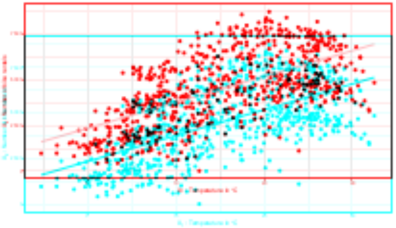
GENERALIZED ADDITIVE MODEL (GAM)

▸ Hastie and Tibshirani (1986)

Problem: LM not great if features act on outcome non-linearly target variable is not linear



GENERALIZED ADDITIVE MODEL (GAM)

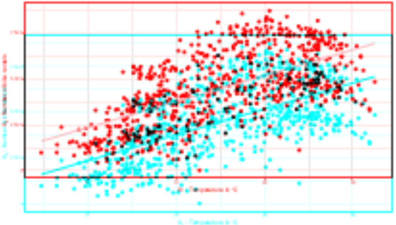
▸ Hastie and Tibshirani (1986)

Problem: LM not great if features act on outcome non-linearly

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)



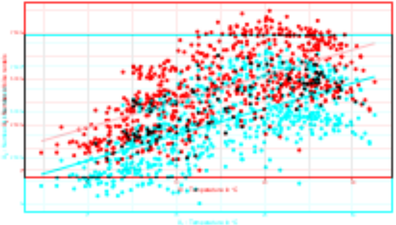
GENERALIZED ADDITIVE MODEL (GAM)

▸ Hastie and Tibshirani (1986)

Problem: LM not great if features act on outcome non-linearly

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

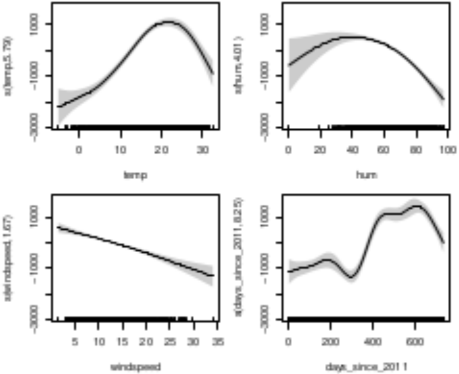
$$g(\mathbb{E}(y \mid \mathbf{x})) = \theta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

- 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, also see PDPs
- Edf (effective degrees of freedom) represents complexity of smoothness

MODEL-BASED BOOSTING ▸ Bühlmann and Yu 2003

- Boosting iteratively combines weak base learners to create powerful ensemble
- Idea: Use simple BLs (e.g univariate, with splines) to ensure interpretability
- Possible to combine BL of same type (with distinct parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$):

$$b^{[j]}(\mathbf{x}, \boldsymbol{\theta}) + b^{[j]}(\mathbf{x}, \boldsymbol{\theta}^\star) = b^{[j]}(\mathbf{x}, \boldsymbol{\theta} + \boldsymbol{\theta}^\star)$$

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

MODEL-BASED BOOSTING

Bühlmann and Yu 2003

- Recall: Boosting iteratively combines weak base learners (BL)
- Boosting iteratively combines weak base learners to create powerful ensemble
- Idea: Use simple BLs (e.g univariate, with splines) to ensure interpretability
- Idea: Use simple linear BL to ensure interpretability (in general also spline BL possible)
- Possible to combine BL of same type (with distinct parameters θ and $\theta^\star$):
- Possible to combine linear BL of same type (with distinct parameters θ and $\theta^\star$):

$$b^{[j]}(\mathbf{x}, \theta) + b^{[j]}(\mathbf{x}, \theta^\star) = b^{[j]}(\mathbf{x}, \theta + \theta^\star)$$

- In each iteration, fit a set of BLs, add best one to model (with step-size ν):

$$\begin{aligned}
\hat{f}^{[1]} &= \hat{f}_0 + \nu b^{[3]}(\mathbf{x}_3, \theta^{[1]}) \\
\hat{f}^{[2]} &= \hat{f}^{[1]} + \nu b^{[3]}(\mathbf{x}_3, \theta^{[2]}) \\
\hat{f}^{[3]} &= \hat{f}^{[2]} + \nu b^{[1]}(\mathbf{x}_1, \theta^{[3]}) \\
&= \hat{f}_0 + \nu \left(b^{[3]}(\mathbf{x}_3, \theta^{[1]} + \theta^{[2]}) + b^{[1]}(\mathbf{x}_1, \theta^{[3]}) \right) \\
&= \hat{f}_0 + \hat{f}_3(\mathbf{x}_3) + \hat{f}_1(\mathbf{x}_1)
\end{aligned}$$

