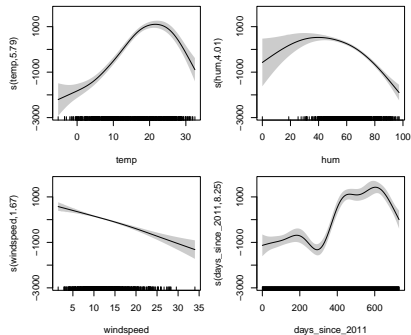


Interpretable Machine Learning

GAM & Boosting



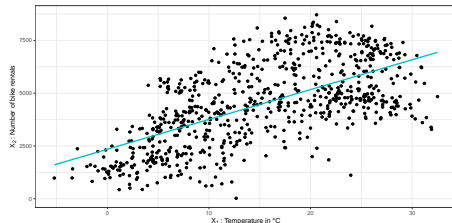
Learning goals

- Generalized additive model
- Model-based boosting with simple base learners
- Feature effect and importance in model-based boosting

GENERALIZED ADDITIVE MODEL (GAM)

► Hastie and Tibshirani (1986)

Problem: LM not suitable if relationship between features and target variable is not linear



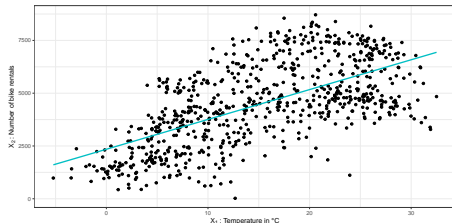
GENERALIZED ADDITIVE MODEL (GAM)

► Hastie and Tibshirani (1986)

Problem: LM not suitable if relationship between features and target variable is not linear

Workaround in LMs / GLMs:

- Feature transformations (e.g., exp or log)
- Including high-order effects
- Categorization of features (i.e., intervals / buckets of feature values)



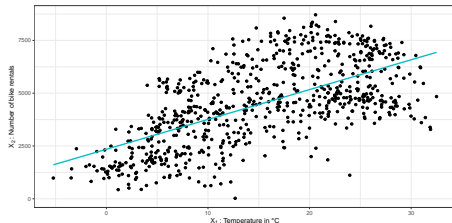
GENERALIZED ADDITIVE MODEL (GAM)

► Hastie and Tibshirani (1986)

Problem: LM not suitable if relationship between features and target variable is not linear

Workaround in LMs / GLMs:

- Feature transformations (e.g., exp or log)
- Including high-order effects
- Categorization of features (i.e., intervals / buckets of feature values)



Idea of GAMs:

- Instead of linear terms $\theta_j x_j$, use flexible functions $f_j(x_j) \rightsquigarrow$ splines

$$g(\mathbb{E}(y \mid \mathbf{x})) = \theta_0 + f_1(x_1) + f_2(x_2) + \dots + f_p(x_p)$$

- Preserves additive structure and allows to model non-linear effects
- Splines have a smoothness parameter to control flexibility (prevent overfitting)
 \rightsquigarrow Needs to be chosen, e.g., via cross-validation

GENERALIZED ADDITIVE MODEL (GAM) - EXAMPLE

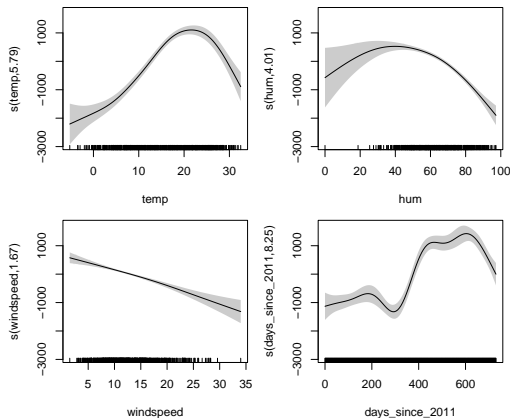
Fit a GAM with smooth splines for four numeric features of bike rental data

↪ more flexible and better model fit but less interpretable than LM

	edf	p-value
s(temp)	5.8	0.00
s(hum)	4.0	0.00
s(windspeed)	1.7	0.00
s(days_since_2011)	8.3	0.00

Interpretation

- Interpretation needs to be done visually and relative to average prediction
- Edf (effective degrees of freedom) represents complexity of smoothness



MODEL-BASED BOOSTING

► Bühlmann and Yu 2003

- Recall: Boosting iteratively combines weak base learners (BL)
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)

