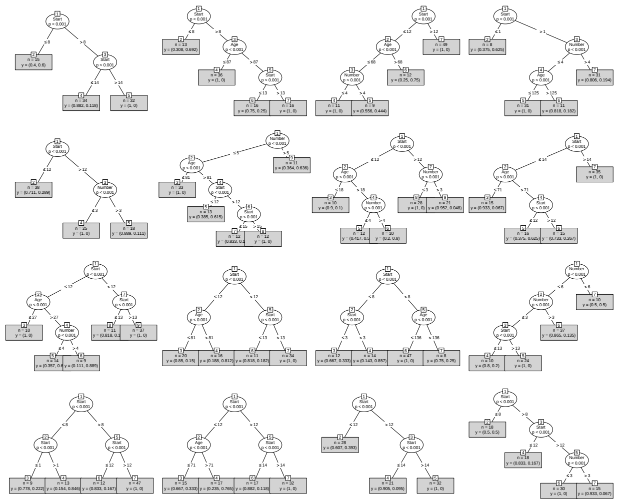
A Comparison of Different Variable Importance Measures

Carolin Strobl (LMU München)

Wien, Jänner 2009

Measuring variable importance in random forests



Measuring variable importance in random forests

Measuring variable importance in random forests

- Gini importance
  mean Gini gain produced by X_j over all trees
  (can be severely biased due to estimation bias and mutiple testing; Strobl et al., 2007)

Measuring variable importance in random forests

- Gini importance
  mean Gini gain produced by X_j over all trees
  (can be severely biased due to estimation bias and mutiple testing; Strobl et al., 2007)
- permutation importance
  mean decrease in classification accuracy after permuting X_j over all trees
  (unbiased when subsampling is used; Strobl et al., 2007)

The permutation importance

within each tree t

$$VI^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$$

$\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i)$ = predicted class before permuting

$\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j})$ = predicted class after permuting X_j

$$\mathbf{x}_{i,\pi_j} = \left(x_{i,1}, \ldots, x_{i,j-1}, \; x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p}\right)$$

Note: $VI^{(t)}(\mathbf{x}_j) = 0$ by definition, if X_j is not in tree t

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

The permutation importance

over all trees:

$$VI(\mathbf{x}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{x}_j)}{ntree}$$

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

What kind of independence corresponds to this kind of permutation?

obs	Y	X_j	Z
1	y_1	$x_{\pi_j(1),j}$	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	$x_{\pi_j(i),j}$	z_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{\pi_j(n),j}$	z_n

$$H_0 : X_j \perp Y, Z \text{ or } X_j \perp Y \wedge X_j \perp Z$$

$$P(Y, X_j, Z) \stackrel{H_0}{=} P(Y, Z) \cdot P(X_j)$$

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

What kind of independence corresponds to this kind of permutation?

the original permutation scheme reflects independence of X_j from both Y and the remaining predictor variables Z

What kind of independence corresponds to this kind of permutation?

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

the original permutation scheme reflects independence of X_j from both Y and the remaining predictor variables Z

$\Rightarrow$ a high variable importance can result from violation of

either one!

Suggestion: Conditional permutation scheme

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

obs	Y	X_j	Z
1	y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
3	y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
27	y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
6	y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
14	y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
33	y_{33}	$x_{\pi_{j\|Z=b}(33),j}$	$z_{33} = b$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$H_0 : X_j \perp Y | Z$$

$$\begin{aligned} & P(Y, X_j | Z) \stackrel{H_0}{=} P(Y|Z) \cdot P(X_j|Z) \\ \text{or} \quad & P(Y|X_j, Z) \stackrel{H_0}{=} P(Y|Z) \end{aligned}$$

Technically

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

- use any partition of the feature space for conditioning

Technically

- use any partition of the feature space for conditioning
- here: use binary partition already learned by tree for each tree
  - determine variables to condition on (via `threshold`)
  - extract their cutpoints
  - generate partition using cutpoints as bisectors

Strobl et al. (2008)

