Debiasing SHAP scores with sample splitting

Markus Loecher

Hochschule für Wirtschaft und Recht Berlin

d-cube Berlin

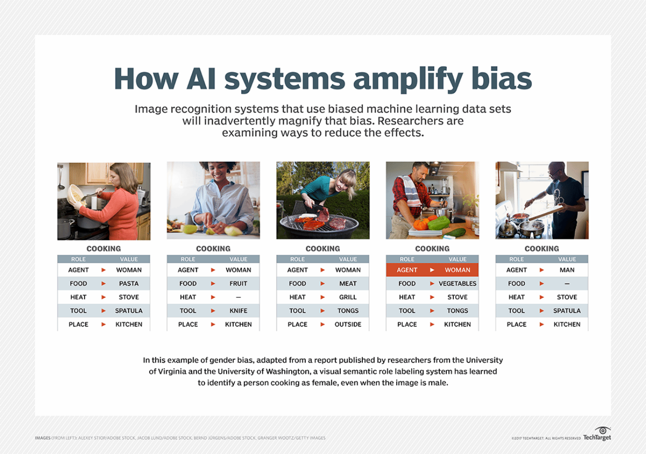
Institute for Data-Driven Digital Transformation

CMStats 2023

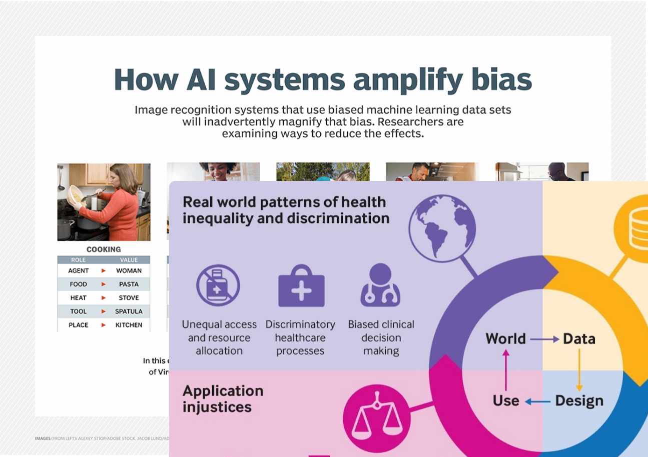
4 days until Winter Solstice

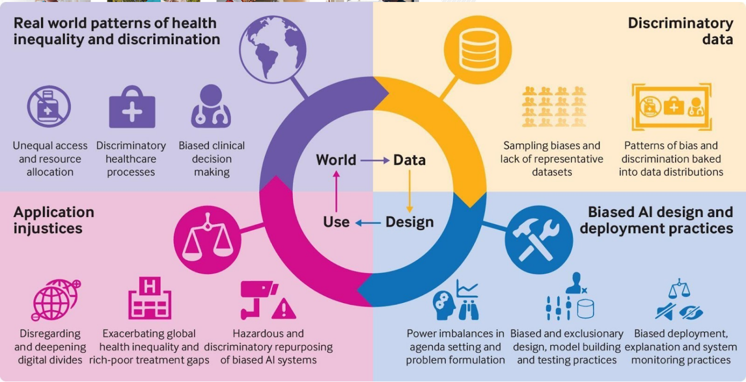
Motivation

Statistical vs. Ethical Bias



Statistical vs. Ethical Bias





US District Courts Data

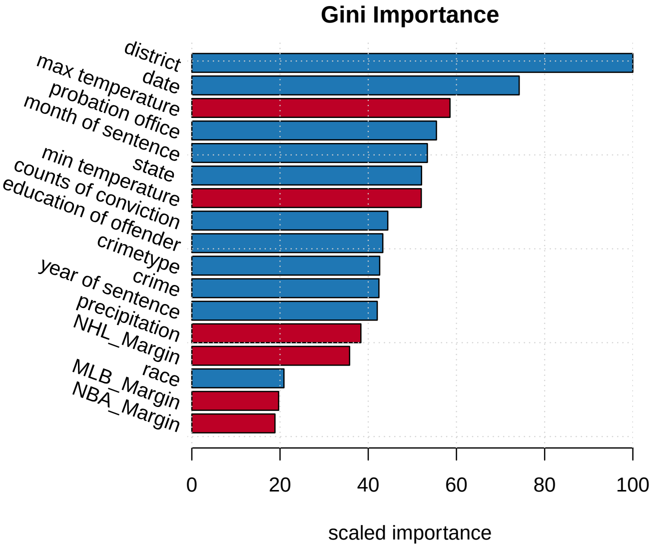
US Federal Sentencing Data

- 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

US District Courts



Goals of this talk

- Review basics of tree ensembles and relevant feature attribution schemes
  - **M**ean **D**ecrease **I**mpurity (MDI)
  - *SHapley Additive exPlanation* (SHAP)
- MDI as well as SHAP values are susceptible to "overfitting" to the training data.
- **Penalized Impurity** measures debias MDI by including OOB samples.
- Extension to SHAP

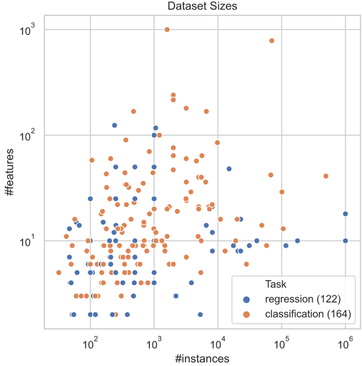
Data

- Sinking of the Titanic

- Simulations

Predictor variables	
X_1	$\sim N(0,1)$
X_2	$\sim M(2)$
X_3	$\sim M(4)$
X_4	$\sim M(10)$
X_5	$\sim M(20)$

- PMLB



Trees vs. Neural Nets I

ARTICLES
https://doi.org/10.1038/s42256-019-0138-9

nature machine intelligence

From local explanations to global understanding with explainable AI for trees

Scott M. Lundberg[1,2], Gabriel Erion[2,3], Hugh Chen[2], Alex DeGrave[2,3], Jordan M. Prutkin[4], Bala Nair[5,6], Ronit Katz[7], Jonathan Himmelfarb[7], Nisha Bansal[7] and Su-In Lee[2*]

Tree-based machine learning models such as random forests, decision trees and gradient boosted trees are popular nonlinear predictive models, yet comparatively little attention has been paid to explaining their predictions. Here we improve the

Advantages of tree-based models

Tree-based models can be more accurate than neural networks in many applications. While deep learning models are more appropriate in fields such as image recognition, speech recognition

and natural language processing, tree-based models consistently outperform standard deep models on tabular-style datasets, where features are individually meaningful and lack strong multiscale temporal or spatial structures[18] (Supplementary Results 1). The

Trees vs. Neural Nets II

Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
ISIR, CNRS, Sorbonne University

