observations):

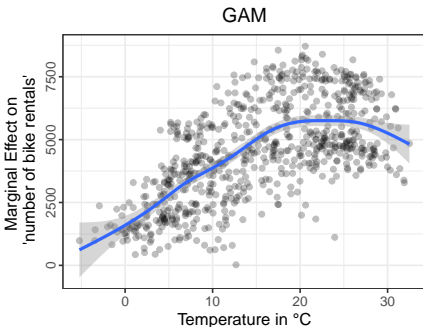
- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1 \hat{\theta}_1$
- Single value $\hat{\theta}_1$ describes global effect

FEATURE EFFECTS - GLOBAL VIEW



LM without interaction: $\hat{\theta}_j$ is linear effect of feature x_j (applies globally to all observations):

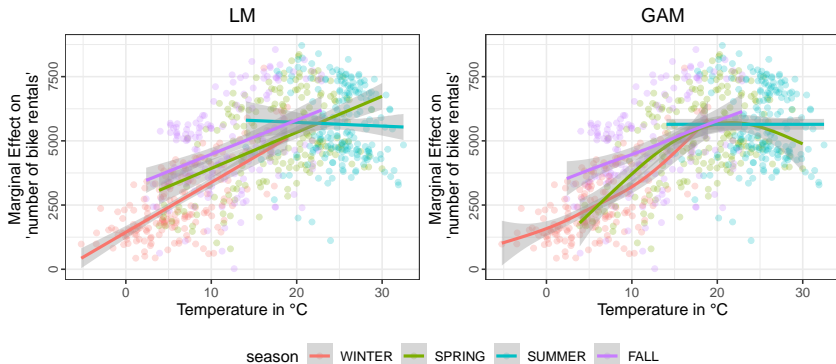
- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1 \hat{\theta}_1$
- Single value $\hat{\theta}_1$ describes global effect



GAM without interaction: $\hat{f}_j(x_j)$ is non-linear effect of feature x_j (applies globally):

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

FEATURE EFFECTS - LOCAL VIEW

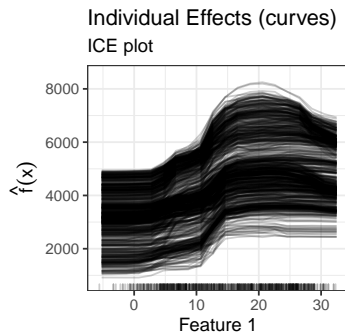


- Interactions: Feature effect is modified by other features and varies across observations
 - ⇒ Effect of temperature varies across seasons
 - ⇒ Multiple values / curves needed to describe effect
- ML models often model non-linear effects and complex interactions
 - ⇒ Need for local feature effect methods, e.g., analyze effect for individual observations
 - ⇒ Analyzing global effects by aggregating local effects

FEATURE EFFECTS

Feature effects visualize or quantify marginal contribution of a feature of interest w.r.t. predictions

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

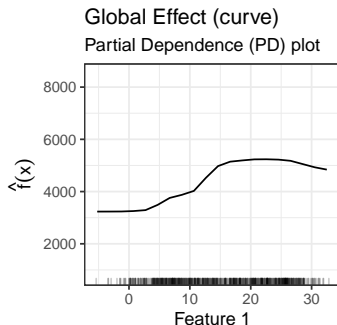
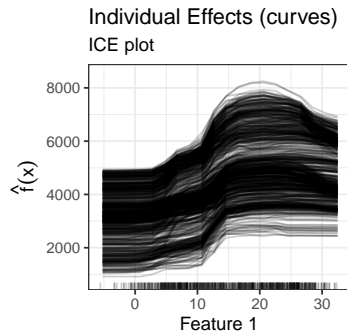


Individual (curves)

FEATURE EFFECTS

Feature effects visualize or quantify marginal contribution of a feature of interest w.r.t. predictions

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



Individual (curves)

aggregate
→
curves

Global (single curve)

FEATURE EFFECTS

Feature effects visualize or quantify marginal contribution of a feature of interest w.r.t. predictions

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