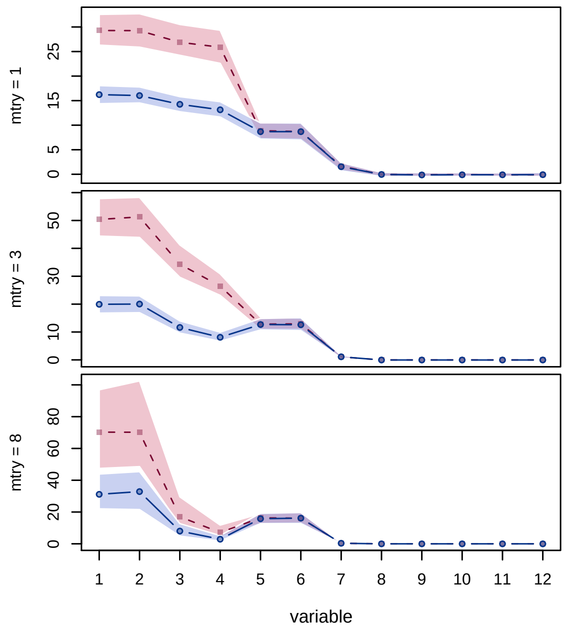
Toy example

spurious correlation between shoe size and reading skills in school-children

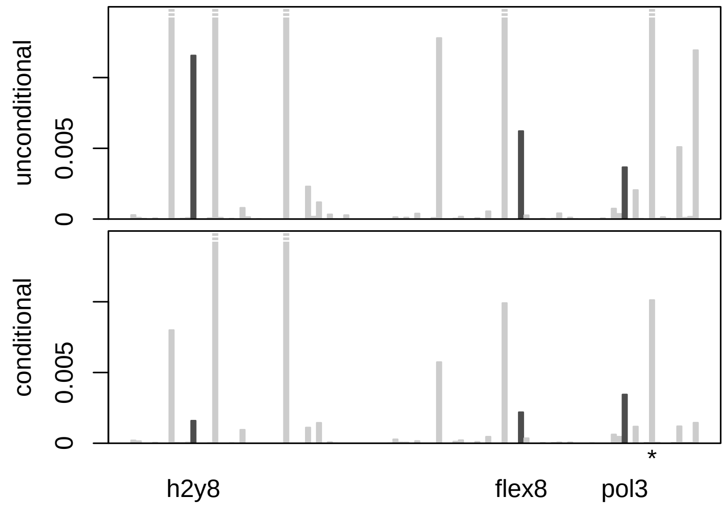
```
> mycf <- cforest(score ~ ., data = readingSkills,
+                 control = cforest_unbiased(mtry = 2))
> varimp(mycf)
nativeSpeaker          age      shoeSize
     12.62926     74.89542      20.01108
> varimp(mycf, conditional = TRUE)
nativeSpeaker          age      shoeSize
    11.808192    46.995336      2.092454
```

from `party 0.9-991`

Simulation results



Peptide-binding data



Other variable importance measures

- partial correlation, standardized beta
  conditional effect of X_j given all other variables
  in the model
- "averaging over orderings"
  - for linear models (`relaimpo`, Grömping, 2006)
    LMG Lindeman, Merenda, and Gold (1980),
    $\approx$ "dominance analysis" Azen and Budescu (2003)
    PMVD Feldman (2005)
  - for GLMs (`hier.part`, Walsh and Nally, 2008)
    "hierarchical partitioning" Chevan and Sutherland (1991)

  R^2 decomposition

Other variable importance measures

- random forest permutation importance
  $\approx$ "averaging over trees"
  unconditional varimp (`randomForest`, `party`, Breiman et al., 2006; Hothorn et al., 2008)
  conditional varimp (`party`, Hothorn et al., 0089)
- elastic net (`elasticnet`, `caret`, Zou and Hastie, 2008; Kuhn, 2008)
  grouping property: correlated predictors get similar (largest) score

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
- *non-negativity*
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG, PMVD
- *non-negativity*
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG, PMVD
- *non-negativity*
  LMG, PMVD, RF varimp (in principle)
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG, PMVD
- *non-negativity*
  LMG, PMVD, RF varimp (in principle)
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
  partial correlation, standardized betas, PMVD,
  RF conditional varimp (in principle), elasticnet?
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG, PMVD
- *non-negativity*
  LMG, PMVD, RF varimp (in principle)
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
  partial correlation, standardized betas, PMVD,
  RF conditional varimp (in principle), elasticnet?
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$
  all

Grömping (2007)

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Simulation study

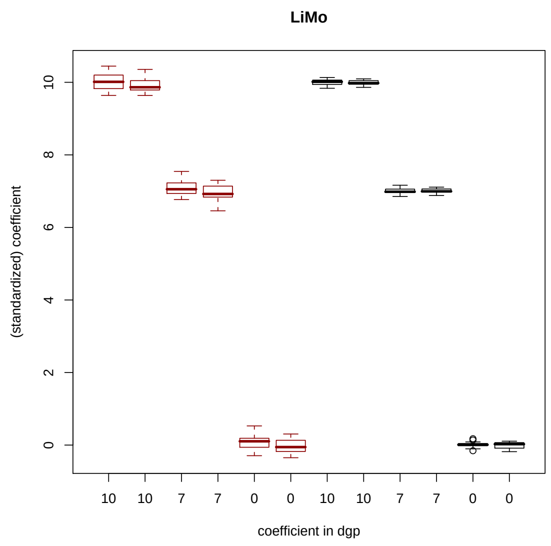
dgp: $y_i = \beta_1 \cdot x_{i,1} + \cdots + \beta_{12} \cdot x_{i,12} + \varepsilon_i,\ \varepsilon_i \overset{i.i.d.}{\sim} N(0,1)$
$X_1, \ldots, X_{12} \sim N(0, \Sigma)$

$$\Sigma = \begin{pmatrix} 1 & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & 1 & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & 1 & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 1 & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 1 & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

X_j	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
β_j	**10**	**10**	**7**	**7**	**0**	**0**	10	10	7	7	0	0

