On functional decompositions, post-hoc machine learning explanations and fairness

Munir Hiabu

Interpretability

Why Interpretability?

- Why apply an algorithm?
  - Prediction
    - We want to predict the response of new observations
  - Inference
    - We want to understand the data generating mechanism

Why Interpretability? Some Points on prediction

Even if prediction is the goal, interpretability can be useful.

- Algorithmic accountability. Are estimates transparent?
  - Ethical responsibility and/or ethical image
  - EU regulation
    - GDPR: " *Individuals have the right to an explanation of the logic behind the decision.* " Kaminski (2021)
    - EU AI Act
- If algorithm is transparent, further benefits:
  - Biases/Irregularities are easier to detect
  - Aiming for causal/plausible effects provides more robustness (compared to fitting based on correlations), since unmeasured confounders can hunt you later under distributional shifts.

Two kinds of interpretability

1. We wish to understand the algorithm (Prediction)
2. We wish to understand the data-generating mechanism (Prediction & Inference)

Overview Machine learning vs GLM

Interpretability in the first part of this talk:

- Understanding the relationship between X and $\widehat{m}(X)$

Quality	Linear Models	Machine Learning
interpretable	✔	✘
interactions manually	✔	✔
interactions	✘	✔
variable selection/sparsity	✔	✔
non-linearity	✘	✔

- Current machine learning algorithms are often highly flexible and can deal with interaction, non-linearity, sparsity and variable selection; but they are not interpretable.

Current post-hoc machine learning interpretations

- Global explanation: Partial Dependence Plots
- local explanations: interventional SHAP

Partial Dependence Plot Toy example I (by Elke Gagelmans, Michel Denuit and Julien Trufin)

- Partial Dependence Plot is a global explanation because it explains features globally.
- Partial Dependence Plot for feature k: $\xi_k(x_k) = \int \hat{m}(x) p_{-k}(x_{-k}) \mathrm{d}x_{-k}$.

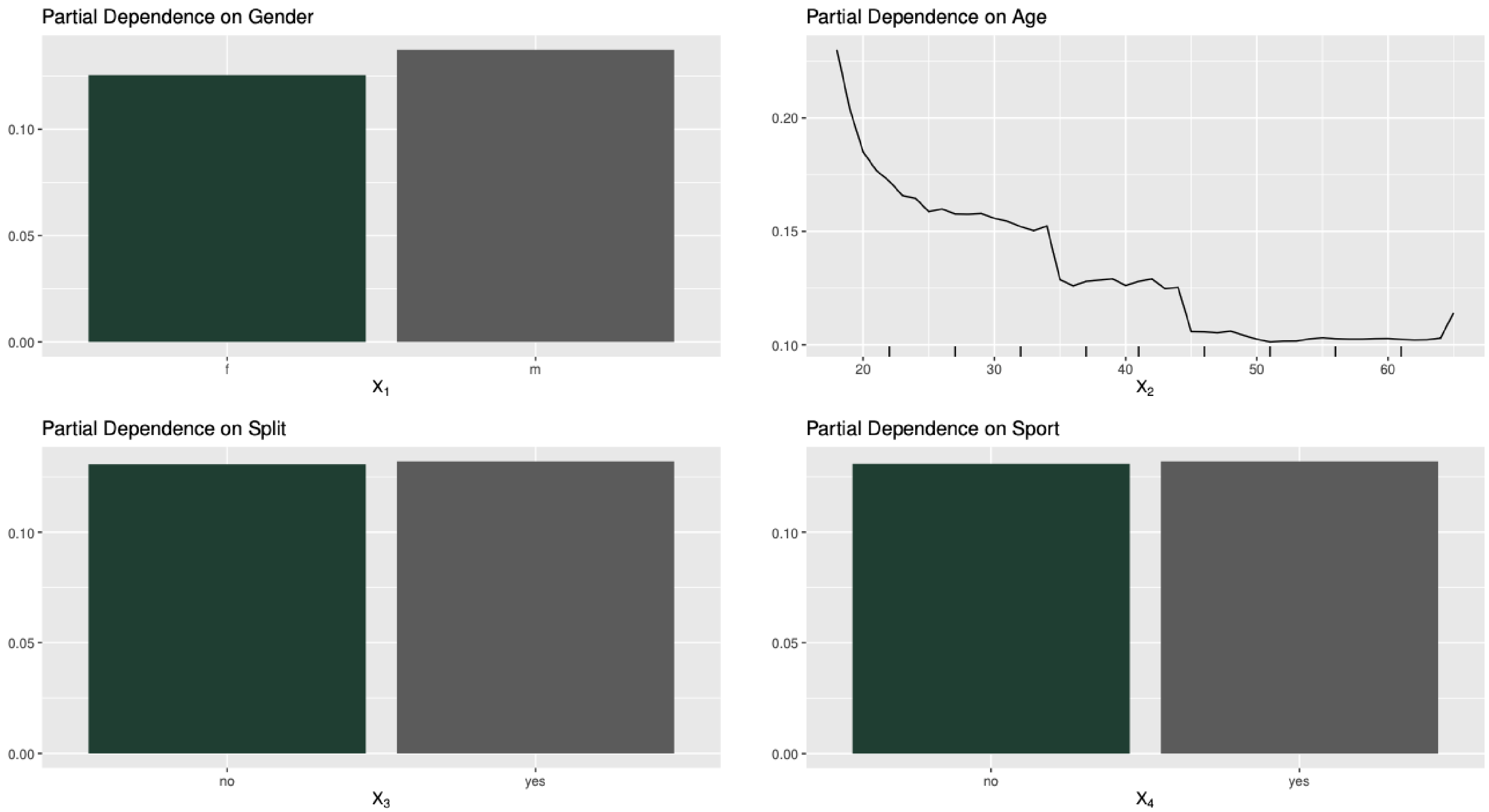
https://detralytics.com/wp-content/uploads/2020/03/Faqctuary_2020-02_Features-with-flat-partial-dependence-plots.pdf