- Final model is additive GAM, we can read off effect curves

MODEL-BASED BOOSTING - LINEAR EXAMPLE

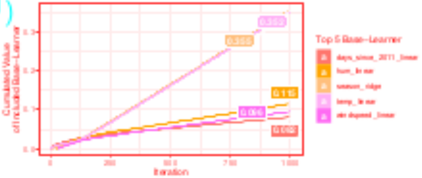
Simple case: Use linear model with single feature (including intercept) as BL

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

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

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	

Relative frequency of selected BLs across iterations



$\Rightarrow$ Converges to solution of LM

- 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 $\theta^\star$):

$$b^{[j]}(\mathbf{x}, \theta) + b^{[j]}(\mathbf{x}, \theta^\star) = b^{[j]}(\mathbf{x}, \theta + \theta^\star)$$

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

$$\hat{f}^{[1]} = \hat{f}_0 + \nu b^{[3]}(\mathbf{x}_3, \theta^{[1]})$$

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 LM
- Early stopping allows feature selection & may 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	

$\Rightarrow$ Converges to solution of LM

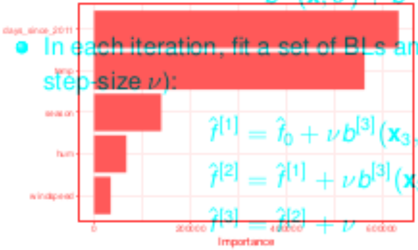
20 iter. with $\nu = 0.1$	Intercept	Weights
days_since_2011	-1210.27	3.3
season	0	WINTER: -276.9 SPRING: 137.6 SUMMER: 112.8 FALL: 20.3
temp	-1118.94	73.2
offset	4504.35	

$\Rightarrow$ 3 BLs selected after 20 iter. (feature selection)

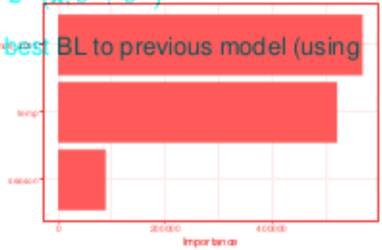
LINEAR EXAMPLE: INTERPRETATION

Feature importance: aggregated change in risk in each iteration per feature

- E.g. iteration 1: days_since_2011 with risk reduction (MSE) of 140,782.94
- For every iteration the change in risk can be attributed to a feature



In-bag-risk: 434,686.0
OOB risk (10-fold CV): 446,450.0



In-bag-risk: 693,505.0
OOB risk (10-fold CV): 705,776.0

$\Rightarrow$ Difference in risk: 258,819.0
Difference in OOB risk: 259,326.0

- 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 $\theta^\star$):

$$b^{[j]}(\mathbf{x},\theta) + b^{[j]}(\mathbf{x},\theta^\star) = b^{[j]}(\mathbf{x},\theta+\theta^\star)$$

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

$$\hat{f}^{[1]} = \hat{f}_0 + \nu b^{[3]}(\mathbf{x}_3,\theta^{[1]})$$
$$\hat{f}^{[2]} = \hat{f}^{[1]} + \nu b^{[3]}(\mathbf{x}_3,\theta^{[2]})$$
$$\hat{f}^{[3]} = \hat{f}^{[2]} + \nu$$
$$= \hat{f}_0 + \nu\left(b^{[3]}(\mathbf{x}_3,\theta^{[1]}+\theta^{[2]}) + \right)$$
$$= \hat{f}_0 + \hat{f}_3(\mathbf{x}_3) +$$

NON-LINEAR EXAMPLE: INTERPRETATION

- Fit model on bike data with different BL types (1000 iter.)
- BLs: linear and centered splines for numeric features, categorical for season

- Recall: Boosting iteratively combines weak base learners (BL) Daniel Schalk et al. 2018
- 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 $\theta^\star$):

$$b^{[j]}(\mathbf{x}, \theta) + b^{[j]}(\mathbf{x}, \theta^\star) = b^{[j]}(\mathbf{x}, \theta + \theta^\star)$$

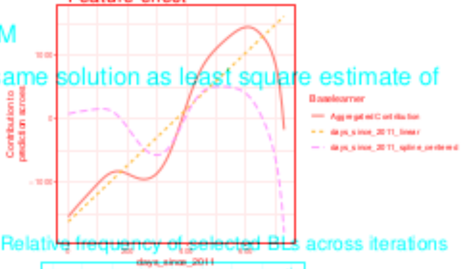
- In each iteration, fit a set of BLs and add the best BL to previous model (using step-size ν):

$$\begin{aligned}
\hat{f}^{[1]} &= \hat{f}_0 + \nu b^{[3]}(\mathbf{x}_3, \theta^{[1]}) \\
\hat{f}^{[2]} &= \hat{f}^{[1]} + \nu b^{[3]}(\mathbf{x}_3, \theta^{[2]}) \\
\hat{f}^{[3]} &= \hat{f}^{[2]} + \nu \\
&= \hat{f}_0 + \nu \left(b^{[3]}(\mathbf{x}_3, \theta^{[1]} + \theta^{[2]}) + \qquad \right) \\
&= \hat{f}_0 + \hat{f}_3(\mathbf{x}_3) +
\end{aligned}$$