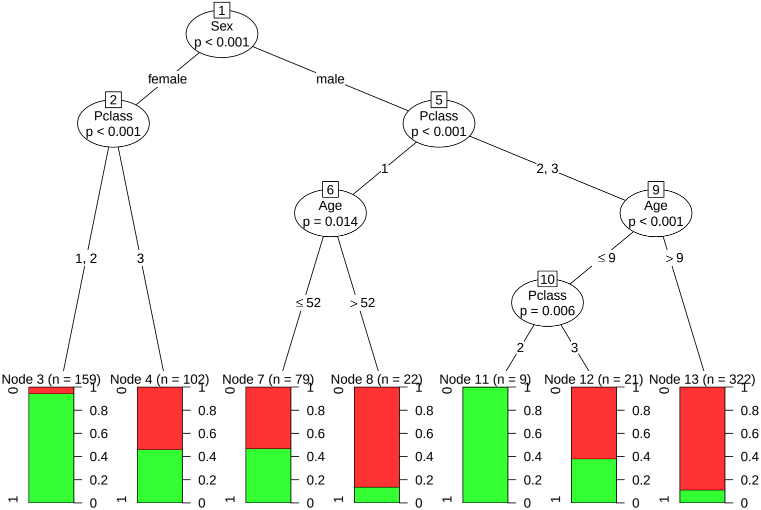
Gaël Varoquaux
Soda, Inria Saclay

Abstract

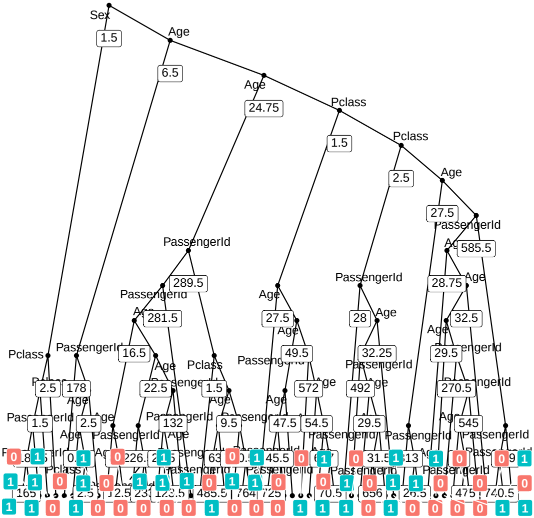
While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data ($\sim$10K samples) even without accounting for their superior speed. To understand this gap, we conduct an

Trees

Shallow trees are **interpretable models**.



Deep Tree



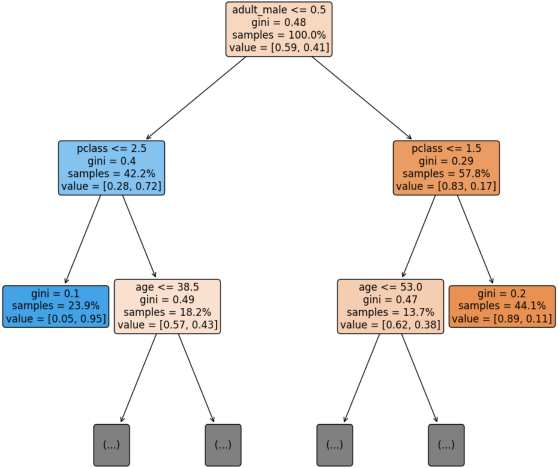
Sex
1.5
Age
6.5
Age
24.75
Pclass
1.5
Pclass
2.5
Age
27.5
PassengerId
585.5
PassengerId
289.5
281.5
16.5
22.5
28.75
32.5
29.5
270.5
545
27.5
49.5
572
47.5
54.5
28
32.25
492
29.5
Pclass
2.5
178
1.5
2.5
132
9.5
45.5

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 Importances: MDI

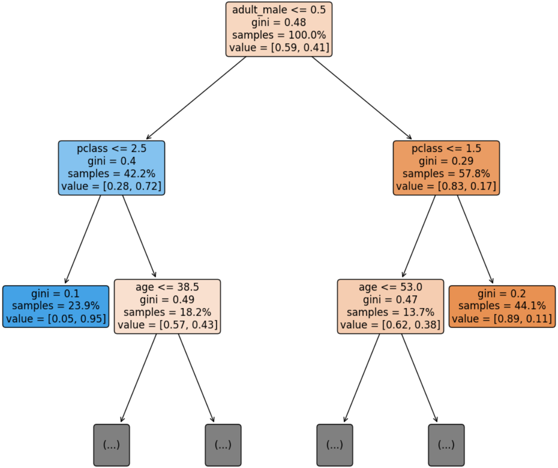


$$\Delta_{\mathcal{I}}(t) := I(t) - \frac{N_n(t^{\text{left}})}{N_n(t)} I(t^{\text{left}})$$

$$- \frac{N_n(t^{\text{right}})}{N_n(t)} I(t^{\text{left}})$$

$$\text{MDI}(k, T) = \sum_{t \in I(T), v(t)=k} \frac{N_n(t)}{n} \Delta_{\mathcal{I}}(t)$$

Variable Importances: MDI



$$\Delta_{\mathcal{I}}(t) := I(t) - \frac{N_n(t^{\text{left}})}{N_n(t)} I(t^{\text{left}}) - \frac{N_n(t^{\text{right}})}{N_n(t)} I(t^{\text{left}})$$

$$\text{MDI}(k, T) = \sum_{t \in I(T), v(t)=k} \frac{N_n(t)}{n} \Delta_{\mathcal{I}}(t)$$