- $X_1 =$ Gender: policyholder's gender (female or male);
- $X_2 =$ Age: policyholder's age (integer values from 18 to 65);
- $X_3 =$ Split: whether the policyholder splits its annual premium or not (yes or no);
- $X_4 =$ Sport: whether the policyholder's car is a sports car or not (yes or no).

$$
\begin{aligned}
\lambda(\boldsymbol{x}) = & 0.1 \times \left(1 + 0.1 I_{\{x_1 = \text{ male }\}}\right) \\
& \times \left(1 + \frac{1}{\sqrt{x_2 - 17}}\right) \\
& \times \left(1 + 0.3 I_{\{18 \le x_2 < 35\}} I_{\{x_4 = \text{ yes }\}} - 0.3 I_{\{45 \le x_2 < 65\}} I_{\{x_4 = \text{ yes }\}}\right)
\end{aligned}
$$

Partial Dependence Plot Toy example I

Random Forest explanations



Partial Dependence Plot Toy example I

- Partial dependence can be misleading because they ignore interaction effects.
- In abstract terms: partial dependence plots are projections into an additive space
  - (higher order) interactions are being ignored
  - Approximation is highly non-trivial and can be very unstable trough extrapolation, see Apley and Zhu (2020)

Current post-hoc machine learning interpretations

- Global explanation: Partial Dependence Plots
- local explanations: interventional SHAP

Shapley values Toy example II

Toy example II

$$m(x_1, x_2) = x_1 + x_2 + 2x_1x_2.$$

The interventional SHAP value for the first feature is

$$\phi_1(x_1, x_2) = x_1 - E[X_1] + x_1x_2 - E[X_1X_2] + x_1E[X_2] - x_2E[X_1].$$

If the features are standardized, i.e., X_1 and X_2 have mean zero and variance one, then

$$\phi_1(x_1, x_2) = x_1 + x_1x_2 - \text{corr}(X_1, X_2).$$

- Assume $\text{corr}(X_1, X_2) = 0.3$
  - individual with $x_1 = 1$ and $x_2 = -0.7$ would see a SHAP value of 0 for the first feature"

$$\phi_1(1, -0.7) = 0.$$

Shapley values Toy example II

- Interventional SHAP values merge interaction effects and main effect and can thereby cancel each other out

glex: Unifying local and global explanations

glex: Unifying local and global explanations

This is joint work with

Joseph Meyer
Heidelberg University

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Marvin Wright
The Leibniz Institute for Prevention Research and Epidemiology - BIP

University of Bremen

Lukas Burk
The Leibniz Institute for Prevention Research and Epidemiology - BIP

University of Bremen

Paper is available on https://arxiv.org/abs/2208.06151

Code is available on https://github.com/PlantedML/glex

glex: Unifying local and global explanations

- Definitions and results
- Toy example I and Toy example II revisited

A functional decomposition

Assume a data set with d features. Also assume that we can approximate the regression function m by a (q-th) order functional decomposition:

$$m(x) \approx m_0 + \sum_{k=1}^{d} m_k(x_k) + \sum_{k_1<k_2} m_{k_1 k_2}(x_{k_1}, x_{k_2}) + \cdots + \sum_{k_1<\cdots<k_q} m_{k_1,...,k_q}(x_{k_1}, \ldots, x_{k_q}).$$

Estimator as collection of components

Instead of seeing an estimator as a high-dimensional function in x, the right-hand-side encourages the view of an estimator as a collection of main-effects, two-way interaction, three-way interactions, …

$$\{\hat{m}_S : S \subseteq 1, \ldots, d\},$$

Interpretability Marginal Identification

Marginal Identification

Let $\hat{m} = \sum_S \hat{m}^*_S$. We say the collection $\{\hat{m}^*_S : S \subseteq 1, \ldots, d\}$ fulfills the marginal identification if for every $S \subseteq \{1, \ldots, d\}$,

$$\sum_{T: T \cap S \neq \emptyset} \int m^*_T(x_T) p_S(x_S) \mathrm{d}x_S = 0 \quad \left(\Leftrightarrow \int \hat{m}^*(x) \mathrm{d}x_S = \sum_{\{T: S \subseteq 1, \ldots, d \setminus S\}} \hat{m}_T \right).$$

Theorem (Rota (1964) Harsanyi (1963) Hiabu, Meyer, and Wright (2022) (Möbius inverse, Harsanyi dividend))

- The marginal identification has a unique solution.
- We have an explicit closed form expression for the components.