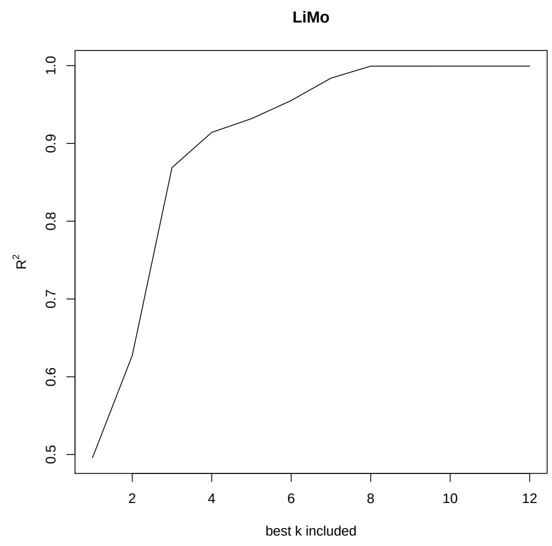
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Linear model



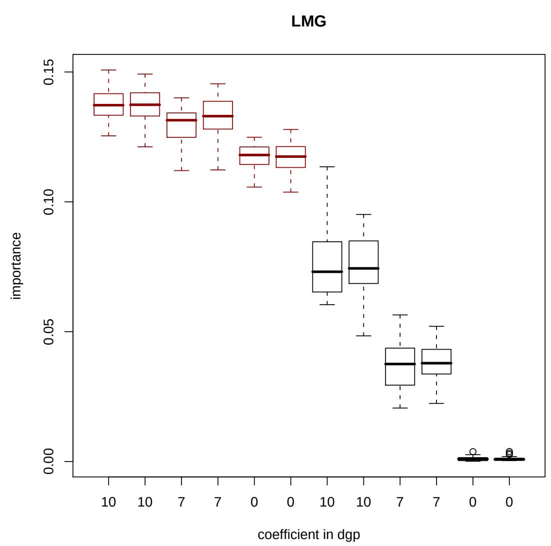
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Linear model



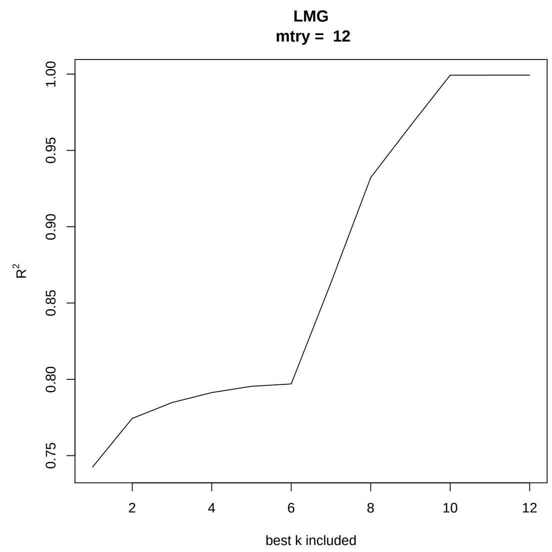
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

LMG



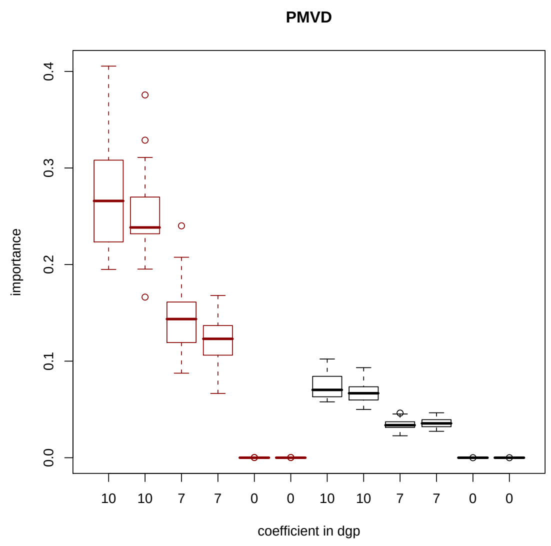
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