$\text{MDI}(\text{Sex}, T) =$

$0.48 - 0.42 \cdot 0.4 - 0.58 \cdot 0.29 = 0.14$

$\text{MDI}(\text{Pclass}, T) =$

$(0.42 \cdot 0.4 - 0.24 \cdot 0.1 - 0.18 \cdot 0.49) +$

$(0.58 \cdot 0.29 - 0.14 \cdot 0.47 - 0.44 \cdot 0.2)$

$= 0.07$

The textbook story

California Housing data

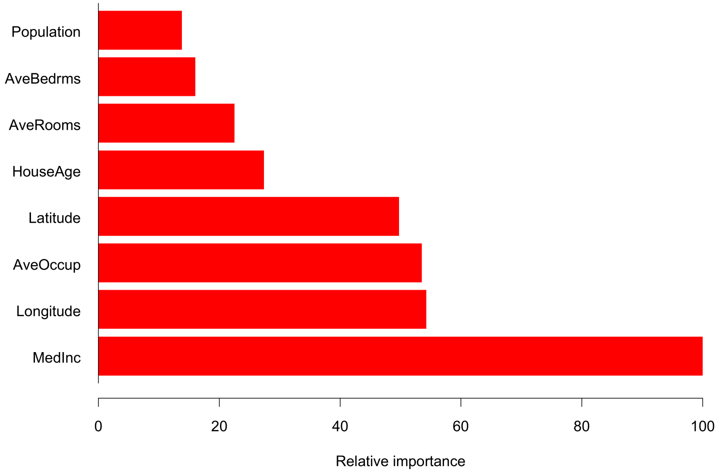
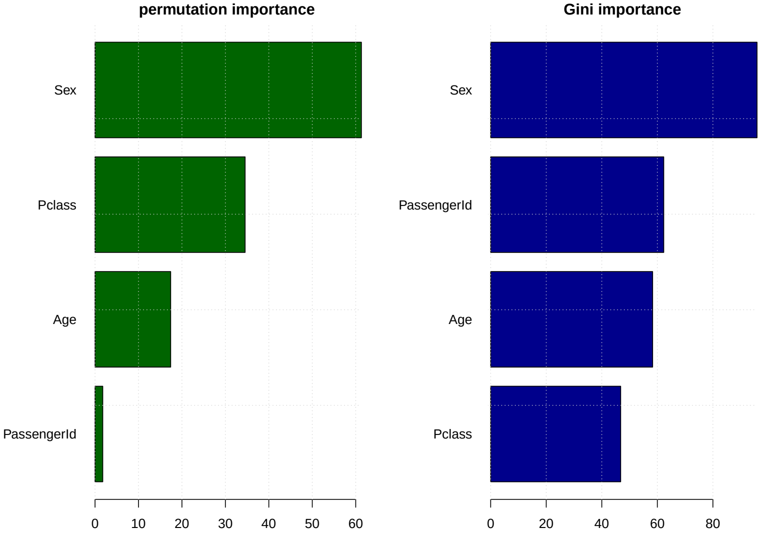


FIGURE 10.14. *Relative importance of the predictors for the California housing data.*