Interpretability Marginal Identification: Summary

Summary(Marginal Identification)

If m is identified via the marginal identification,

$$\hat{m}(x) = \hat{m}_0^* + \sum_j \hat{m}_j^*(x_j) + \sum_{j<k} \hat{m}_{j,k}^*(x_j, x_k) + \cdots$$

then

- Interventional SHAP values are:

$$\phi_k(x) = \hat{m}_k^*(x_k) + \frac{1}{2} \sum_j \hat{m}_{kj}^*(x_{kj}) + \cdots$$

- Partial dependence plots are:

$$\xi_k(x_k) = \hat{m}_0^* + \hat{m}_k^*(x_k).$$

- Plugin-Debiasing: If components are well estimated and if S are protected variables and all components that contain a subset of S are dropped, we derive a plugin de-biased estimator.

glex: Unifying local and global explanations

- Definitions and results
- Toy example I and Toy example II revisited

Interpretability Partial Dependence Plot: Toy example I

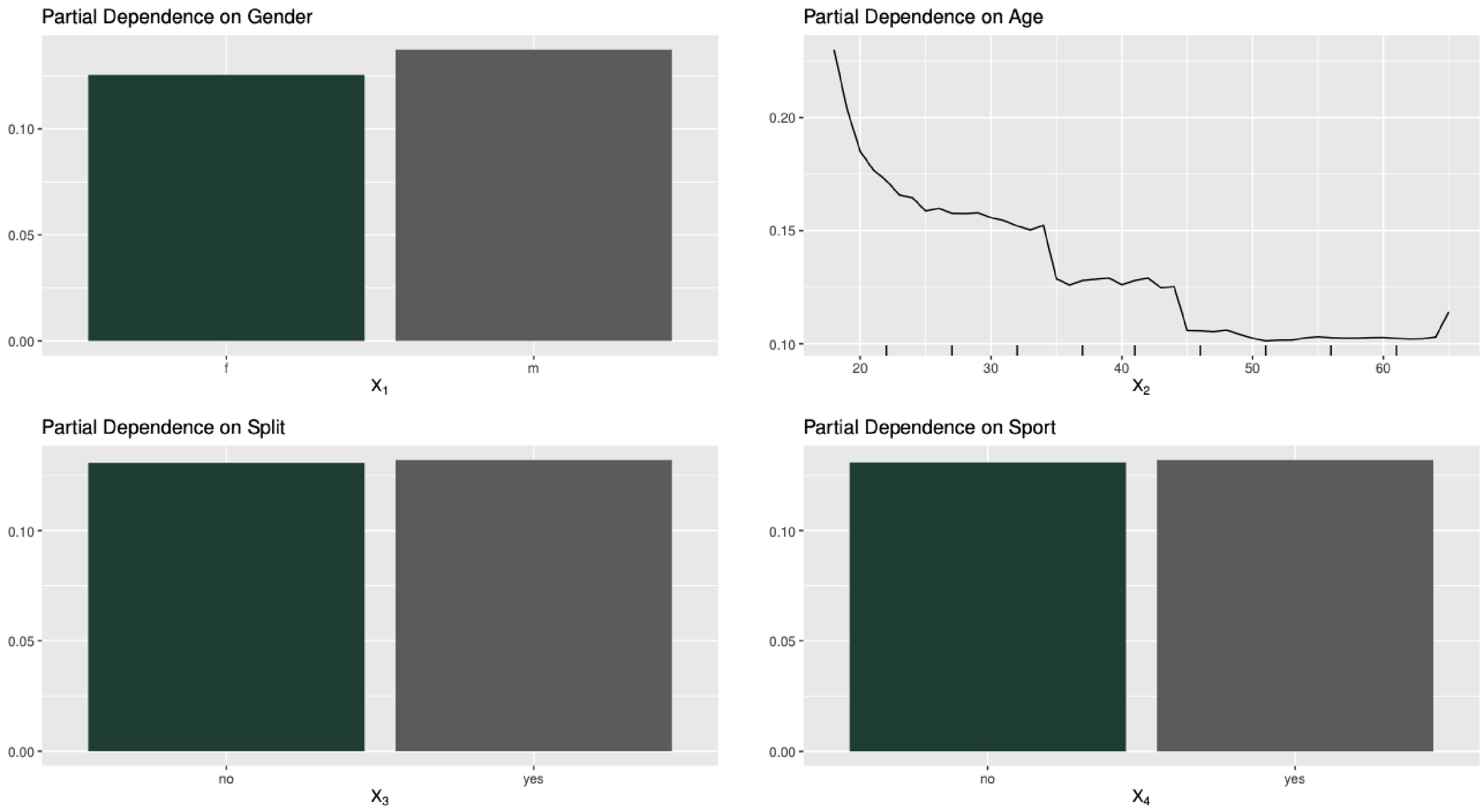
https://detralytics.com/wp-content/uploads/2020/03/Faqctuary_2020-02_Features-with-flat-partial-dependence-plots.pdf

- $X_1 =$ Gender: policyholder's gender (female or male);
- $X_2 =$ Age: policyholder's age (integer values from 18 to 65);
- $X_3 =$ Split: whether the policyholder splits its annual premium or not (yes or no);
- $X_4 =$ Sport: whether the policyholder's car is a sports car or not (yes or no).

$$
\begin{aligned}
\lambda(\boldsymbol{x}) =& 0.1 \times \left(1 + 0.1 I_{\{x_1 = \text{ male }\}}\right) \\
& \times \left(1 + \frac{1}{\sqrt{x_2 - 17}}\right) \\
& \times \left(1 + 0.3 I_{\{18 \leq x_2 < 35\}} I_{\{x_4 = \text{ yes }\}} - 0.3 I_{\{45 \leq x_2 < 65\}} I_{\{x_4 = \text{ yes }\}}\right)
\end{aligned}
$$

Interpretability Partial Dependence Plot: Toy example I

Random Forest explanaitions



Interpretability Marginal Identification: Toy example I

We now fit and interpret the data with our new toolbox:

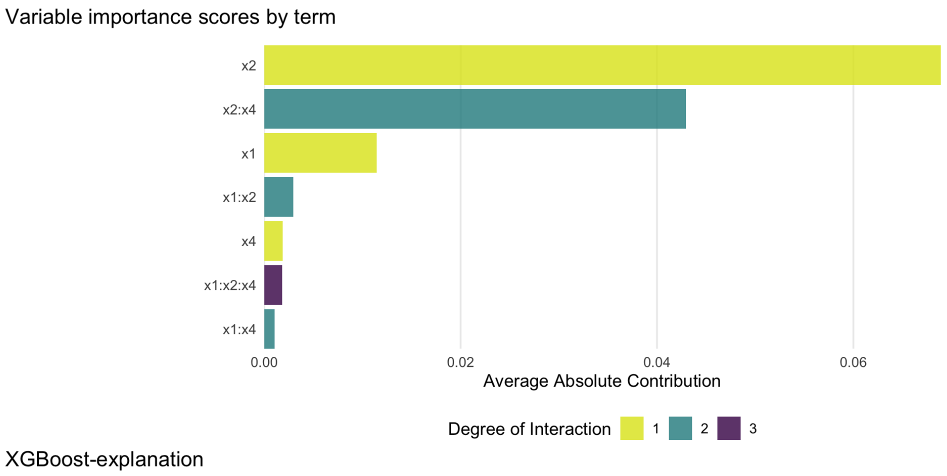
```r
library(xgboost)
library(glex)
xg <- xgboost(data = cbind(x1,x2,x3,x4), label = y, params = list(max_depth = 4, eta = .01, objective = "count:poisson"), nrounds = 100,verb
res <- glex(xg, x)
res$m
```

```
                 x1          x2           x4         x1:x2         x1:x4
     1: -0.01152380  0.25194600 -0.0028184073  0.0115238001  0.0011399114
     2: -0.01152380 -0.07413661  0.0009307946 -0.0032819044 -0.0010556617
     3: -0.01152380 -0.07662543  0.0009307946 -0.0007930862 -0.0010556617
     4: -0.01152380 -0.01418692  0.0009307946 -0.0022103293 -0.0010556617
     5:  0.01136923 -0.07257039  0.0009307946  0.0016583172  0.0009999517
    ---
499996:  0.01136923 -0.01418692  0.0009307946  0.0021572104  0.0009999517
499997: -0.01152380  0.06099980 -0.0028184073  0.0021134221  0.0011399114
499998: -0.01152380  0.07283587 -0.0028184073  0.0021134221  0.0011399114
499999: -0.01152380 -0.07909062  0.0009307946 -0.0032988089 -0.0010556617
500000:  0.01136923 -0.07909062 -0.0028184073  0.0032432671 -0.0010815641
               x2:x4      x1:x2:x4
     1:  0.065182180 -0.0011399114
     2:  0.054061705 -0.0025864581
     3:  0.051655529 -0.0001802821
     4: -0.002821577 -0.0005253192
     5:  0.055561892  0.0010237235
    ---
499996: -0.002821577  0.0005248302
499997:  0.049072182 -0.0027803520
499998:  0.060346019 -0.0027803520
499999:  0.049267178 -0.0026073710
500000: -0.049237372 -0.0025832539
```

Interpretability Partial Dependence Plot: Toy example I

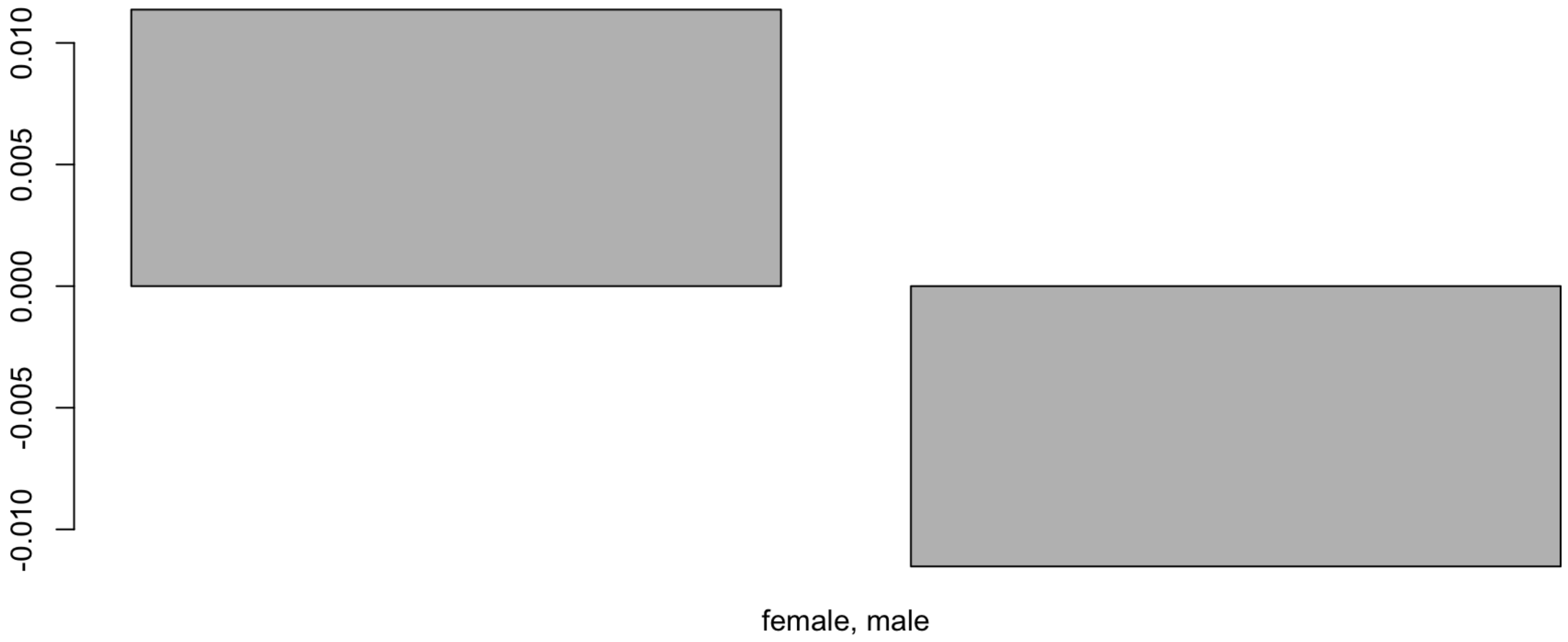
```r
#| code-fold: true
library(glex)
# Model fitting
library(xgboost)
# Visualization
library(ggplot2)
library(patchwork)
theme_set(theme_minimal(base_size = 13))
vi_xgb <- glex_vi(res)

p_vi <- autoplot(vi_xgb, threshold = 0) +
  labs(title = NULL, tag = "XGBoost-explanation")

p_vi+
  plot_annotation(title = "Variable importance scores by term") &
  theme(plot.tag.position = "bottomleft")
```

```
 ## X1 = female/male  ## X2 = age ## X3 = split no/yes  ## X4 = sports car no/yes
```



