Unbiased variable importance for random forests

Markus Loecher
Berlin School of Economics and Law

Psychoco 2020

1. Motivation
2. Pitfalls
   - Linear Models
   - Trees
   - Explanatory Overfitting
3. OOB Gini
4. Outlook
5. Appendix: Residualization
   - Statistical View
   - Residualization

Motivation

Data

- Sinking of the Titanic

- Kaggle's NY Rental Listing Inquiries competition:

Two Sigma CONNECT X renthop

- US Federal Sentencing

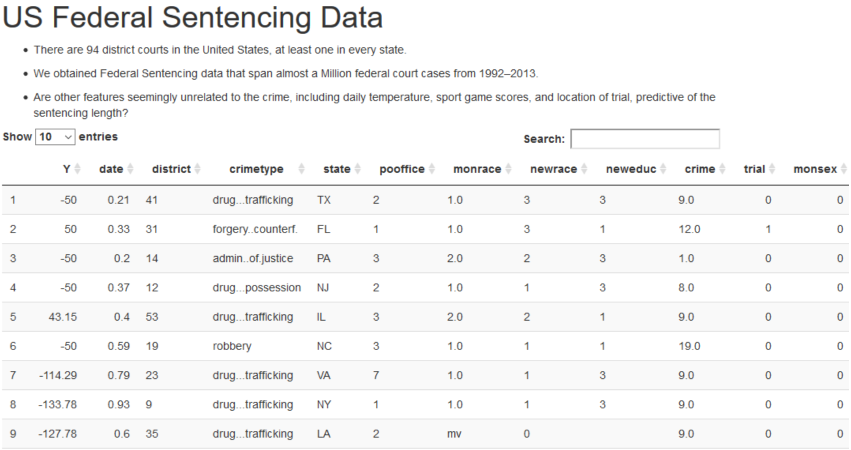
Interest in new rental on RentHop

- bathrooms: number of bathrooms
- bedrooms: number of bathrooms
- latitude
- longitude
- **price**: in USD
- **interest_level**: 'high', 'medium', 'low'
- street_address
- photos: a list of photo links.
- building_id

- created
- description
- display_address
- features: a list of features about this apartment

US District Courts Data

Data Header



US Federal Sentencing Data

- There are 94 district courts in the United States, at least one in every state.
- We obtained Federal Sentencing data that span almost a Million federal court cases from 1992–2013.
- Are other features seemingly unrelated to the crime, including daily temperature, sport game scores, and location of trial, predictive of the sentencing length?

Show 10 entries

Search:

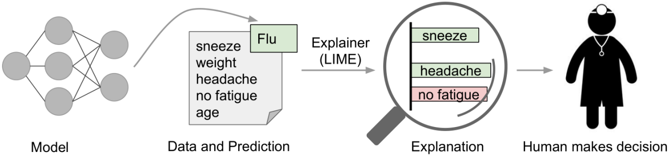
	Y	date	district	crimetype	state	pooffice	monrace	newrace	neweduc	crime	trial	monsex
1	-50	0.21	41	drug...trafficking	TX	2	1.0	3	3	9.0	0	0
2	50	0.33	31	forgery..counterf.	FL	1	1.0	3	1	12.0	1	0
3	-50	0.2	14	admin..of.justice	PA	3	2.0	2	3	1.0	0	0
4	-50	0.37	12	drug...possession	NJ	2	1.0	1	3	8.0	0	0
5	43.15	0.4	53	drug...trafficking	IL	3	2.0	2	1	9.0	0	0
6	-50	0.59	19	robbery	NC	3	1.0	1	1	19.0	0	0
7	-114.29	0.79	23	drug...trafficking	VA	7	1.0	1	3	9.0	0	0
8	-133.78	0.93	9	drug...trafficking	NY	1	1.0	1	3	9.0	0	0
9	-127.78	0.6	35	drug...trafficking	LA	2	mv	0		9.0	0	0

Prediction versus Understanding

- Variables are seldom equally relevant
- Find ranking in “impact”
- Relative importance of regressor variables is an old topic in statistics
- Variable Selection itself a research area
- GDPR

Prediction versus Understanding

- Variables are seldom equally relevant
- Find ranking in "impact"
- Relative importance of regressor variables is an old topic in statistics
- Variable Selection itself a research area
- GDPR
- **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations (**Lime**): Explaining the predictions of any machine learning classifier



