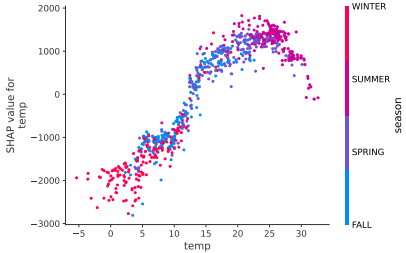
Interpretable Machine Learning

Global SHAP



Learning goals

- Get an intuition of additive feature attributions
- Understand the concept of Kernel SHAP
- Ability to interpret SHAP plots
- Global SHAP methods

GLOBAL SHAP ▸ Lundberg et al. 2018

Idea:

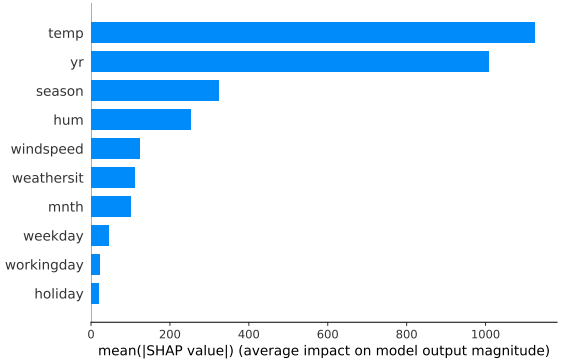
- Run SHAP for every observation and thereby get a matrix of Shapley values
- The matrix has one row per data observation and one column per feature
- We can interpret the model globally by analyzing the Shapley values in this matrix

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

FEATURE IMPORTANCE

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

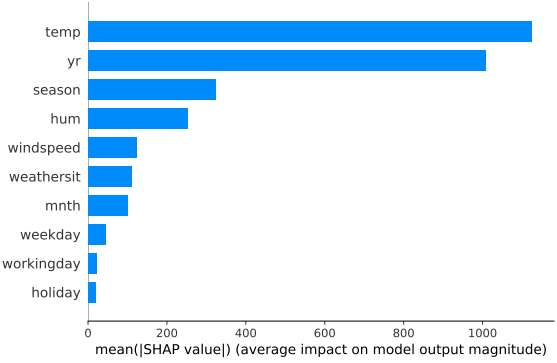
$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right|$$



FEATURE IMPORTANCE

Interpretation:

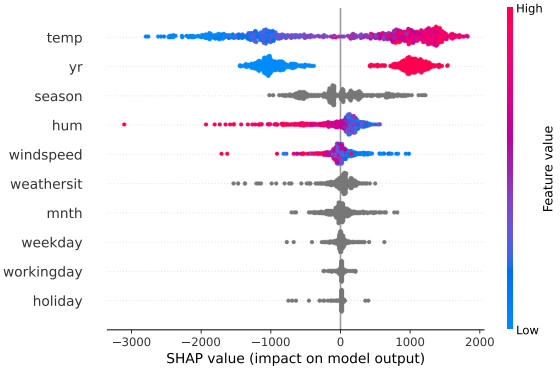
- The features temperature and year have by far the highest influence on the model's prediction
- Compared to Shapley values, no effect direction is provided, but instead a feature ranking similar to PFI
- However, Shapley FI is based on the model's predictions only while PFI is based on the model's performance (loss)



SUMMARY PLOT

Combines feature importance with feature effects

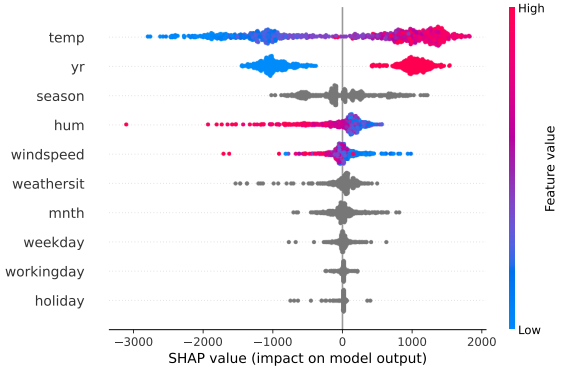
- Each point is a Shapley value for a feature and an observation
- The color represents the value of the feature from low to high
- Overlapping points are jittered in y-axis direction