LMG

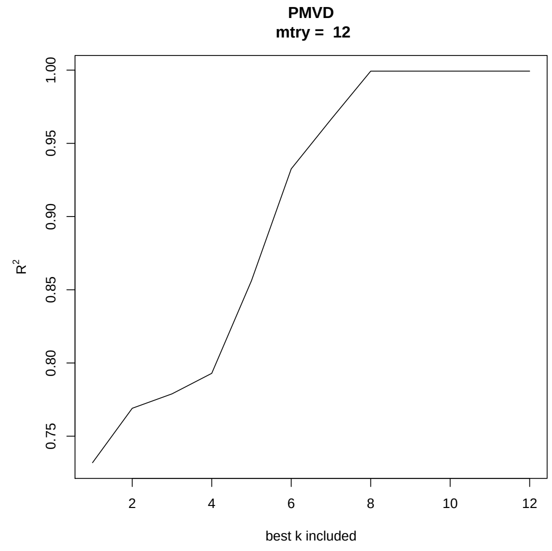


Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

PMVD

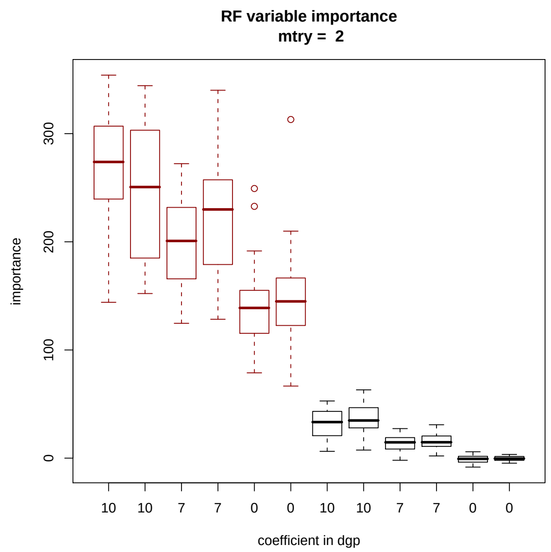


PMVD



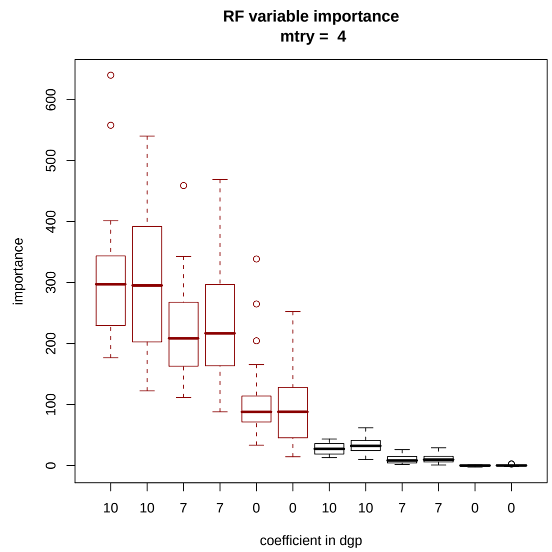
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF unconditional importance



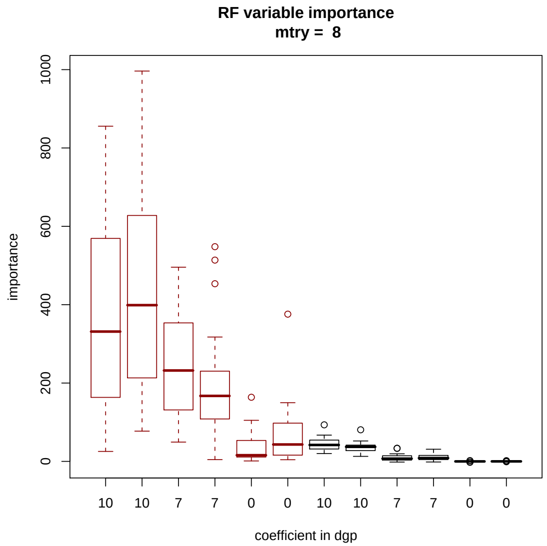
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF unconditional importance



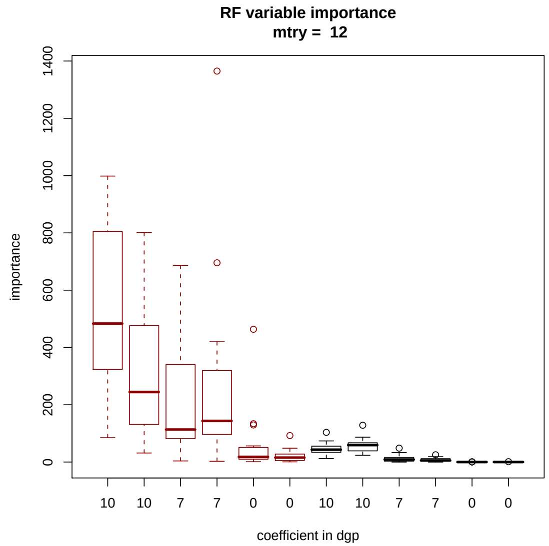
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF unconditional importance



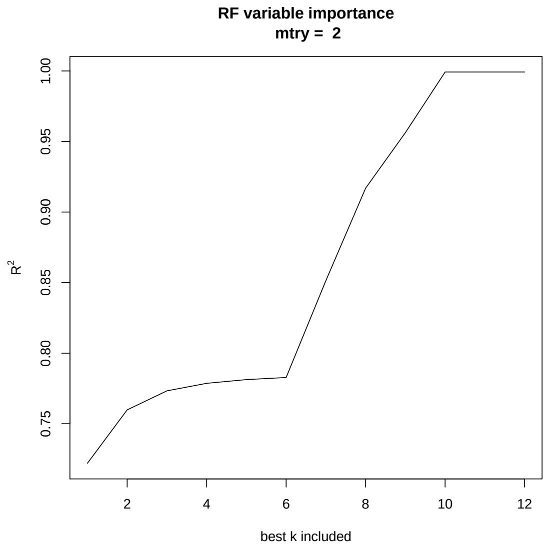
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF unconditional importance