The textbook story

California Housing data

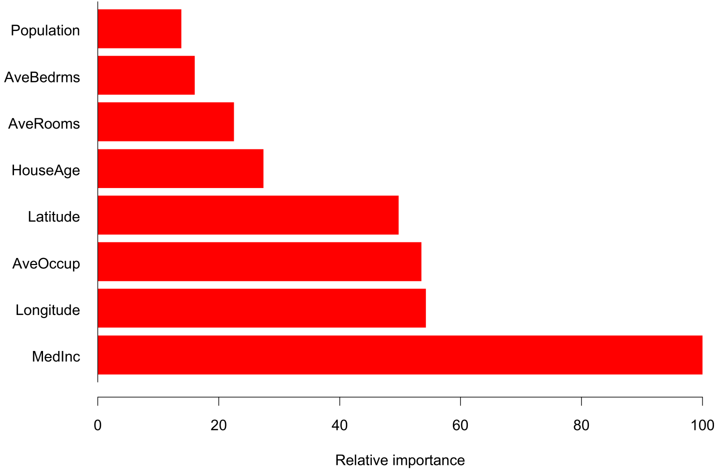


FIGURE 10.14. *Relative importance of the predictors for the California housing data.*

Bonus: Partial Dependence

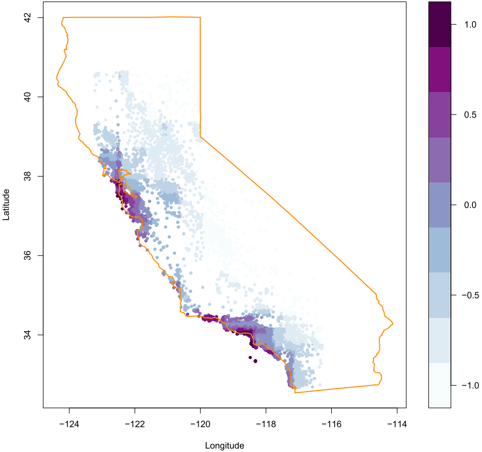


FIGURE 10.17. *Partial dependence of median house value on location in California. One unit is $100,000, at 1990 prices, and the values plotted are relative to the overall median of $180,000.*

Pitfalls

Linear Models

NY rent data set I

“Location, Location, Location, ..?”

	Dependent variable:
	price
longitude	−0.131 (0.134)
latitude	−0.131 (0.134)
bathrooms	0.022* (0.012)
bedrooms	0.017 (0.012)
Constant	0.000 (0.010)
Observations	10,000
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	1.000 (df = 9995)
F Statistic	3.173** (df = 4; 9995)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

VI, Linear Models

- *"Variable importance is not very well defined as a concept. There is no theoretically defined variable importance metric..."*
- Change in R^2 when the variable is added to the model last
- Average order-dependent R^2 allocations over all p! orderings (**LMG**)

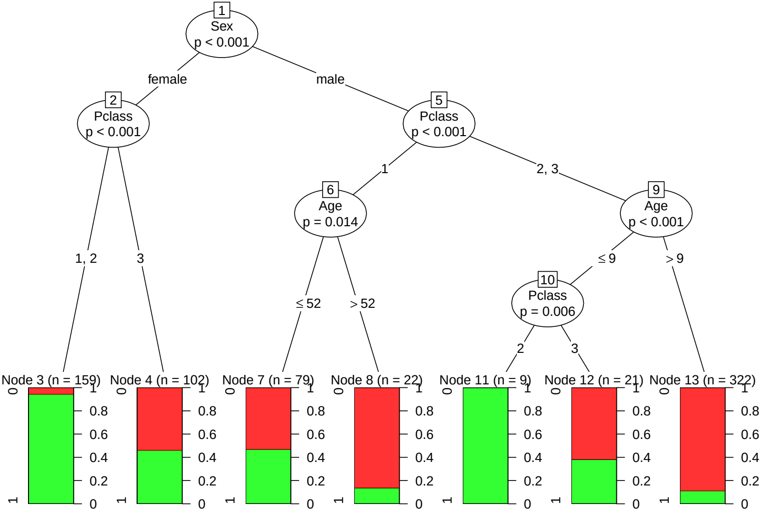
+ Direction/sign of contribution
+ Uncertainty "for free"
+ Easy to understand !?
- Marginal versus conditional
- Confounding effects
- Slave to linearity
- Interactions must be coded apriori