Explanatory Overfitting

Gini importance can be highly misleading

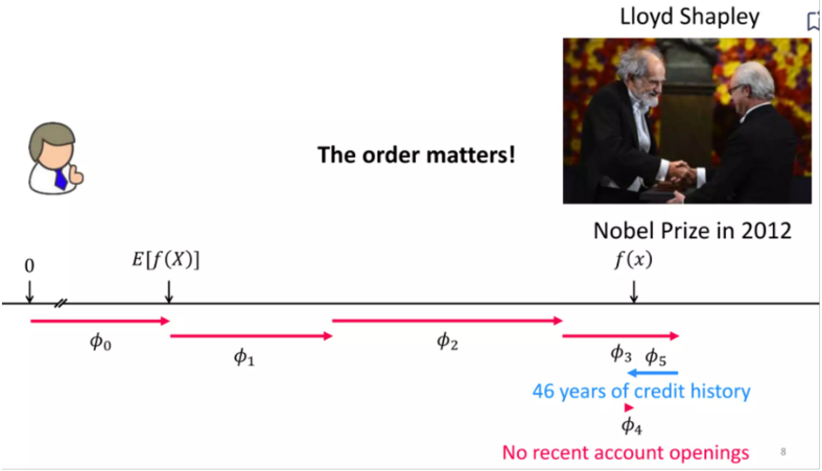


Global vs. Local Explanations

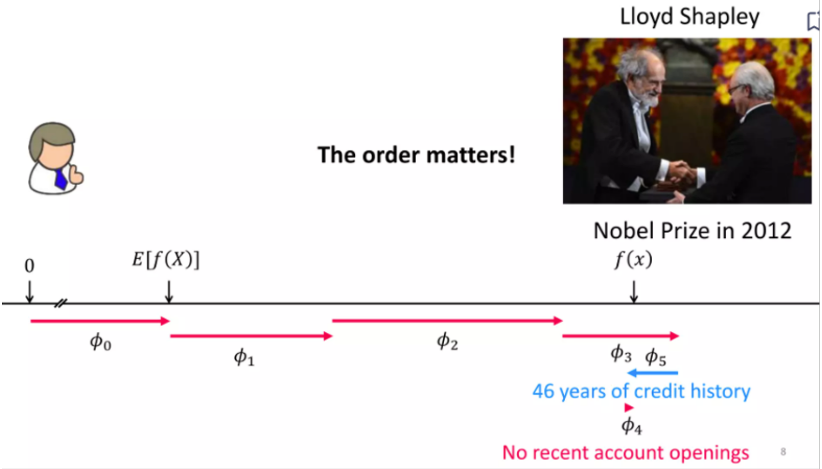
Credit Allocation

$$\phi_i(f, x) = \sum_{R \in \mathcal{R}} \frac{1}{M!} \left[f_x(P_i^R \cup i) - f_x(P_i^R) \right]$$

Credit Allocation



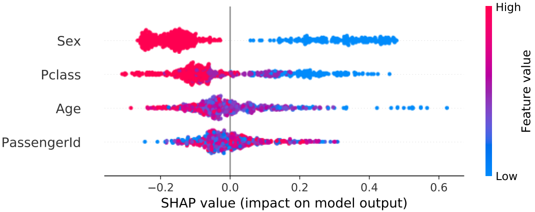
Credit Allocation

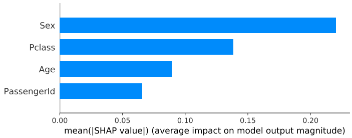


Averaging over all N! possible orderings !

SHAP values

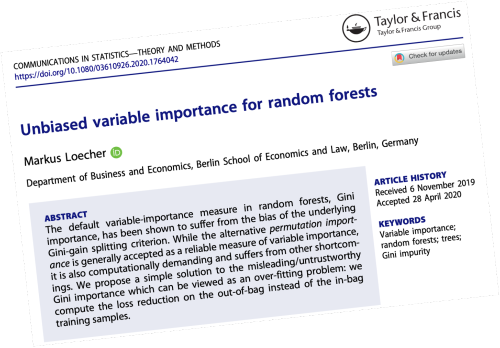
- Appealing properties: Additivity and Consistency.
- “TreeExplainer” computes local explanations based on exact **Shapley values** in polynomial time





Debiasing MDI

The OOB idea


COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2020.1764042

Taylor & Francis
Taylor & Francis Group

Check for updates

Unbiased variable importance for random forests

Markus Loecher
Department of Business and Economics, Berlin School of Economics and Law, Berlin, Germany

ABSTRACT
The default variable-importance measure in random forests, Gini importance, has been shown to suffer from the bias of the underlying Gini-gain splitting criterion. While the alternative *permutation importance* is generally accepted as a reliable measure of variable importance, it is also computationally demanding and suffers from other shortcomings. We propose a simple solution to the misleading/untrustworthy Gini importance which can be viewed as an over-fitting problem: we compute the loss reduction on the out-of-bag instead of the in-bag training samples.

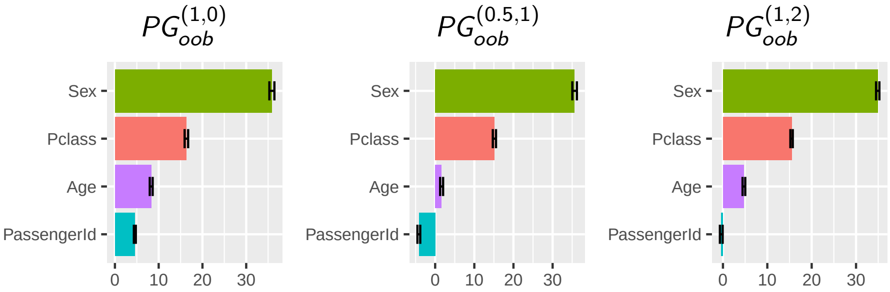
ARTICLE HISTORY
Received 6 November 2019
Accepted 28 April 2020

