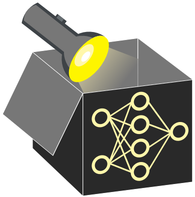
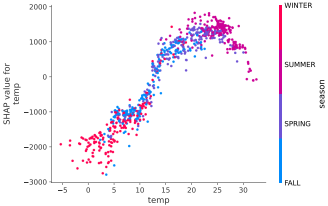
Interpretable Machine Learning



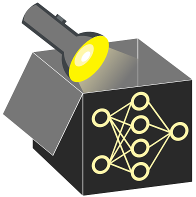
Global SHAP



Learning goals

- Understand how SHAP values can be aggregated for global model interpretation
- Learn global SHAP visualizations: feature importance, summary, and dependence plots
- Recognize advantages and limitations of global SHAP explanations

GLOBAL SHAP ▸ Lundberg et al. 2018

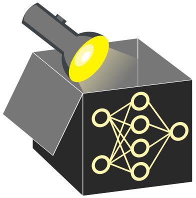


Idea:

- Run SHAP for every observation and thereby get a matrix of Shapley values
- The matrix has one row per data observation and one column per feature
- We can interpret the model globally by analyzing the Shapley value matrix

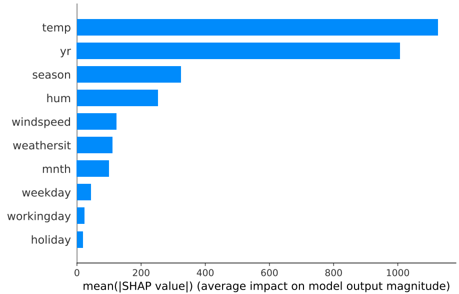
$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_{22} & \phi_{23} & \cdots & \phi_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \phi_{n3} & \cdots & \phi_{np} \end{bmatrix}$$

FEATURE IMPORTANCE

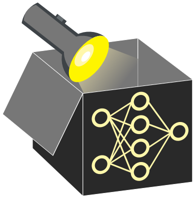


Idea: Average the absolute Shapley values of each feature over all observations.
This corresponds to calculating averages column by column in matrix Φ

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right|$$

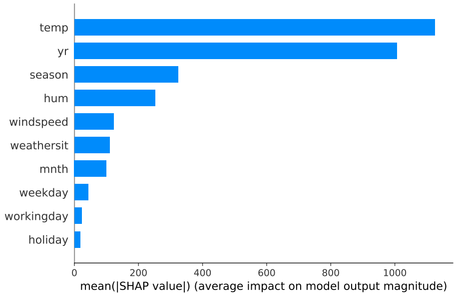


FEATURE IMPORTANCE

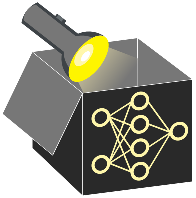


Interpretation:

- Features "temp" and "year" have highest influence on the model's prediction
- Shapley FI does not provide information on direction of the effect
  $\rightsquigarrow$ Provides a feature ranking based on the magnitude of the Shapley values
- Shapley FI is based only on model predictions
  Note: Other FI measures are based on model's performance (loss)

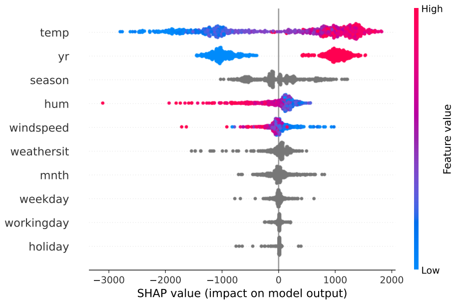


SUMMARY PLOT

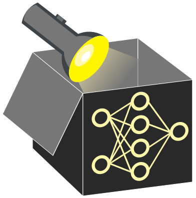


Combines feature importance with feature effects

- Each point is a Shapley value for a feature and an observation
- The color represents the value of the feature from low to high
- Overlapping points are jittered in y-axis direction