Trees

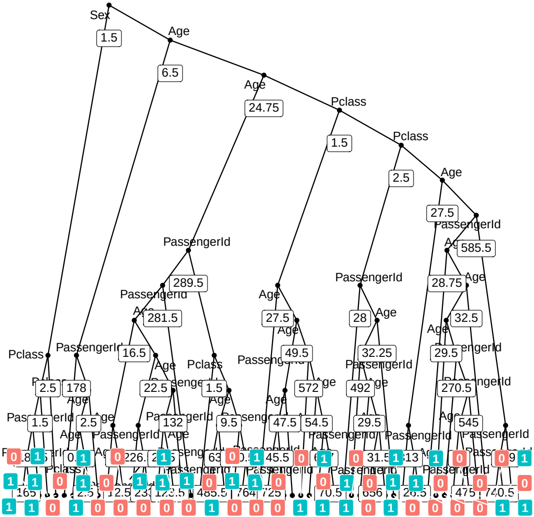
Trees details

- **Greedy**: At each split we minimize *squared error* or *node impurity*
- **All Interactions**: data “thin out” exponentially fast.
- **Piecewise Constant**: no smoothness, inferior for regression.
- **Model complexity**: depth of tree, typically single pruned trees
- **Boosting**: many shallow trees sequentially minimize loss

Titanic Tree



Deep Tree


Sex
1.5
Age
6.5
Age
24.75
Pclass
1.5
Pclass
2.5
Age
27.5
PassengerId
585.5
PassengerId
289.5
PassengerId
281.5
Age
16.5
Age
27.5
Age
49.5
PassengerId
28
Age
32.25
Age
28.75
Age
32.5
Age
29.5
PassengerId
Pclass
2.5
PassengerId
178
22.5
Pclass
1.5
572
492
270.5
1.5
2.5
132
9.5
47.5
54.5
29.5
545
PassengerId
45.5
31.5
0
1

Random Forests

- Many **deep** trees grown in parallel on **bootstrapped** samples.
- **Column sampling** leads to additional parameter *mtry*.

- The tuning parameter *mtry* can have profound effects on prediction quality as well as the variable importance measures outlined below.
- RF rarely suffer from *prediction overfitting*
- Not true for *explanatory overfitting*

Variable Importance I

- **Gini importance**: the mean decrease in impurity $I(= 2\hat{p} \cdot (1 - \hat{p}))$ of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors)
- For a single parent node m:

$$\Delta I(m) = I(m) - \frac{N_{left}}{N_m} \cdot I_{left} - \frac{N_{right}}{N_m} \cdot I_{right}$$

For the tree/forest:

$$MDI(j,t) = \sum_{t=1}^{J-1} \frac{N_m}{N} \Delta I(m)_t (v(t) = j), \Rightarrow MDI(j) = \frac{1}{M} \sum_{t=1}^{M} MDI(j,t)$$

as a measure of relevance for each predictor variable X_j.The sum is over the $J - 1$ internal nodes of the tree.

Bootstrap: OOB

Due to the CART bootstrap row sampling, 36.8% of the observations are (on average) not used for an individual tree; those **“out of bag” (OOB)** samples can serve as a validation set to estimate the test error, e.g.:

$$E\left(Y - \hat{Y}\right)^2 \approx OOB_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{\hat{y}}_{i,OOB}\right)^2$$

where $\overline{\hat{y}}_{i,OOB}$ is the average prediction for the *i*th observation from those trees for which this observation was OOB.

Variable Importance II

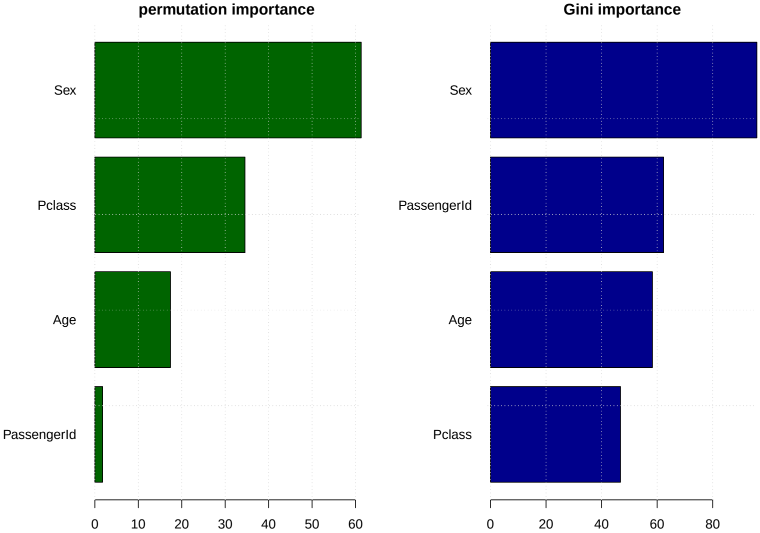
The default method to compute variable importance is the *mean decrease in impurity* (or *gini importance*) mechanism: At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable. Note that this measure is quite like the R^2 in regression on the training set.