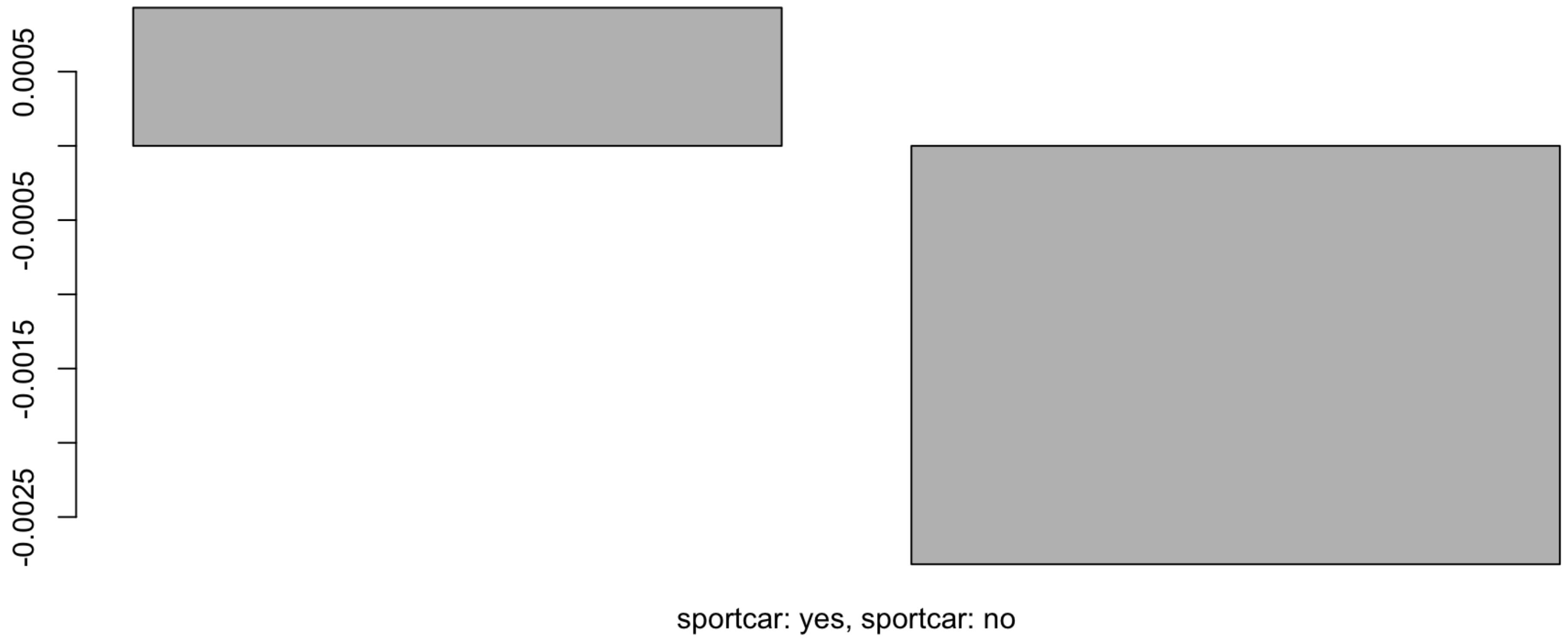
Interpretability Partial Dependence Plot: Toy example I

```
#| code-fold: true
barplot(as.numeric(c(res$m$x1[which.min(x1==0)], res$m$x1[which.min(x1==1)])), names.arg=("female, male") )
```



