Interprétabilité et Explicabilité

Arthur Charpentier[1]

[1] Université du Québec à Montréal

Équité des modèles, COVEA, Septembre 2022

– *Je me demande ce que cherche ce type là-bas*, demandai-je, désignant un grand individu habillé simplement qui suivait l'autre côté de la rue, en examinant anxieusement les numéros. Il tenait à la main une grande enveloppe bleue et, de toute évidence, portait un message.

– *Vous parlez de ce sergent d'infanterie de marine ? dit Sherlock Holmes*.

– *Comment diable avez-vous pu deviner cela ?* demandai-je.
– *Deviner quoi ?* fit-il sans aménité.
– *Eh bien, qu'il était un sergent de marine en retraite ?*
– *Je n'ai pas de temps à perdre en bagatelles !* répondit-il avec brusquerie avant d'ajouter dans un sourire : *excusez ma rudesse ! Vous avez rompu le fil de mes pensées. Mais c'est peut-être aussi bien. Ainsi donc vous ne voyiez pas que cet homme était un sergent de marine ?*
– *Non, certainement pas !*

–Décidément, l'explication de ma méthode me coûte plus que son application ! Si l'on vous demandait de prouver que deux et deux font quatre, vous seriez peut-être embarrassé ; et cependant, vous êtes sûr qu'il en est ainsi. Malgré la largeur de la rue, j'avais pu voir une grosse ancre bleue tatouée sur le dos de la main du gaillard. Cela sentait la mer. Il avait la démarche militaire et les favoris réglementaires ; c'était, à n'en pas douter, un marin. Il avait un certain air de commandement et d'importance. Rappelez-vous son port de tête et le balancement de sa canne ! En outre, son visage annonçait un homme d'âge moyen, sérieux, respectable. Tous ces détails m'ont amené à penser qu'il était sergent.

Conan-Doyle (1887) Une étude en rouge

Agenda

Exemple(s) en classification d'images

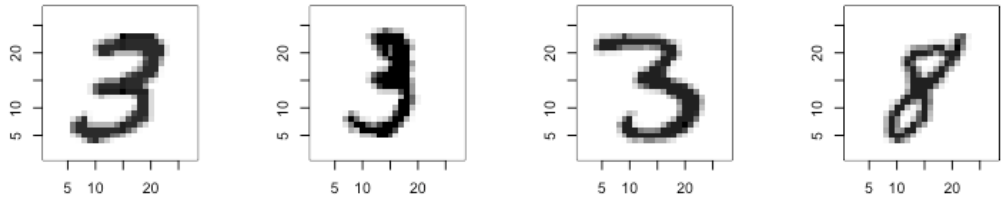
Régression et interprétation

Interprétation et identifiabilité

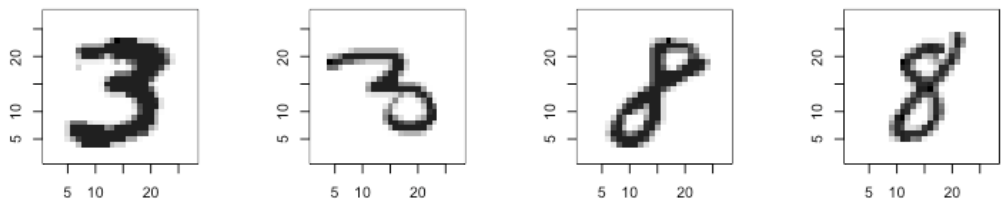
Comprendre sans prévoir, prévoir sans comprendre