The widely used alternative *reduction in MSE when permuting a variable* as a measure of variable importance or short **permutation importance** is defined as follows:

$$\mathrm{VI} = OOB_{MSE,perm} - OOB_{MSE}$$

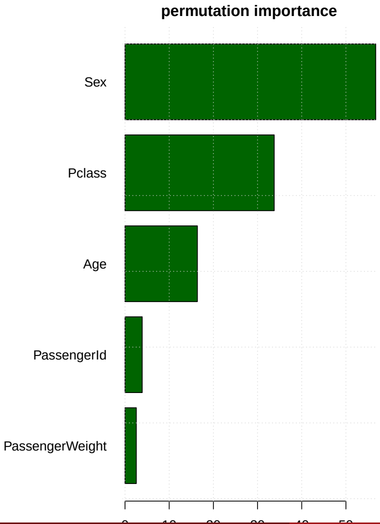
Explanatory Overfitting

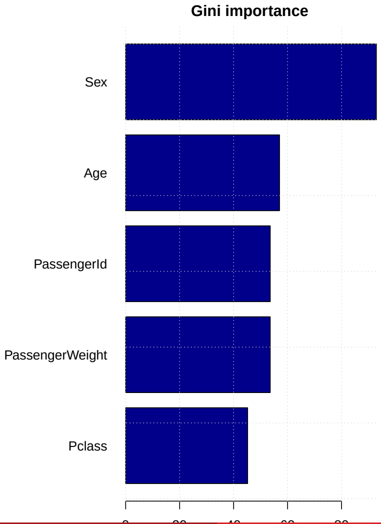
Gini importance can be highly misleading



Noise Feature

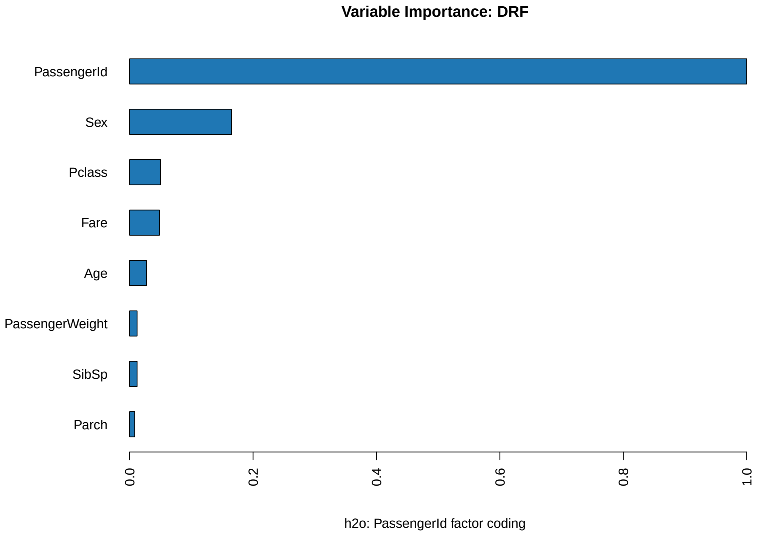
Let us go one step further and add a Gaussian noise feature, which we call PassengerWeight:



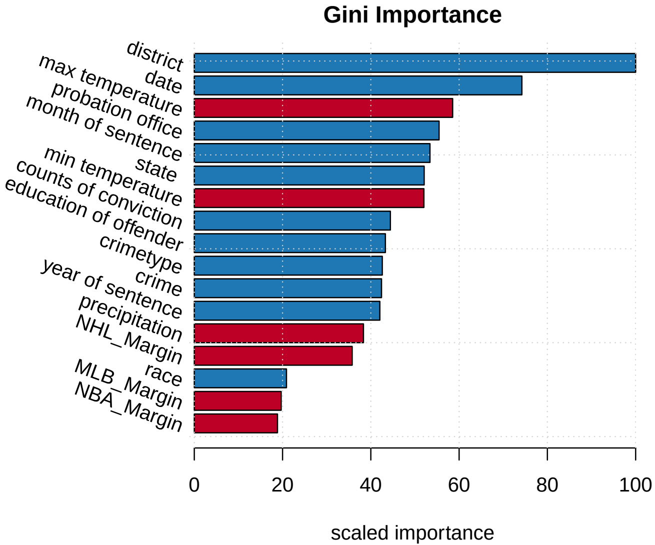


Categorical Features

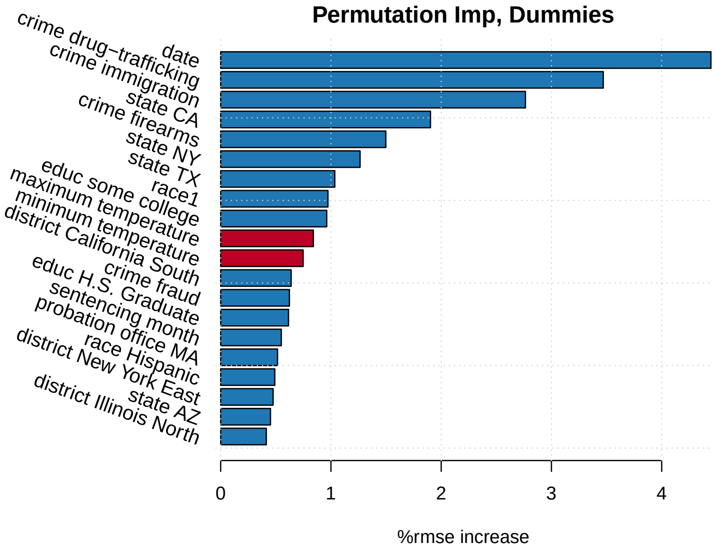
Coding passenger ID as factor makes matters worse:



US District Courts VI-Gini



US District Courts VI-OOB I



Prior Work

- Distribution of *maximally selected statistics*: Strobl et al. (2007), Shih and Tsai (2004), Shih (2004)
- Several authors (Loh and Shih, 1997, Hothorn et al., 2006) argue that the criterion for split variable and split point selection should be separated.
- Conditional Importance: Strobl et al. (2008)
- Permuting the response variable: Altmann et al. (2010), Hapfelmeier and Ulm (2013)
- Adding pseudo variables to a dataset, which are permuted versions of the original variables (Sandri and Zuccolotto, 2008, Nembrini et al., 2018). R package *ranger* (Wright and Ziegler, 2015)
- Cross Validation Janitza et al. (2018)
- Using OOB samples to compute a debiased version of the Gini importance (Li et al., 2019, Zhou and Hooker, 2019)

OOB Gini

OOB impurity reduction

Main idea: increase impurity $I(m)$ for node m by a penalty that is proportional to the difference $\Delta = (\hat{p}_{OOB} - \hat{p}_{inbag})^2$.

$$PG_{OOB}^{\alpha,\lambda} = \alpha \cdot I_{OOB} + (1-\alpha) \cdot I_{inbag} + \lambda \cdot (\hat{p}_{OOB} - \hat{p}_{inbag})^2$$

OOB impurity reduction

Main idea: increase impurity $I(m)$ for node m by a penalty that is proportional to the difference $\Delta = (\hat{p}_{OOB} - \hat{p}_{inbag})^2$.

$$PG_{OOB}^{\alpha,\lambda} = \alpha \cdot I_{OOB} + (1-\alpha) \cdot I_{inbag} + \lambda \cdot (\hat{p}_{OOB} - \hat{p}_{inbag})^2$$

$$\begin{cases} PG_{OOB}^{(0)} & = 2 \cdot \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) \\ PG_{OOB}^{(1)} & = 2 \cdot \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + (\hat{p}_{OOB} - \hat{p}_{inbag})^2 \\ PG_{OOB}^{(2)} & = \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + \hat{p}_{inbag} \cdot (1 - \hat{p}_{inbag}) + (\hat{p}_{OOB} - \hat{p}_{inbag})^2 \\ PG_{OOB}^{(3)} & = \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + \hat{p}_{inbag} \cdot (1 - \hat{p}_{inbag}) + \frac{1}{2}(\hat{p}_{OOB} - \hat{p}_{inbag})^2 \end{cases}$$

OOB impurity reduction

Main idea: increase impurity $I(m)$ for node m by a penalty that is proportional to the difference $\Delta = (\hat{p}_{OOB} - \hat{p}_{inbag})^2$.

