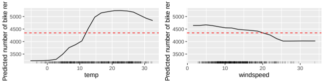
Interpretable Machine Learning

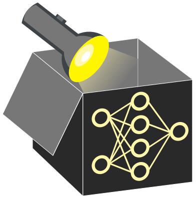
Partial Dependence Feature Importance



Learning goals

- Introduction to PDP feature importance
- Numerical and Categorical Measures of flatness

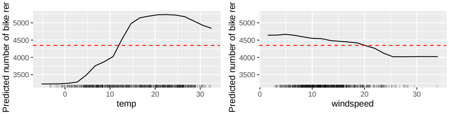
PARTIAL DEPENDENCE - REVISION



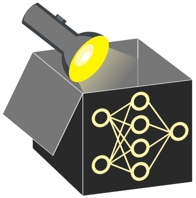
The partial dependence (PD) is the expectation of ICE curves w.r.t. the marginal distribution of complementary features $\mathbf{x}_{-S}$.
The PD function $\hat{f}$ is estimated by the point-wise average of the ICE curves at x_S^*:

$$\hat{f}_{S,PD}(x_S^*) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S^*, \mathbf{x}_{-S}^{(i)})$$

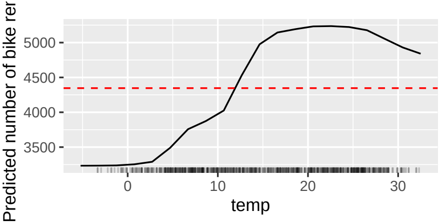
PD plot: Visualizes the **average effect of a feature**, i.e., how the expected prediction changes if the feature value is changed.

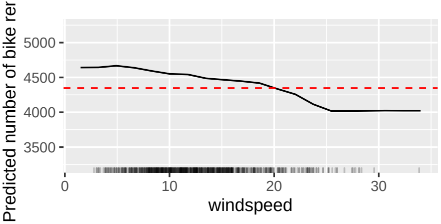


PD FEATURE IMPORTANCE - MOTIVATION



- The basic motivation is that a flat PDP indicates that the feature is not important
- The more the PDP varies, the more important the feature is



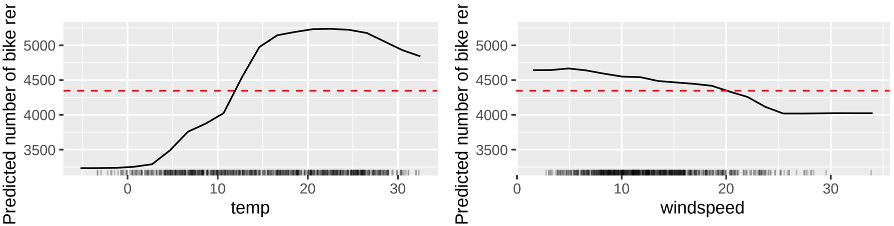


The notion of variable importance is based on any measure of the **flatness** $F(\cdot)$ of the partial dependence function $\hat{f}$.

$$I(x) = F\left(\hat{f}_S\left(\mathbf{z}_S\right)\right)$$

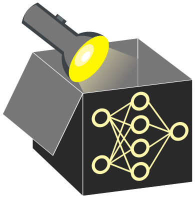
PD FEATURE IMPORTANCE - IDEA

- Therefore, we focus our attention to the surface that spans between the PDP curve itself and the average of all feature values (red dashed line). $\sum_{k=1}^{K} \hat{f}_S\left(x_S^{(k)}\right)$

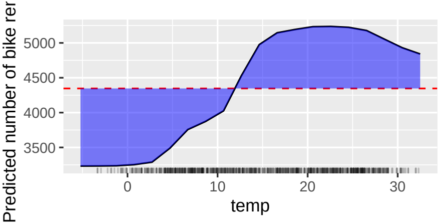


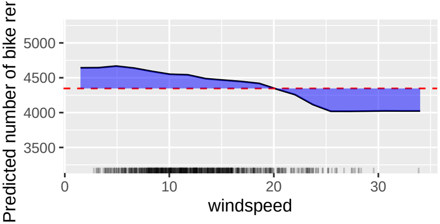
The function $F(\cdot)$ determines the exact quantification of the **flatness** measure.

PD FEATURE IMPORTANCE - IDEA



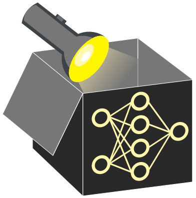
- Therefore, we focus our attention to the surface that spans between the PDP curve itself and the average of all feature values (red dashed line). $\sum_{k=1}^{K} \hat{f}_S \left(x_S^{(k)} \right)$





The function $F(\cdot)$ determines the exact quantification of the **flatness** measure.

MEASURES OF FLATNESS



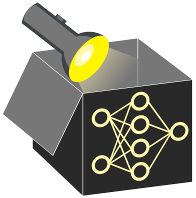
For **numerical** features, importance is defined as the deviation of each unique feature value from the average curve

$$I\left(x_S\right) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{f}_S\left(x_S^{(k)}\right) - \frac{1}{K}\sum_{k=1}^{K}\hat{f}_S\left(x_S^{(k)}\right)\right)^2}$$

For **categorical** we achieve a rough estimate of the deviation by applying the range rule: This is the range of the PDP values for the unique categories divided by four.

$$I\left(x_S\right) = \left(\max_k\left(\hat{f}_S\left(x_S^{(k)}\right)\right) - \min_k\left(\hat{f}_S\left(x_S^{(k)}\right)\right)\right)/4$$

DISCUSSION



Advantages

variable importance measure that is...

- suitable for use with any supervised learning algorithm, provided new predictions can be obtained
- model-based and takes into account the effect of all the features in the model
- consistent and has the same interpretation regardless of the learning algorithm employed
- has the potential to help identify possible interaction effects.

Disadvantages

- This PDP-based feature importance should be interpreted with care.
- It captures only the main effect of the feature and ignores possible feature interactions.
- variable importance metric relies on the fitted model; hence, it is crucial to properly tune and train the model to attain the best performance possible.