KEYWORDS
Variable importance; random forests; trees; Gini impurity

$$PG_{oob}^{\alpha,\lambda} = \alpha \cdot G_{oob} + (1-\alpha) \cdot G_{in} + \lambda \cdot (\hat{p}_{oob} - \hat{p}_{in})^2$$

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

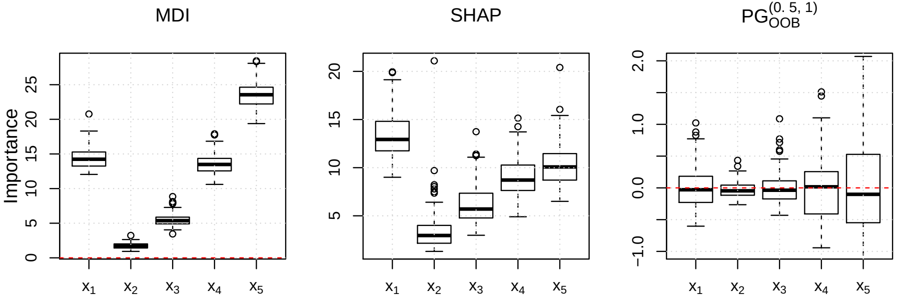
OOB Titanic



$PG_{oob}^{(1,0)}$
$PG_{oob}^{(0.5,1)}$
$PG_{oob}^{(1,2)}$
Sex
Pclass
Age
PassengerId
0
10
20
30

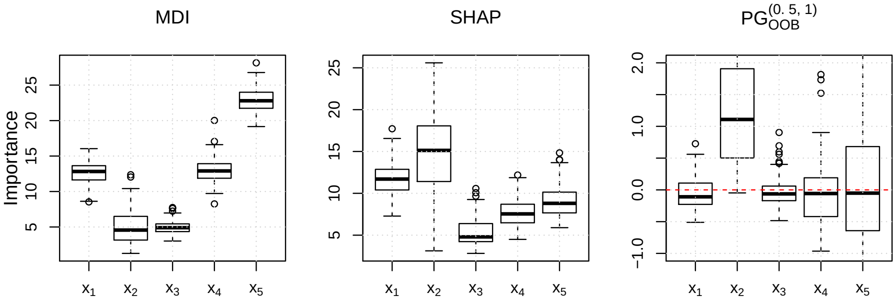
Simulated Data, Null

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, 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:

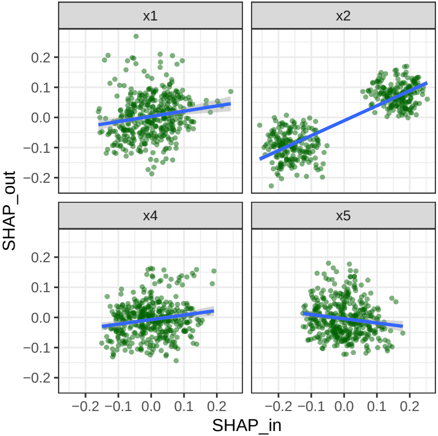
$$P(Y=1|X) = \text{Logistic}\left(\frac{2}{5}\sum_{j \in S} x_j/j - 1\right)$$

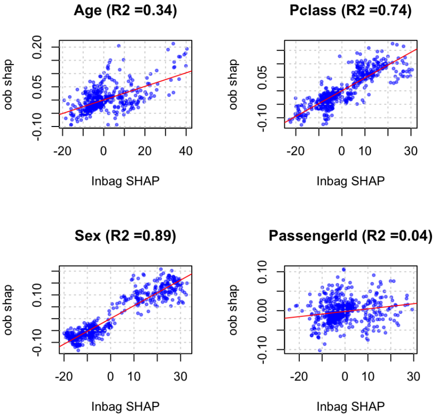
$\widehat{PG}_{oob}^{(1,0)}$	$PG_{oob}^{(1,0)}$	$\widehat{PG}_{oob}^{(0.5,1)}$	$PG_{oob}^{(0.5,1)}$	SHAP	SHAP_{in}	SHAP_{oob}	MDA	MDI
0.66	0.28	0.92	0.78	0.66	0.56	0.73	0.65	0.10

Table 1: Average AUC scores for noisy feature identification. MDA = permutation importance, MDI = (default) Gini impurity. The $\widehat{PG}_{oob}$ scores apply the variance bias correction $n/(n-1)$. The SHAP_{in}, SHAP_{oob} scores are based upon separating the inbag from the oob data.