$$PG_{OOB}^{\alpha,\lambda} = \alpha \cdot I_{OOB} + (1-\alpha) \cdot I_{inbag} + \lambda \cdot (\hat{p}_{OOB} - \hat{p}_{inbag})^2$$

$$\begin{cases} PG_{OOB}^{(0)} & = 2 \cdot \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) \\ PG_{OOB}^{(1)} & = 2 \cdot \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + (\hat{p}_{OOB} - \hat{p}_{inbag})^2 \\ PG_{OOB}^{(2)} & = \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + \hat{p}_{inbag} \cdot (1 - \hat{p}_{inbag}) + (\hat{p}_{OOB} - \hat{p}_{inbag})^2 \\ PG_{OOB}^{(3)} & = \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + \hat{p}_{inbag} \cdot (1 - \hat{p}_{inbag}) + \frac{1}{2}(\hat{p}_{OOB} - \hat{p}_{inbag})^2 \end{cases}$$

Note that $PG_{OOB}^{(2)} = \hat{p}_{OOB} + \hat{p}_{inbag} - 2\hat{p}_{OOB} \cdot \hat{p}_{inbag}$ and $E(PG_{OOB}^{(2)}) = 0$!

OOB impurity reduction

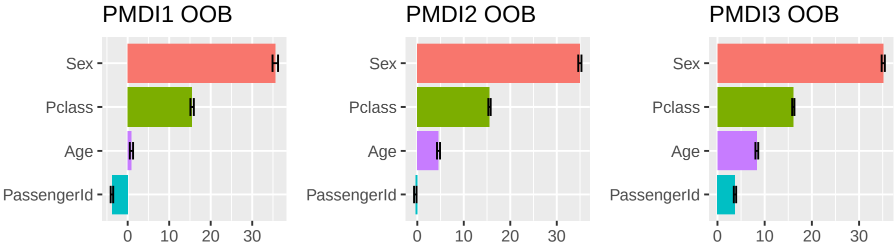
Main idea: increase impurity $I(m)$ for node m by a penalty that is proportional to the difference $\Delta = (\hat{p}_{OOB} - \hat{p}_{inbag})^2$.

$$PG_{OOB}^{\alpha,\lambda} = \alpha \cdot I_{OOB} + (1-\alpha) \cdot I_{inbag} + \lambda \cdot (\hat{p}_{OOB} - \hat{p}_{inbag})^2$$

$$\begin{cases} PG_{OOB}^{(0)} & = 2 \cdot \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) \\ PG_{OOB}^{(1)} & = 2 \cdot \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + (\hat{p}_{OOB} - \hat{p}_{inbag})^2 \\ PG_{OOB}^{(2)} & = \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + \hat{p}_{inbag} \cdot (1 - \hat{p}_{inbag}) + (\hat{p}_{OOB} - \hat{p}_{inbag})^2 \\ PG_{OOB}^{(3)} & = \hat{p}_{OOB} \cdot (1 - \hat{p}_{OOB}) + \hat{p}_{inbag} \cdot (1 - \hat{p}_{inbag}) + \frac{1}{2}(\hat{p}_{OOB} - \hat{p}_{inbag})^2 \end{cases}$$

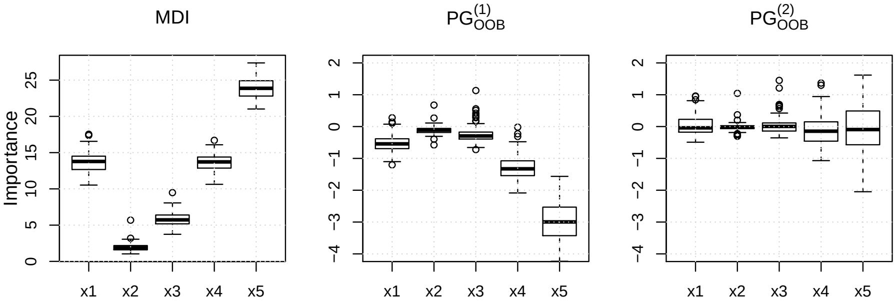
Note that $PG_{OOB}^{(2)} = \hat{p}_{OOB} + \hat{p}_{inbag} - 2\hat{p}_{OOB} \cdot \hat{p}_{inbag}$ and $E(PG_{OOB}^{(2)}) = 0$!
Now a package rfVarImpOOB on CRAN