SUMMARY PLOT

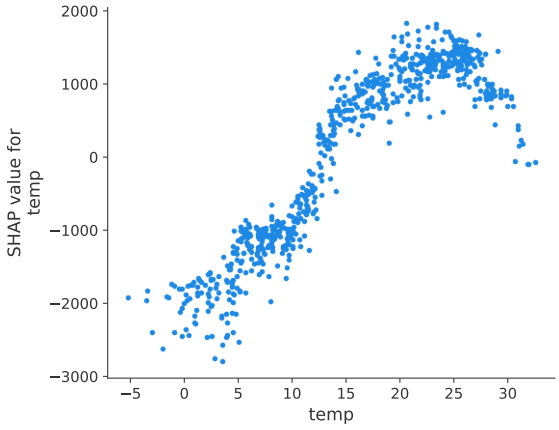
Interpretation:

- Low temperatures have a negative impact while high temperatures lead to more bike rentals
- Year: two point clouds for 2011 and 2012 (other categorical features are gray)
- A high humidity has a huge, negative impact on the bike rental, while low humidity has a rather minor positive impact on bike rentals



DEPENDENCE PLOT

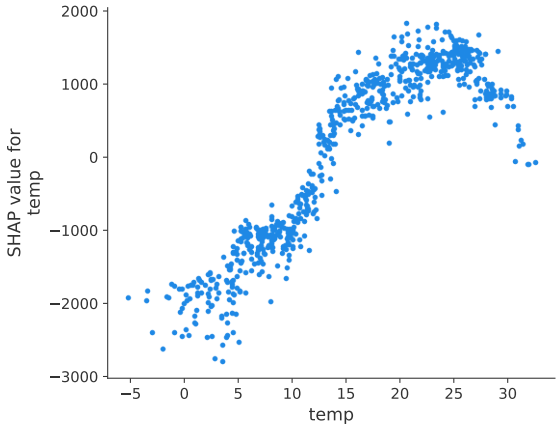
- Visualize the marginal contribution of a feature similar to the PDP
- Plot a point with the feature value on the x-axis and the corresponding Shapley value on the y-axis



DEPENDENCE PLOT

Interpretation:

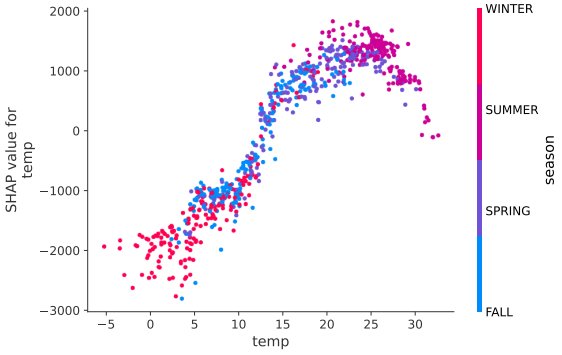
- Increasing temperatures induce increasing bike rentals until 25°C
- If it gets too hot, the bike rentals decrease



DEPENDENCE PLOT

Interpretation:

- We can colour the observations by a second feature to detect interactions
- Visibly the temperatures interaction with the season is very strong



DISCUSSION

Advantages

- All the advantages of Shapley values
- Unify the field of interpretable machine learning in the class of additive feature attribution methods
- Has a fast implementation for tree-based models
- Various global interpretation methods

Disadvantages

- Disadvantages of Shapley values also apply to SHAP
- KernelSHAP is slow (TreeSHAP can be used as a faster alternative for tree-based models ▸ Lundberg et al 2018 – and for an intuitive explanation ▸ see Sukumar: TreeSHAP)
- KernelSHAP ignores feature dependence