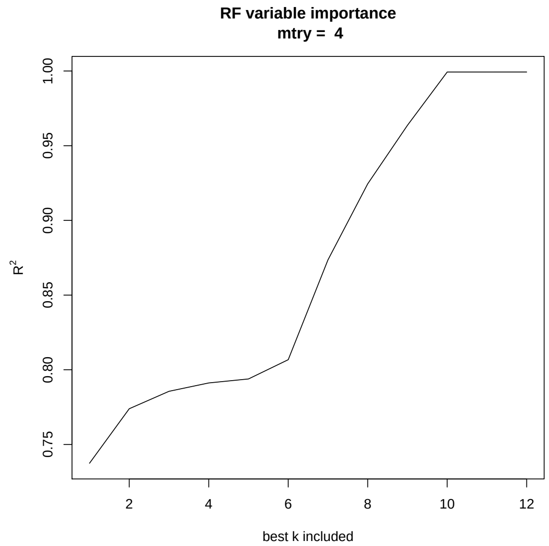
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF unconditional importance



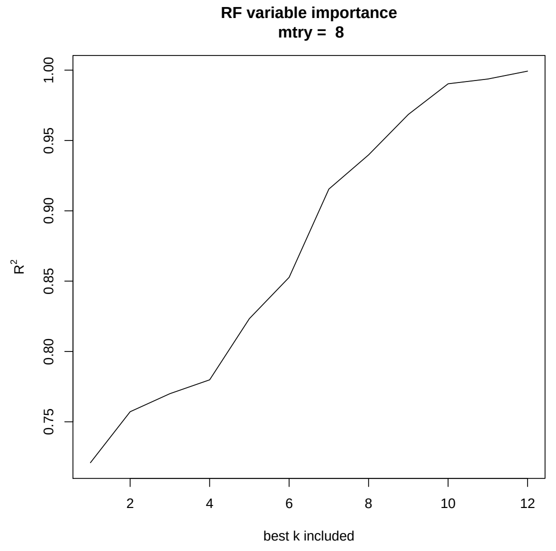
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF unconditional importance