OOB Titanic


PMDI1 OOB
PMDI2 OOB
PMDI3 OOB
Sex
Pclass
Age
PassengerId
0
10
20
30

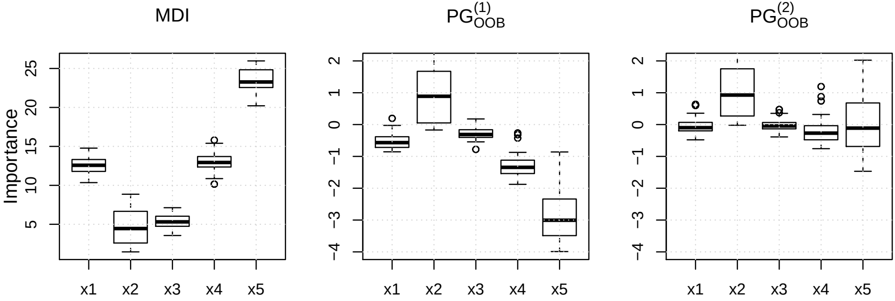
Simulated Data, Null

X_1 is continuous, while the other predictor variables $X_2, \ldots, X_5$ are multinomial with $2, 4, 10, 20$ categories, respectively. (sample size, $\text{n} = 120$).



Simulated Data, Power Study

Response is a binomial process with probabilities that depend on the value of x_2, namely $P(y = 1|X_2 == 1) = 0.35, P(y = 1|X_2 == 2) = 0.65$



Noisy feature identification

The data has 1000 samples with 50 features. All features are discrete, with the jth feature containing $j+1$ distinct values $0, 1, \ldots, j$. We randomly select a set S of 5 features from the first ten as relevant features. The remaining features are noisy features. All samples are i.i.d. and all features are independent. We generate the outcomes using the following rule:

$$P(Y = 1|X) = \text{Logistic}(\frac{2}{5} \sum_{j \in S} x_j/j - 1)$$

$PG^{(0)}_{oob}$	$PG^{(1)}_{oob}$	$PG^{(2)}_{oob}$	$PG^{(3)}_{oob}$	AIR	MDA	MDI
0.28	**0.91**	0.78	0.37	0.68	0.65	0.10

Table 1: Average AUC scores for noisy feature identification. MDA = permutation importance, MDI = (default) Gini impurity

Outlook

Summary/Recommendations

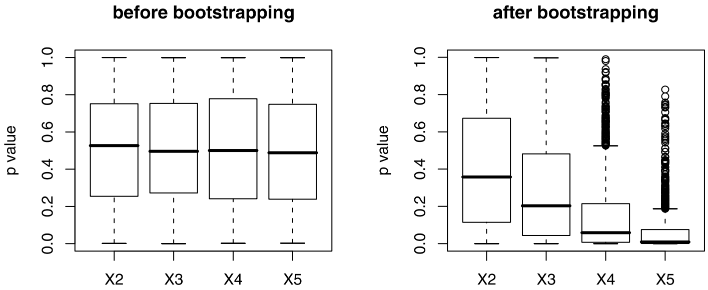
- RF default importance not reliable
- Permutation importance very expensive
- Everything else is evaluated on hold-out-set.
- Boosting or extremely randomized trees for VI
- Careful about conditional versus marginal importance: Residualizing seems to work
- Social Sciences, Bioinformatics and Economics

https://explained.ai/rf-importance

Appendix: Residualization

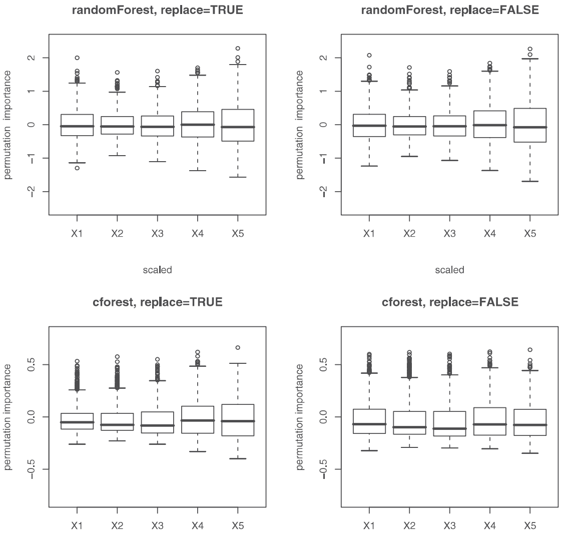
Statistical View

Effects induced by bootstrapping



Distribution of the p values of χ^2 tests of each categorical variable $X_2, \ldots, X_5$ and the binary response for the null case simulation study, where none of the predictor variables is informative. (Multinomials $X_{2,3,4,5} \sim M(2, 4, 10, 20)$, $n = 120$)

Solution: Permutation Importance?