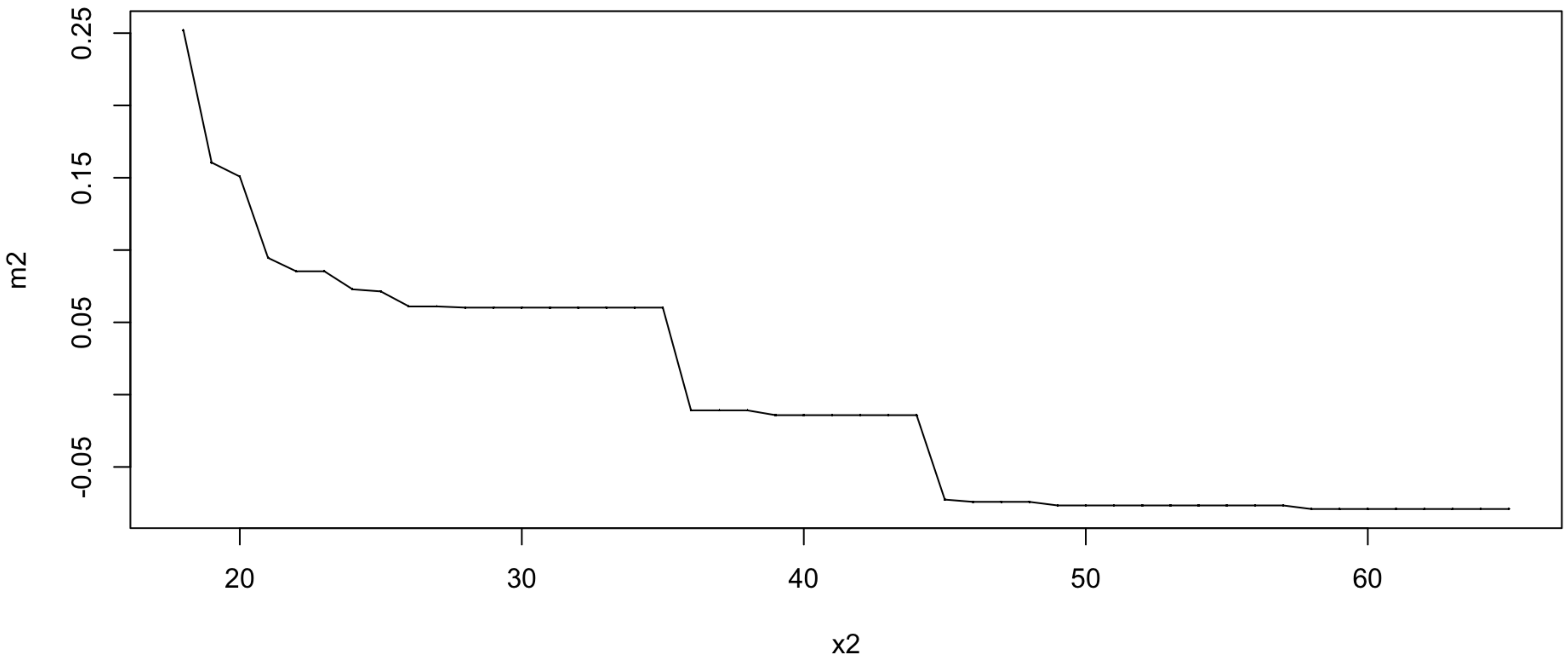
Interpretability Partial Dependence Plot: Toy example I

```
barplot(as.numeric(c(res$m$x4[which.min(x4==1)], res$m$x4[which.min(x4==0)])), names.arg=("sportcar: yes, sportcar: no") )
```



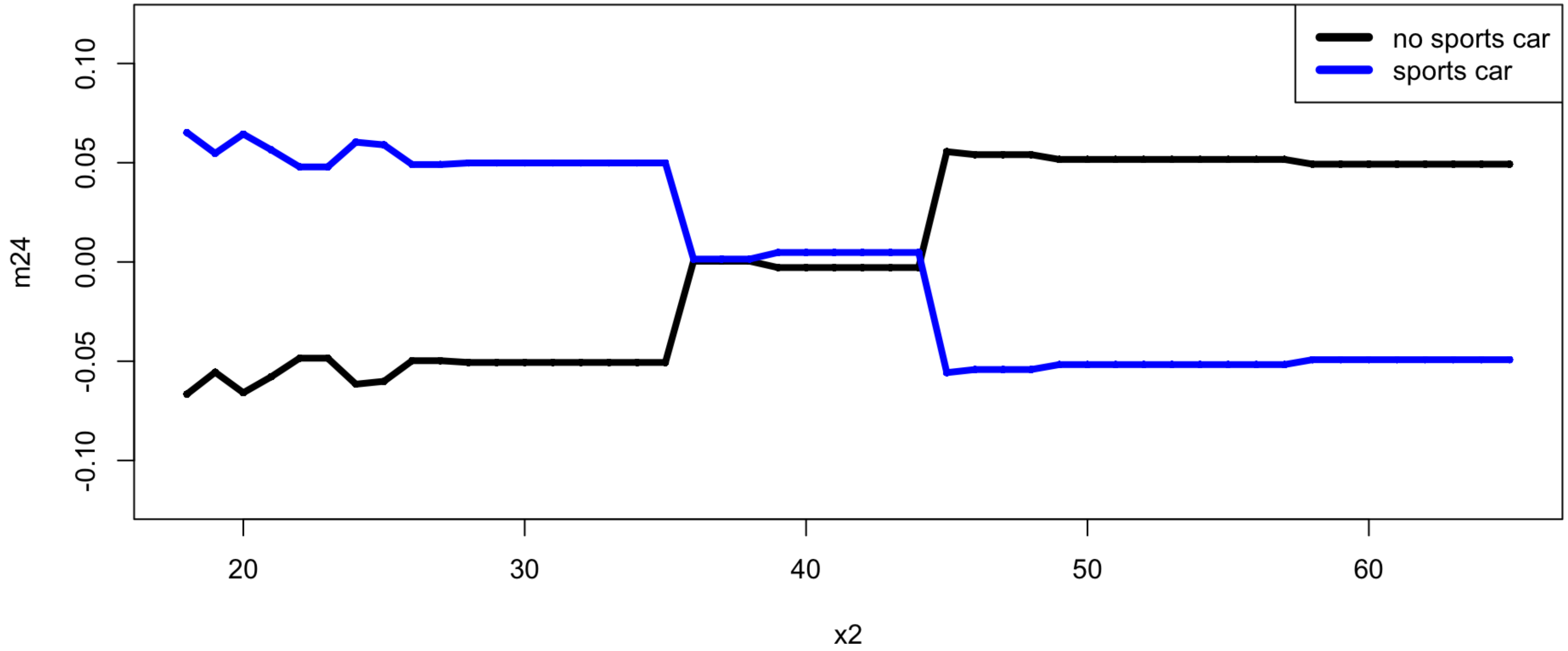
Interpretability Partial Dependence Plot: Toy example I

```r
plot(as.numeric(x2[order(x2)]), as.numeric(res$m$x2[order(x2)]),type="l", xlab="x2", ylab="m2")
```