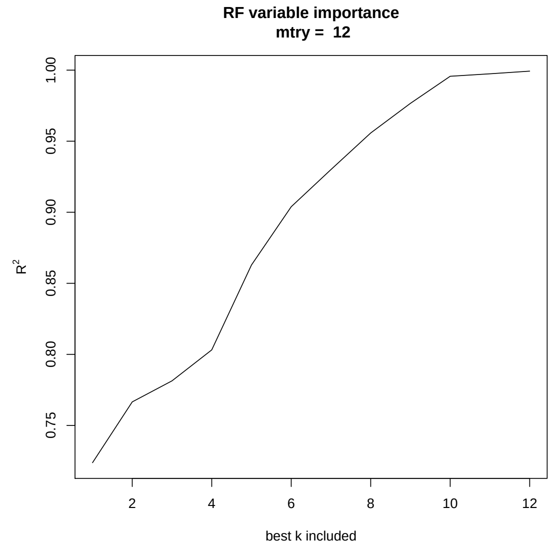


Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

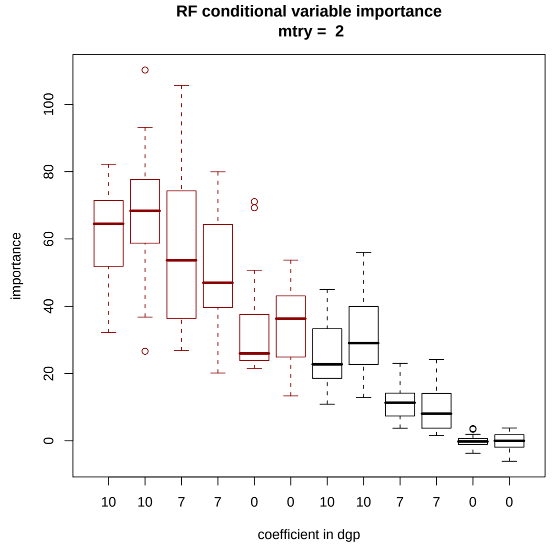
RF unconditional importance



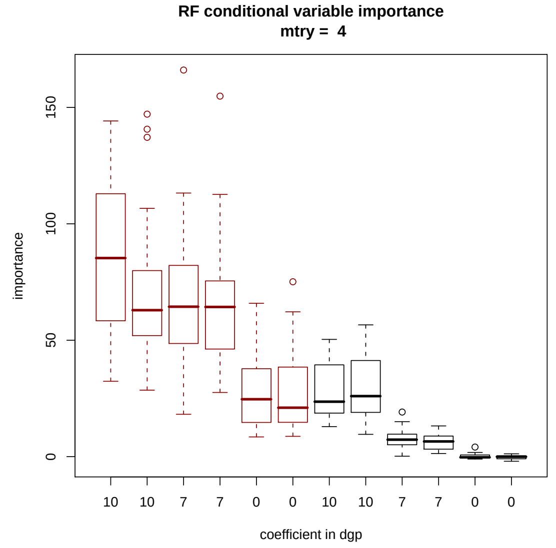
RF unconditional importance



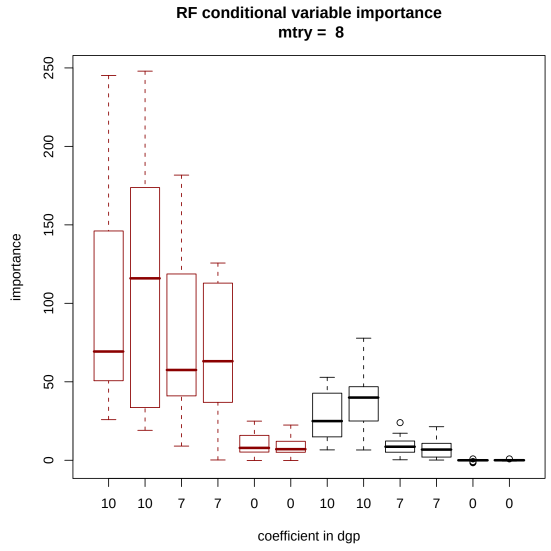
RF conditional importance



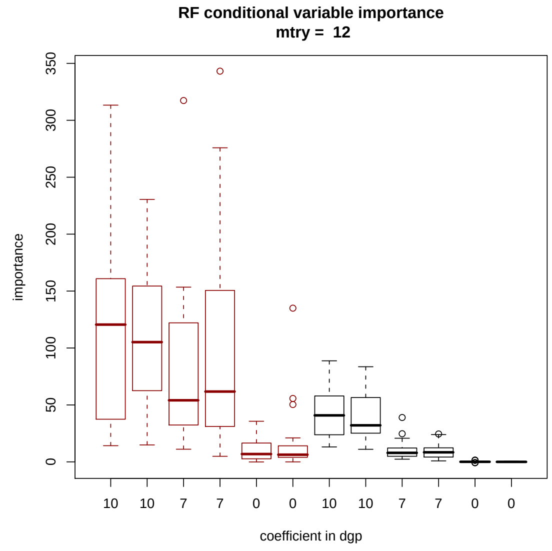
RF conditional importance



RF conditional importance

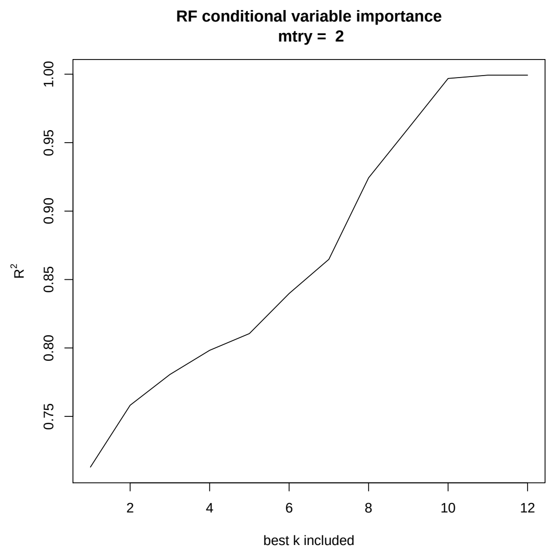


RF conditional importance



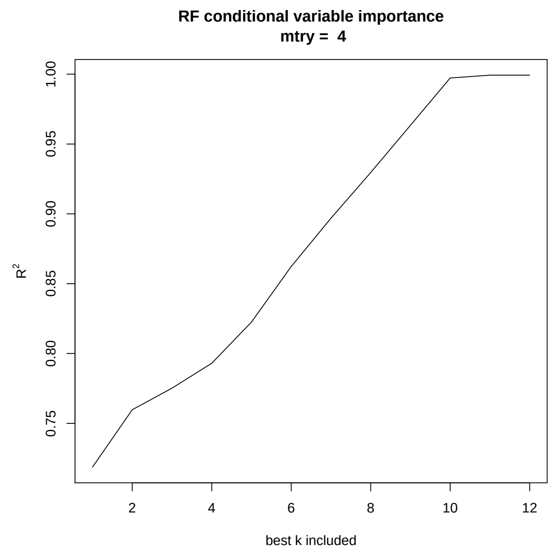
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF conditional importance