Debiasing SHAP

Sample Splitting





Shrunk SHAP

1. Compute SHAP scores separately for inbag/oob.
2. Fit a linear model $SHAP_{j,oob} = \beta_j \cdot SHAP_{j,inbag} + u$
3. Use the estimates $\widehat{SHAP}_{j,oob} = \hat{\beta}_j \cdot SHAP_{j,inbag}$ as local explanations instead of the original $SHAP_j$ which mix inbag and oob values.

Shrunk SHAP

1. Compute SHAP scores separately for inbag/oob.
2. Fit a linear model $SHAP_{j,oob} = \beta_j \cdot SHAP_{j,inbag} + u$
3. Use the estimates $\widehat{SHAP}_{j,oob} = \hat{\beta}_j \cdot SHAP_{j,inbag}$ as local explanations instead of the original $SHAP_j$ which mix inbag and oob values.

- Rescale scores by multiplying with

$$\sum_{j=1}^{M} \phi_j(f,x) / \sum_{j=1}^{M} \hat{\beta}_j \cdot \phi_j(f,x)$$

preserves *local accuracy*: $f(x) = E(f) + \sum_{i=1}^{M} \phi_i(f,x)$

Shrunk SHAP

1. Compute SHAP scores separately for inbag/oob.
2. Fit a linear model $SHAP_{j,oob} = \beta_j \cdot SHAP_{j,inbag} + u$
3. Use the estimates $\widehat{SHAP}_{j,oob} = \hat{\beta}_j \cdot SHAP_{j,inbag}$ as local explanations instead of the original $SHAP_j$ which mix inbag and oob values.

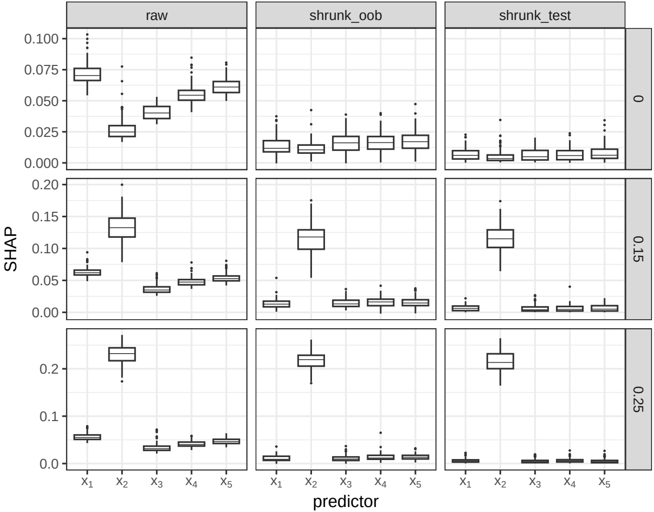
- Rescale scores by multiplying with

$$\sum_{j=1}^{M} \phi_j(f,x) / \sum_{j=1}^{M} \hat{\beta}_j \cdot \phi_j(f,x)$$

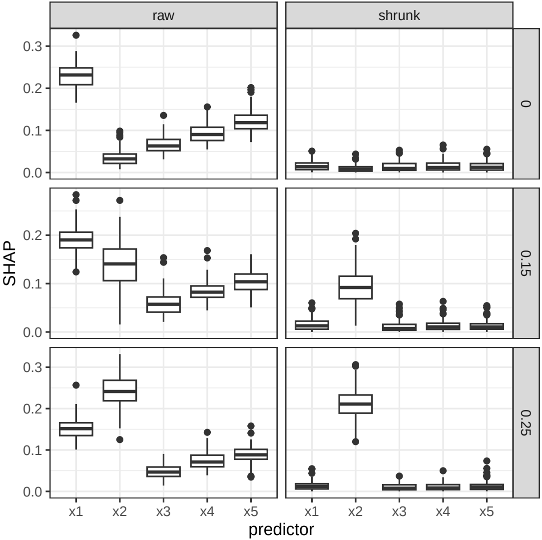
preserves *local accuracy*: $f(x) = E(f) + \sum_{i=1}^{M} \phi_i(f,x)$