Formalisations mathématiques

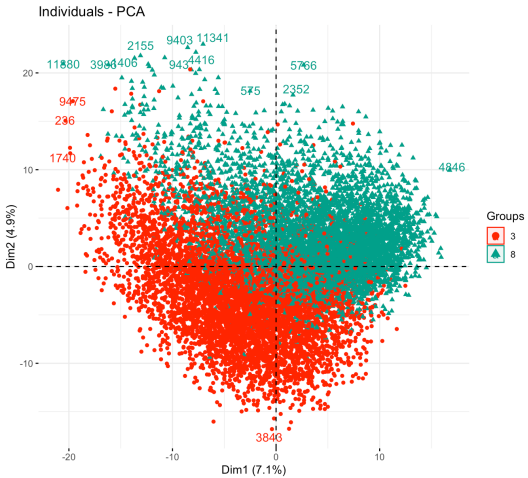
Interprétation et classification d'images

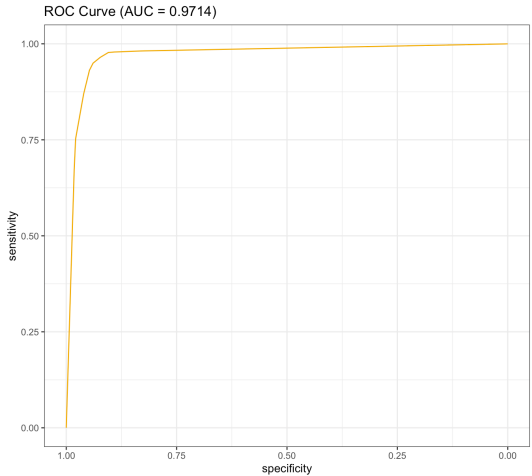


Quelles images correspondent à un 3, et à un 8 ?

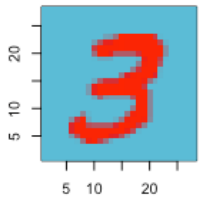


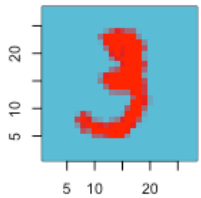
Interprétation et classification d'images