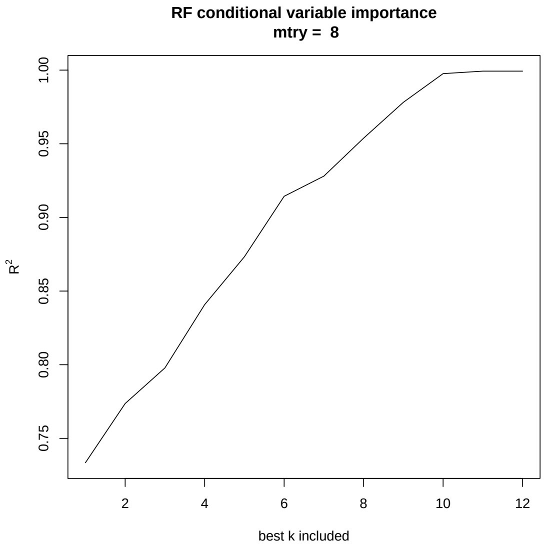
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF conditional importance



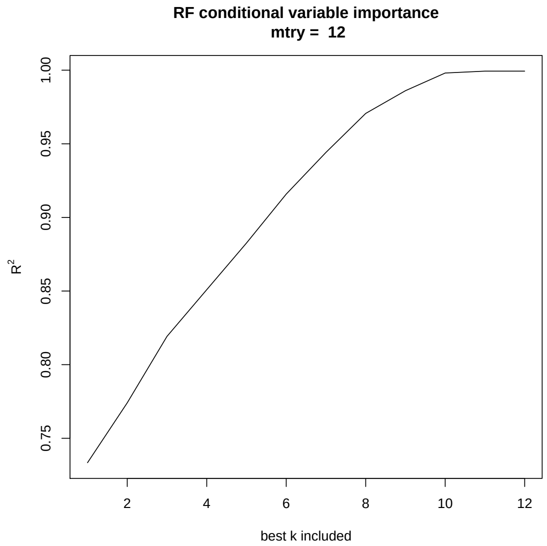
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF conditional importance



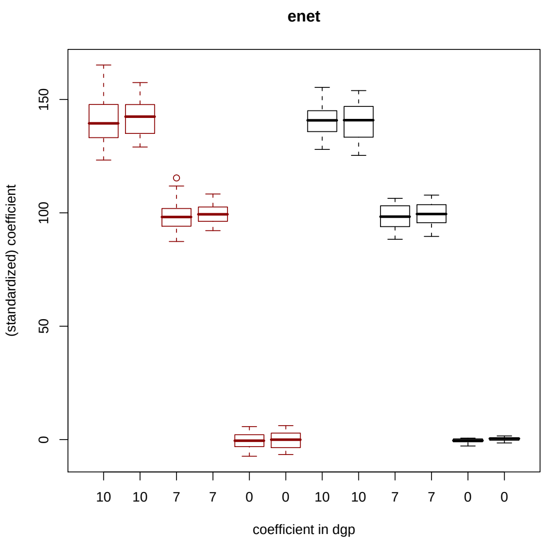
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

RF conditional importance



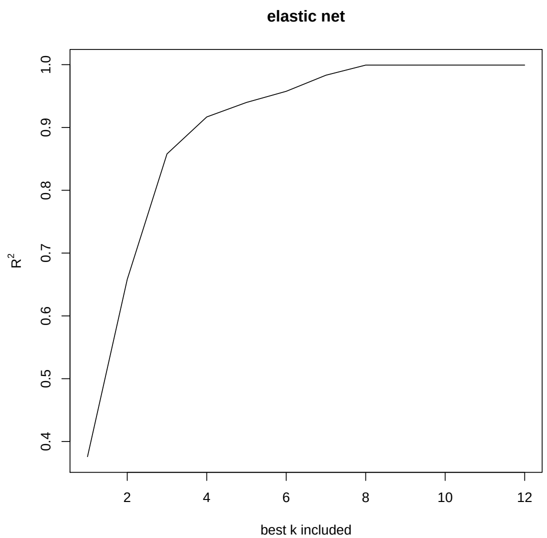
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Elastic net



Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Elastic net



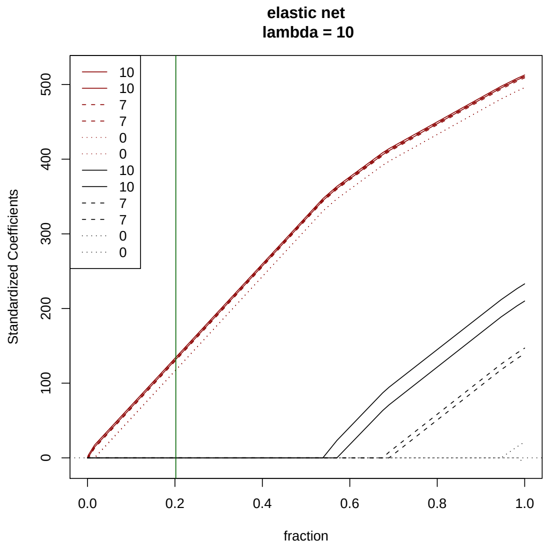
Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Now wait a second...

what about elastic net's grouping property?