Residualization

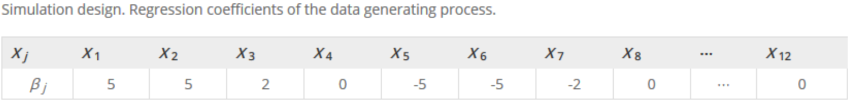
Conditional vs. Marginal Importance

- $H_0 : X_j \perp Y \wedge X_j \perp Z$

Conditional vs. Marginal Importance

- $H_0 : X_j \perp Y \wedge X_j \perp Z$
- Linear Model for sake of illustration

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{12} x_{12} + u$$

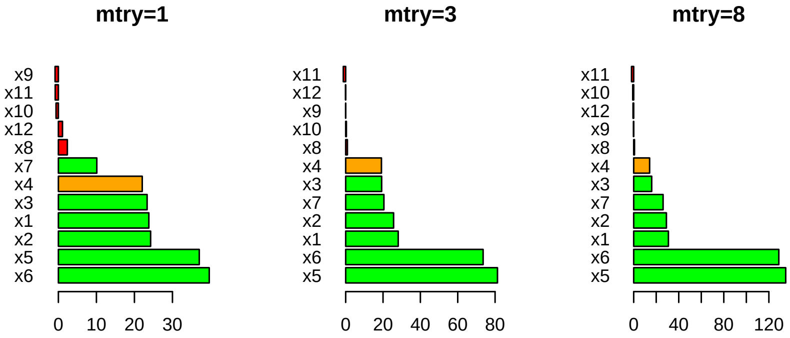


Simulation design. Regression coefficients of the data generating process.

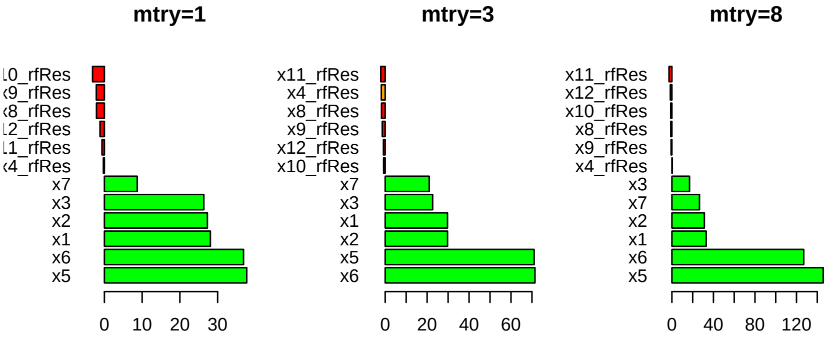
x_j	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	...	x_{12}
β_j	5	5	2	0	-5	-5	-2	0	...	0

- $X_1, \ldots, X_{12} \sim N(0, \Sigma)$ with $\sigma_{j,j} = 1$, $\sigma_{j,j'} = 0.9$ for $j \neq j' \leq 4$, else $\sigma_{j,j'} = 0$.

Conditional"izing" I



Conditional“izing” II



References

André Altmann, Laura Toloși, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10): 1340–1347, 04 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq134. URL https://doi.org/10.1093/bioinformatics/btq134.

Alexander Hapfelmeier and Kurt Ulm. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60:50–69, 2013.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15 (3):651–674, 2006.

Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 12(4):885–915, 2018.

Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased mdi feature importance measure for random forests. *arXiv preprint arXiv:1906.10845*, 2019.

Wei-Yin Loh and Yu-Shan Shih. Split selection methods for classification trees. *Statistica sinica*, pages 815–840, 1997.

Stefano Nembrini, Inke R König, and Marvin N Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

Marco Sandri and Paola Zuccolotto. A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628, 2008.

Y-S Shih. A note on split selection bias in classification trees. *Computational statistics & data analysis*, 45(3):457–466, 2004.

Yu-Shan Shih and Hsin-Wen Tsai. Variable selection bias in regression trees with constant fits. *Computational statistics & data analysis*, 45(3):595–607, 2004.

Carolin Strobl, Anne-Laure Boulesteix, and Thomas Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52(1):483–501, 2007.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1): 307, 2008.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

Zhengze Zhou and Giles Hooker. Unbiased measurement of feature importance in tree-based methods. *arXiv preprint arXiv:1903.05179*, 2019.