- Recall: Boosting iteratively combines weak base learners (BL)
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)
- Possible to combine linear BL of same type (with distinct parameters θ and θ^*):

$$b_j(x, \theta) + b_j(x, \theta^*) = b_j(x, \theta + \theta^*)$$

- Recall: Boosting iteratively combines weak base learners (BL)
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)
- Possible to combine linear BL of same type (with distinct parameters θ and θ^*):

$$b_j(x, \theta) + b_j(x, \theta^*) = b_j(x, \theta + \theta^*)$$

- In each iteration, fit a set of BLs and add the best BL to previous model (using step-size ν):

$$\hat{f}^{[1]}(x) = \hat{f}_0 + \nu b_3(x_3, \theta^{[1]})$$

- Recall: Boosting iteratively combines weak base learners (BL)
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)
- Possible to combine linear BL of same type (with distinct parameters θ and θ^*):

$$b_j(x, \theta) + b_j(x, \theta^*) = b_j(x, \theta + \theta^*)$$

- In each iteration, fit a set of BLs and add the best BL to previous model (using step-size ν):

$$\widehat{f}^{[1]}(x) = \widehat{f}_0 + \nu b_3(x_3, \theta^{[1]})$$

$$\widehat{f}^{[2]}(x) = \widehat{f}^{[1]}(x) + \nu b_3(x_3, \theta^{[2]})$$

- Recall: Boosting iteratively combines weak base learners (BL)
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)
- Possible to combine linear BL of same type (with distinct parameters θ and θ^*):

$$b_j(x, \theta) + b_j(x, \theta^*) = b_j(x, \theta + \theta^*)$$

- In each iteration, fit a set of BLs and add the best BL to previous model (using step-size ν):

$$\hat{f}^{[1]}(x) = \hat{f}_0 + \nu b_3(x_3, \theta^{[1]})$$

$$\hat{f}^{[2]}(x) = \hat{f}^{[1]}(x) + \nu b_3(x_3, \theta^{[2]})$$

$$\hat{f}^{[3]}(x) = \hat{f}^{[2]}(x) + \nu b_1(x_1, \theta^{[3]})$$

$$= \hat{f}_0 + \nu \left(b_3(x_3, \theta^{[1]} + \theta^{[2]}) + b_1(x_1, \theta^{[3]}) \right)$$

$$= \hat{f}_0 + \hat{f}_3(x_3) + \hat{f}_1(x_1)$$

- Recall: Boosting iteratively combines weak base learners (BL)
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)
- Possible to combine linear BL of same type (with distinct parameters θ and θ^*):

$$b_j(x, \theta) + b_j(x, \theta^*) = b_j(x, \theta + \theta^*)$$

- In each iteration, fit a set of BLs and add the best BL to previous model (using step-size ν):

$$\hat{f}^{[1]}(x) = \hat{f}_0 + \nu b_3(x_3, \theta^{[1]})$$

$$\hat{f}^{[2]}(x) = \hat{f}^{[1]}(x) + \nu b_3(x_3, \theta^{[2]})$$

$$\hat{f}^{[3]}(x) = \hat{f}^{[2]}(x) + \nu b_1(x_1, \theta^{[3]})$$

$$= \hat{f}_0 + \nu \left(b_3(x_3, \theta^{[1]} + \theta^{[2]}) + b_1(x_1, \theta^{[3]}) \right)$$

$$= \hat{f}_0 + \hat{f}_3(x_3) + \hat{f}_1(x_1)$$

- Final model is additive (as GAMs), where each component function is interpretable

MODEL-BASED BOOSTING - LINEAR EXAMPLE

Simple case: Use linear model with single feature (including intercept) as BL

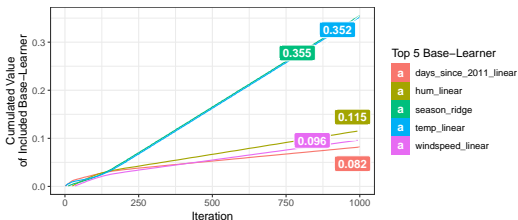
$$b_j(x_j, \theta) = x_j \theta + \theta_0 \quad \text{for } j = 1, \dots, p \quad \rightsquigarrow \text{ordinary linear regression}$$

- Here: Interpretation of weights as in LM
- After many iterations, it converges to same solution as least square estimate of LMs

1000 iter. with $\nu = 0.1$	Intercept	Weights
days_since_2011	-1791.06	4.9
hum	1953.05	-31.1
season	0	WINTER: -323.4 SPRING: 539.5 SUMMER: -280.2 FALL: 67.2
temp	-1839.85	120.4
windspeed	725.70	-56.9
offset	4504.35	