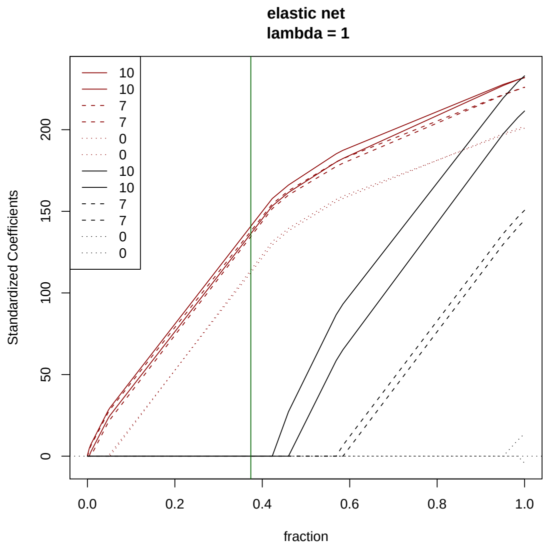
Elastic net

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References



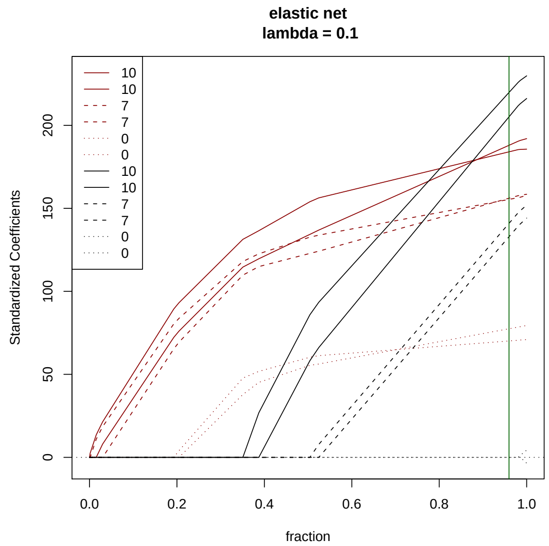
Elastic net

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

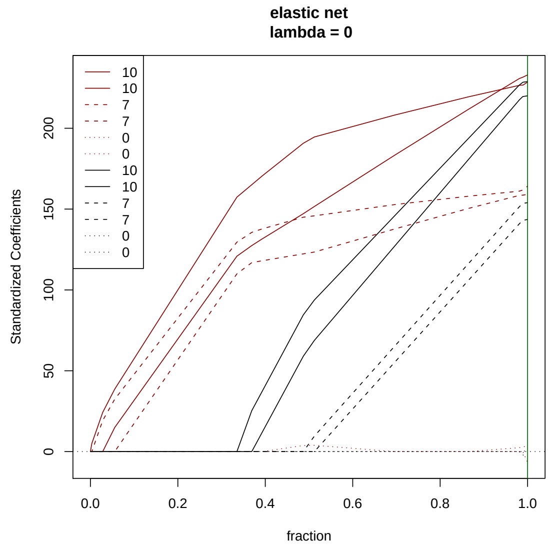


Elastic net

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References



Elastic net



Summary

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Summary

- w.r.t. prediction accuracy: importance measures following the *exclusion* principle rule

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Summary

- w.r.t. prediction accuracy: importance measures following the *exclusion* principle rule standardized betas, PMVD (not quite), RF conditional importance (especially with large `mtry`) and elastic net (tuned!)

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Summary

- w.r.t. prediction accuracy: importance measures following the *exclusion* principle rule
  standardized betas, PMVD (not quite), RF conditional importance (especially with large `mtry`) and elastic net (tuned!)
- RF: not limited to linear model, interactions included, applicable even if $p > 30$

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Summary

- w.r.t. prediction accuracy: importance measures following the *exclusion* principle rule
  standardized betas, PMVD (not quite), RF conditional importance (especially with large `mtry`) and elastic net (tuned!)
- RF: not limited to linear model, interactions included, applicable even if $p > 30$
- if you want elastic net to group: don't tune!?

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Azen, R. and D. V. Budescu (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods 8*(2), 129–48.

Breiman, L., A. Cutler, A. Liaw, and M. Wiener (2006). *Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.5-16.

Chevan, A. and M. Sutherland (1991). Hierarchical partitioning. *The American Statistician 45*(2), 90–96.

Feldman, B. (2005). Relative importance and value. Technical report.

Grömping, U. (2006). *relaimpo: Relative Importance of Regressors in Linear Models*. R package version 2.1.

Grömping, U. (2007). Estimators of relative importance for linear regression based on variance decomposition. *The American Statistician 61*(2), 139–147.

Variable importance in RF
Conditional variable importance in RF
Other variable importance measures
Summary
References

Kuhn, M. (2008). *caret: Classification and Regression Training*. R package version 3.51.

Lindeman, R., P. Merenda, and R. Gold (1980). *Introduction to Bivariate and Multivariate Analysis*. Glenview: Scott Foresman & Co.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9:307*.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8:25*.

Walsh, C. and R. M. Nally (2008). *hier.part: Hierarchical Partitioning*. R package version 1.0-3.

Zou, H. and T. Hastie (2008). *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.0-5.