Interpretability Partial Dependence Plot: Toy example I

```
plot(as.numeric(x2[which(x4==0)][order(x2[which(x4==0)])]),as.numeric(res$m$`x2:x4`[which(x4==0)][order(x2[which(x4==0)])]), type="l" ,ylim=
lines(as.numeric(x2[which(x4==1)][order(x2[which(x4==1)])]),as.numeric(res$m$`x2:x4`[which(x4==1)][order(x2[which(x4==1)])]), type="l", col=
legend("topright", legend=c("no sports car", "sports car"),
       col=c("black", "blue"), cex=1, lwd=5)
```



Interpretability Interventional SHAP: Toy example II

$$m(x_1, x_2) = x_1 + x_2 + x_1 x_2$$

If X_1, X_2 have each mean zero and variance one, then

- Marginal identification:

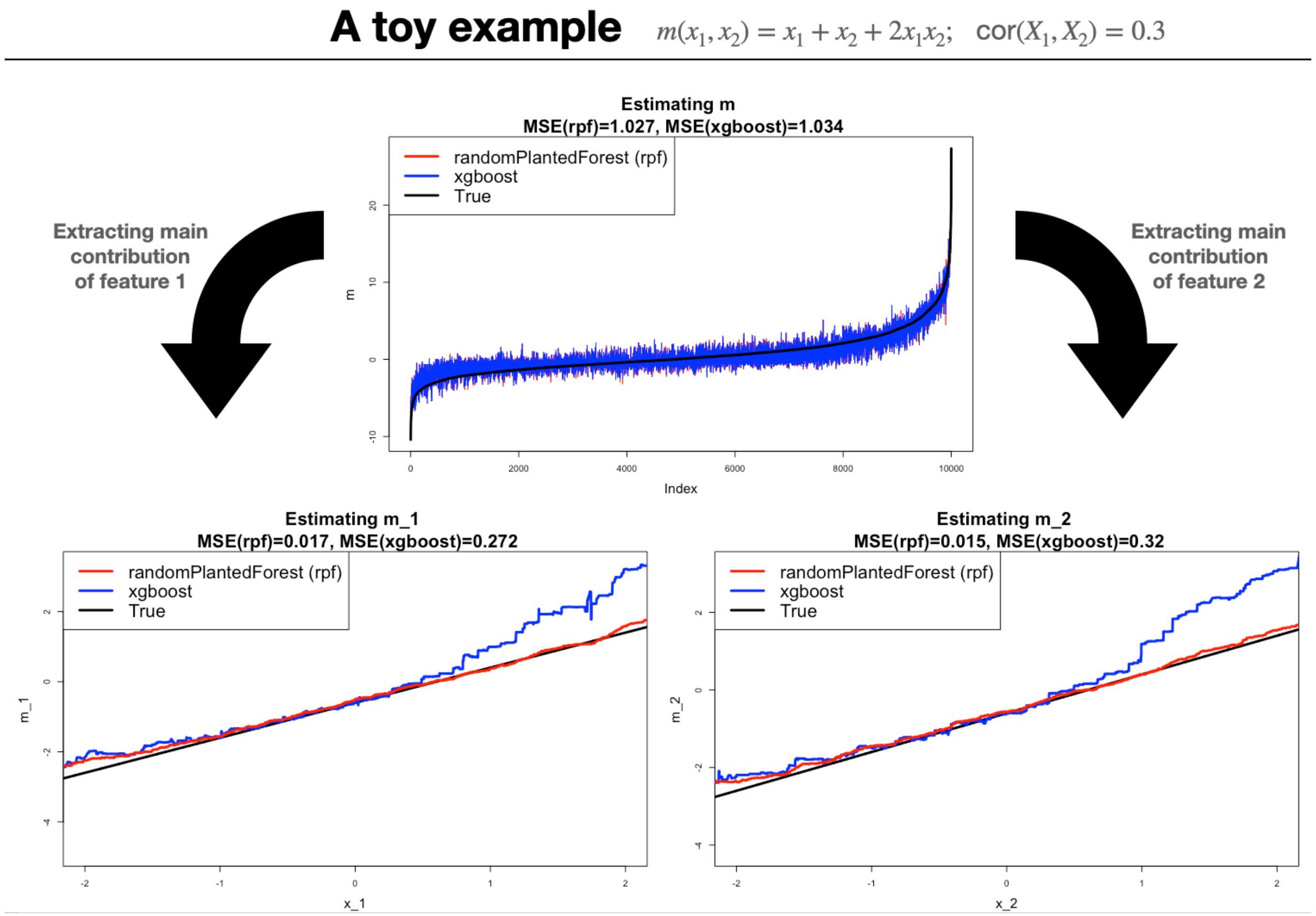
$$\begin{aligned} m_0 &= 2corr(X_1, X_2) \\ m_1^*(x_1) &= x_1 - 2corr(X_1, X_2) \\ m_2^*(x_2) &= x_2 - 2corr(X_1, X_2) \\ m_{12}^*(x_1, x_2) &= 2x_1x_2 + 2corr(X_1, X_2). \end{aligned}$$