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

NON-LINEAR EXAMPLE: INTERPRETATION

- Fit model on bike data with different BL types (1000 iter.)
- BLs: linear and centered splines for numeric features, categorical for season



⇒ In-bag-risk: 250,202.0 ; OOB risk (10-fold CV): 267,497.0 (difference to lin. example: 178,953.0)

⇒ In-bag-risk: 434,686.0 ; OOB risk (10-fold CV): 446,450.0 (previous lin. example with 1000 iter.)

- Feature importance (risk reduction over iter.)
  ⇝ days_since_2011 most important
- Total effect for days_since_2011
  ⇝ Combination of partial effects of linear BL and centered spline BL

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

$b^{(j)}(x_j, \theta) = x_j\theta + \theta_0$ for $j = 1, \dots, p$ ⇝ 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
[illegible]	[illegible]	[illegible]
season	0	[illegible] SPRING: 539.5 SUMMER: [illegible] FALL: 67.2
temp	[illegible]	[illegible]
windspeed	725.70	-56.9
offset	[illegible]	

Relative frequency of selected BLs across iterations

Converges to solution of LM

MODEL-BASED BOOSTING - LINEAR EXAMPLE

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

$$b^{[j]}(x_j, \theta) = x_j\theta + \theta_0 \quad \text{for } j = 1, \ldots p \qquad \rightsquigarrow \text{ordinary linear regression}$$

- 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	

$\Rightarrow$ 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	

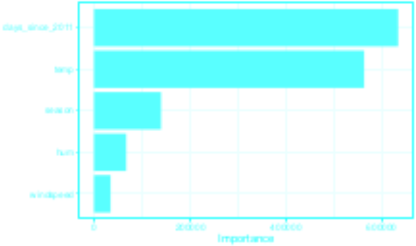
$\Rightarrow$ 3 BLs selected after 20 iter. (feature selection)

LINEAR EXAMPLE: INTERPRETATION

Feature importance: aggregated change in risk in each iteration per feature.

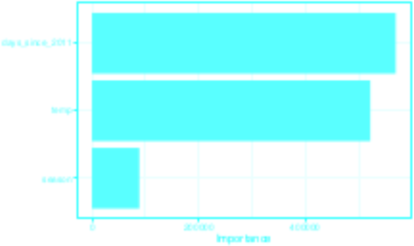
- E.g. iteration 1: `days_since_2011` causes a risk reduction (MSE) of 140,782.94
- For every iteration the change in risk can be attributed to a feature

1000 iterations



Overall risk: 434,686.0
OOB risk (10-fold CV): 446,450.0

20 iterations



Overall risk: 693,505.0
OOB risk (10-fold CV): 705,776.0

$\Rightarrow$ Difference in risk: 258,819.0
Difference in OOB risk: 259,326.0

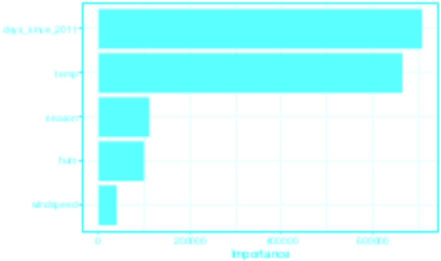
NON-LINEAR EXAMPLE: INTERPRETATION

- Fit model on bike data with different BL types (1000 iter.) ▸ Daniel Schalk et al. 2018
- BLs: linear and centered splines for numeric features, categorical for season

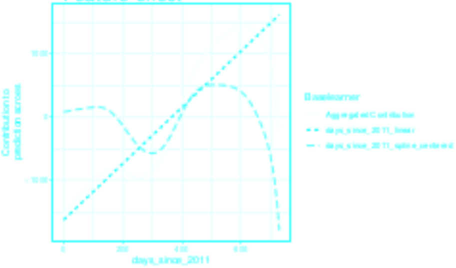
NON-LINEAR EXAMPLE: INTERPRETATION

- Fit model on bike data with different BL types (1000 iter.) ▸ Daniel Schalk et al. 2018
- BLs: linear and centered splines for numeric features, categorical for season

Feature importance



Feature effect



⇒ In-bag-Risk: 250,202.0 ; OOB risk (10-fold CV): 267,497.0 (Difference: 178,953.0)

- Feature importance (risk reduction over iter.) ⇝ `days_since_2011` most important
- Total effect for `days_since_2011`
  ⇝ Combination of partial effects of linear BL and centered spline BL