⇒ Converges to solution of LM

Relative frequency of selected BLs across iterations



MODEL-BASED BOOSTING - LINEAR EXAMPLE

Simple case: Use linear model with single feature (including intercept) as BL

$$b_j(x_j, \theta) = x_j \theta + \theta_0 \quad \text{for } j = 1, \dots, p \quad \rightsquigarrow \text{ordinary linear regression}$$

- Here: Interpretation of weights as in LM
- After many iterations, it converges to same solution as least square estimate of LMs
- Early stopping allows feature selection and might prevent overfitting (regularization)

1000 iter. with $\nu = 0.1$	Intercept	Weights
days_since_2011	-1791.06	4.9
hum	1953.05	-31.1
season	0	WINTER: -323.4 SPRING: 539.5 SUMMER: -280.2 FALL: 67.2
temp	-1839.85	120.4
windspeed	725.70	-56.9
offset	4504.35	

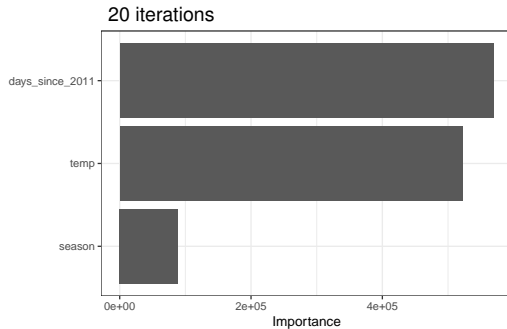
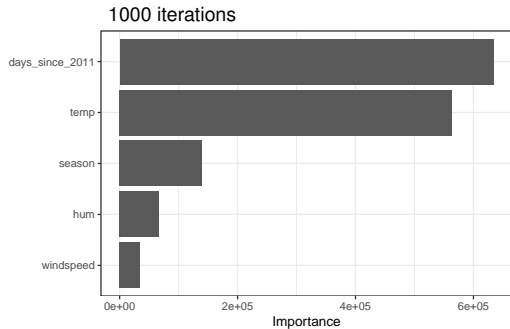
⇒ Converges to solution of LM

20 iter. with $\nu = 0.1$	Intercept	Weights
days_since_2011	-1210.27	3.3
season	0	WINTER: -276.9 SPRING: 137.6 SUMMER: 112.8 FALL: 20.3
temp	-1118.94	73.2
offset	4504.35	

⇒ 3 BLs selected after 20 iter. (feature selection)

MODEL-BASED BOOSTING - LINEAR EXAMPLE: INTERPRETATION

Feature importance (risk reduction over iterations)



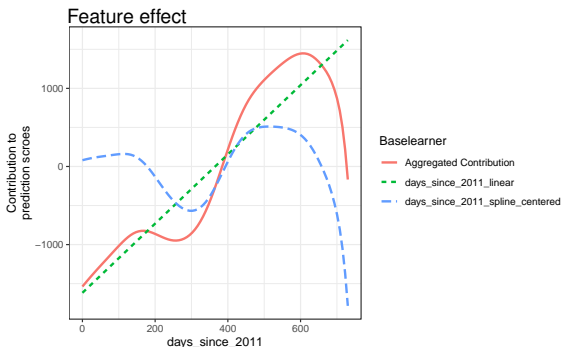
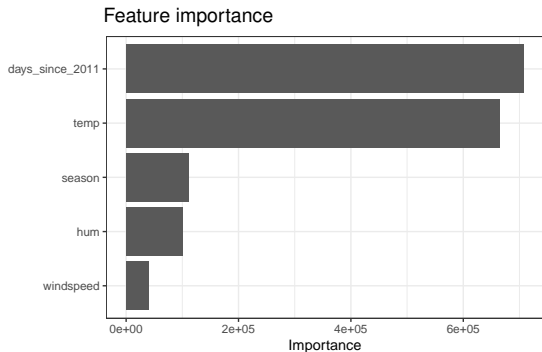
⇒ `days_since_2011` most important, followed by `temp`

MODEL-BASED BOOSTING - INTERPRETATION

- Fit model on bike data with different BL types ▶ Daniel Schalk et al. 2018
- BLs: linear and centered splines for numeric features, categorical for season

MODEL-BASED BOOSTING - INTERPRETATION

- Fit model on bike data with different BL types ► Daniel Schalk et al. 2018
- BLs: linear and centered splines for numeric features, categorical for season



- Total effect for days_since_2011
 ↪ Combination of partial effects of linear BL and centered spline BL