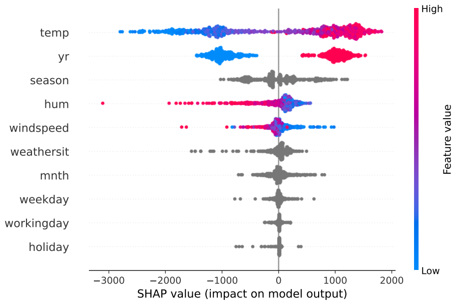


SUMMARY PLOT



Interpretation:

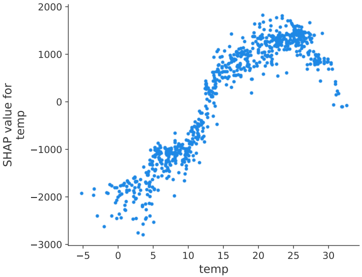
- Low temp have a negative impact; high temp lead to more bike rentals
- Year: two point clouds for 2011 (low value) and 2012 (high value)
- Categorical features are gray (no low/high value)
- High humidity has a huge negative impact on bike rentals
- Low humidity has a rather minor positive impact on bike rentals



DEPENDENCE PLOT: EFFECT + INTERACTION

Interpretation of SHAP Dependence Plot (Feature = Temperature)

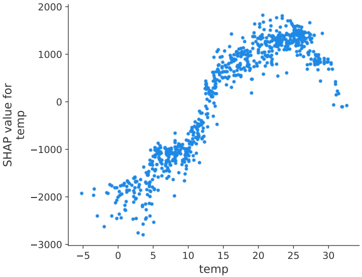
- Plot points with feature value on x-axis and corresponding SHAP value on y-axis



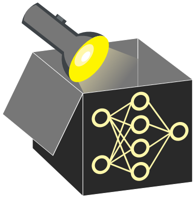
DEPENDENCE PLOT: EFFECT + INTERACTION

Interpretation of SHAP Dependence Plot (Feature = Temperature)

- Plot points with feature value on x-axis and corresponding SHAP value on y-axis
- Shows how temp influences bike rentals $\rightsquigarrow$ Marginal effect similar to PD plot
- SHAP values increase with temp until ≈25 ℃: higher temp $\rightsquigarrow$ higher predictions
- After ≈25 ℃, SHAP values decrease slightly

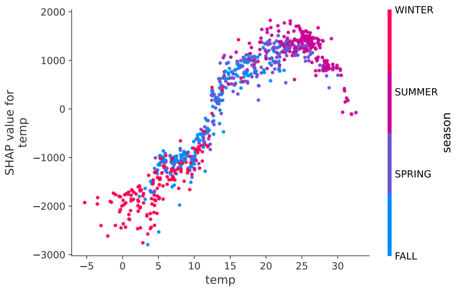


DEPENDENCE PLOT: EFFECT + INTERACTION

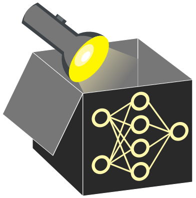


Interpretation of SHAP Dependence Plot (Feature = Temperature)

- Plot points with feature value on x-axis and corresponding SHAP value on y-axis
- Shows how temp influences bike rentals $\rightsquigarrow$ Marginal effect similar to PD plot
- SHAP values increase with temp until $\approx$25 ℃: higher temp $\rightsquigarrow$ higher predictions
- After $\approx$25 ℃, SHAP values decrease slightly
- Interaction with **season** is visible (via color-encoded observations):
    - In summer, higher temperatures decrease bike rentals
    - In winter, higher temperatures increase bike rentals



DISCUSSION



Advantages

- Retains local accuracy: SHAP values exactly decompose predictions
- Aggregating local SHAP values yields global model insights
  $\rightsquigarrow$ Visual diagnostics: feature importance, summary plot, dependence plots
- Efficient for tree-based models via TreeSHAP
  (See ▸ Lundberg et al 2018 and for intuitive explanation ▸ Sukumar: TreeSHAP)
- Unifies feature attribution under a consistent additive framework
- Can be used for images ▸ SHAP image examples and text ▸ SHAP text examples

Disadvantages

- KernelSHAP is inefficient for large datasets or complex models
- Ignores feature dependencies in marginal sampling (interventional SHAP)
- Conditional sampling (observational SHAP) is difficult in practice (would require estimating a conditional distribution)