- Partial dependence plot of x_1 is $m_1^*(x_1)$
- Interventional SHAP of x_1 is $m_1^*(x_1) + 0.5 \times m_{12}^*(x_1, x_2)$

An open problem

Interpretability An open problem

We simulate 10,000 noisy observations and got the following fits:



Random Planted Forest (rpf) joint work with Joseph Meyer, Enno Mammen (Heidelberg University)

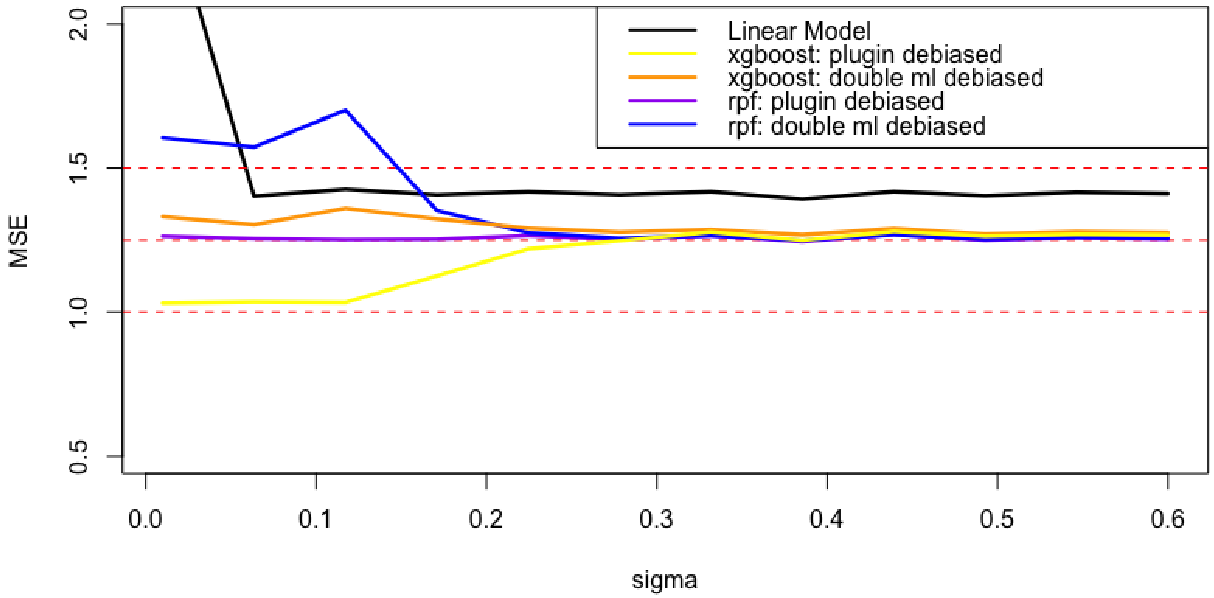
Further information: https://plantedml.com/randomPlantedForest

Interpretability An open problem

- Data generating mechansim:
  - $D = Bern(0.5)$ (protected)
  - $X_1 = D + N(0, \sigma)$
  - $X_2 = N(0, 1)$
  - $Y = \sin(X_2) + D + N(0, 1)$

Red dotted lines:

- Optimal MSE using (D, X): 1 $(= Var(N(0, 1)))$
- Optimal MSE using X: 1.25 $(= Var(D + N(0, 1)))$
- Optimal MSE of constant predictor: 1.5 $(= Var(Y))$



Thank You!

References

Apley, Daniel W., and Jingyu Zhu. 2020. “Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (4): 1059–86. https://doi.org/https://doi.org/10.1111/rssb.12377.

Harsanyi, John C. 1963. “A Simplified Bargaining Model for the n-Person Cooperative Game.” In *International Economic Review*, 4:194–220. 2.

Hiabu, Munir, Joseph T Meyer, and Marvin N Wright. 2022. “Unifying Local and Global Model Explanations by Functional Decomposition of Low Dimensional Structures.” *arXiv Preprint arXiv:2208.06151*.

Kaminski, Margot E. 2021. “The Right to Explanation, Explained.” In *Research Handbook on Information Law and Governance*, 278–99. Edward Elgar Publishing.

Rota, Gian-Carlo. 1964. “On the Foundations of Combinatorial Theory I. Theory of Möbius Functions.” *Zeitschrift für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 2 (4): 340–68.