- Alternatively, fit two forests

Shrunk SHAP, RF

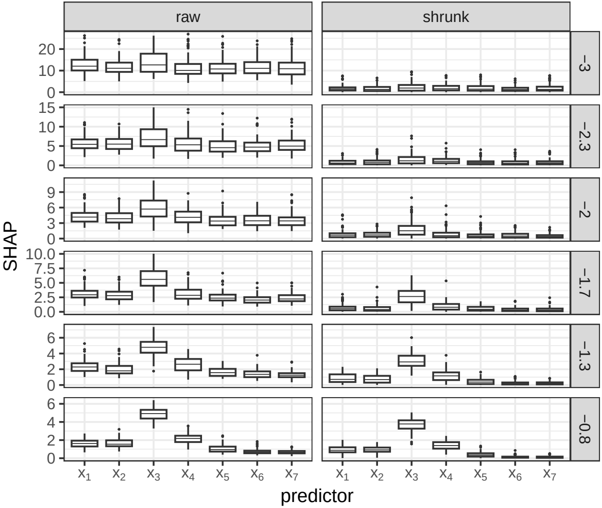


Shrunk SHAP, XGBoost

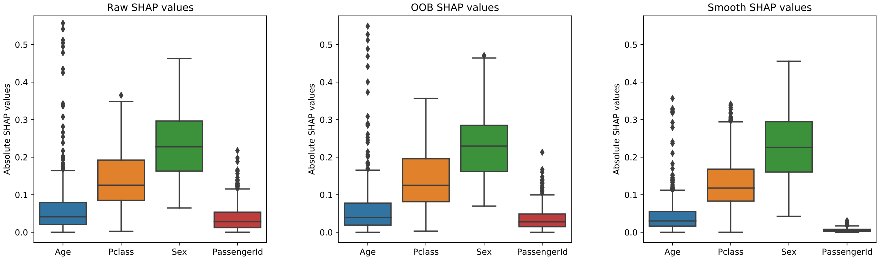


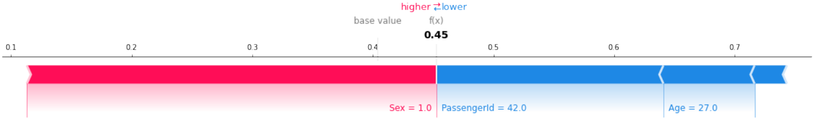
Shrunk SHAP, Complex Data

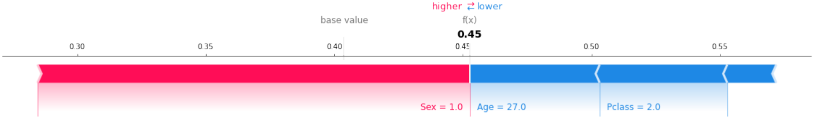
$$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.05)^2 + 10X_4 + 5X_5 + \epsilon$$



Shrunk SHAP, Titanic







Feature ranking as a classification task

SHAP	$\widehat{\text{SHAP}}_{in}^{shrunk}$	SHAP_{oob}	MDA	MDI
0.66	0.89	0.73	0.65	0.10

Table 2: Average AUC scores for relevant feature identification. *MDA* = permutation importance, *MDI* = (default) Gini impurity, SHAP_{oob} scores are based upon only the oob data. The $\widehat{\text{SHAP}}_{in}^{shrunk}$ scores outperform all other methods.

Summary

- **We are interpreting/explaining models not data directly !**
- Hence, *SHapley Additive exPlanation* (SHAP) values are susceptible to "explanatory overfitting" to the training data as is MDI.
- Combining inbag and OOB data debiases MDI and SHAP.
- Feature-wise detection of over-fitting

Paper

Home > AStA Advances in Statistical Analysis > Article

Debiasing SHAP scores in random forests

Original Paper | Open access | Published: 22 August 2023 | (2023)

Download PDF

You have full access to this open access article

AStA Advances in Statistical Analysis

Aims and scope →

Submit manuscript →