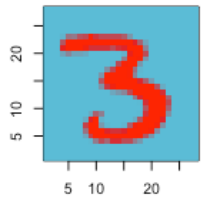


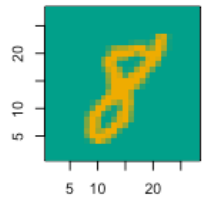


Interprétation et classification d'images

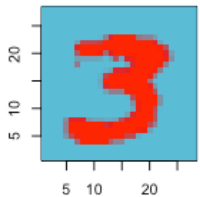


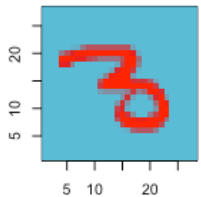


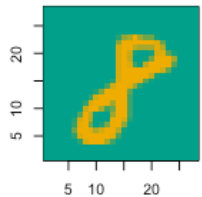


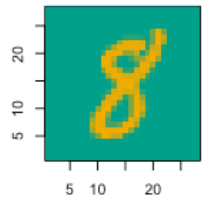


Quelles images correspondent à un 3, et à un 8 ? ■ = 3 et ■ = 8

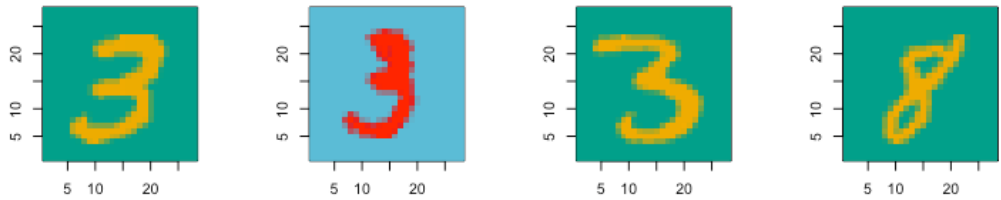




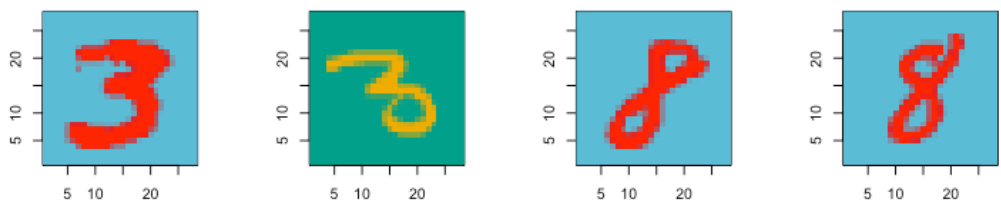




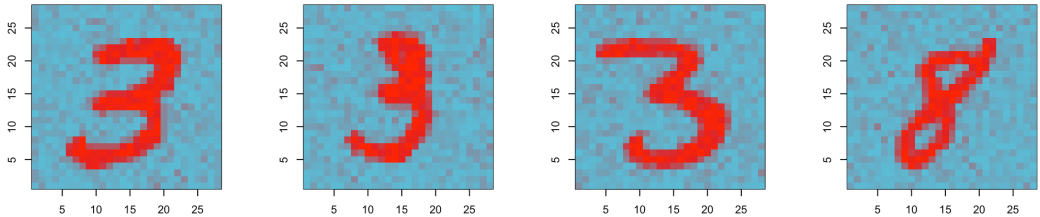
Interprétation et classification d'images



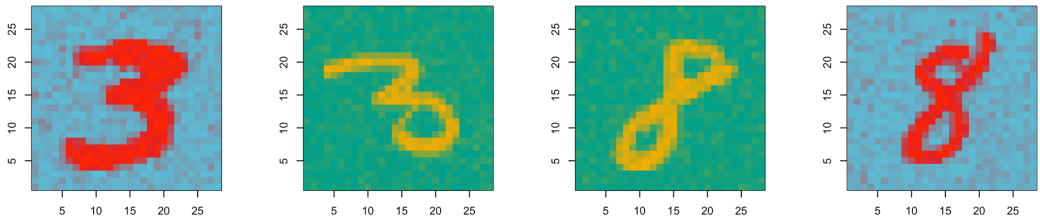
Quelles images correspondent à un 3, et à un 8 ?



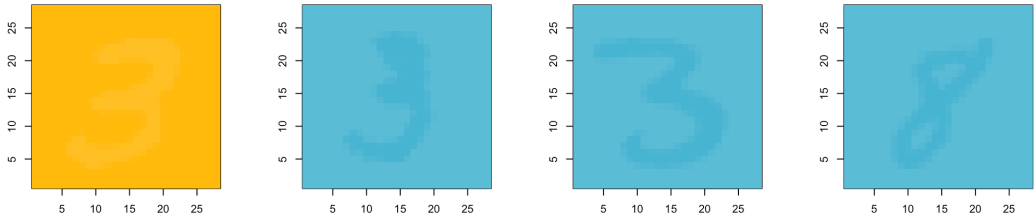
Interprétation et classification d'images



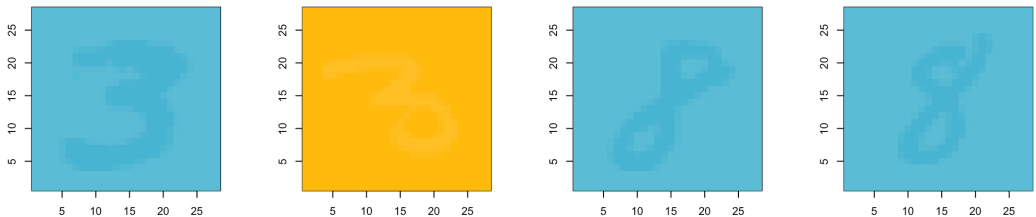
Un peu de bruit peu (beaucoup) perturber l'algorithme (cf Elsayed et al. (2018))



Interprétation et classification d'images



Parfois, l'algorithme peut être bien plus efficace que l'oeil humain



Interprétation et classification d'images

Le cas de l'apprentissage semi-supervisé est encore plus compliquée. Tenenbaum et al. (2011) a introduit les "tufas". Voici un exemple de tufa

Interprétation et classification d'images

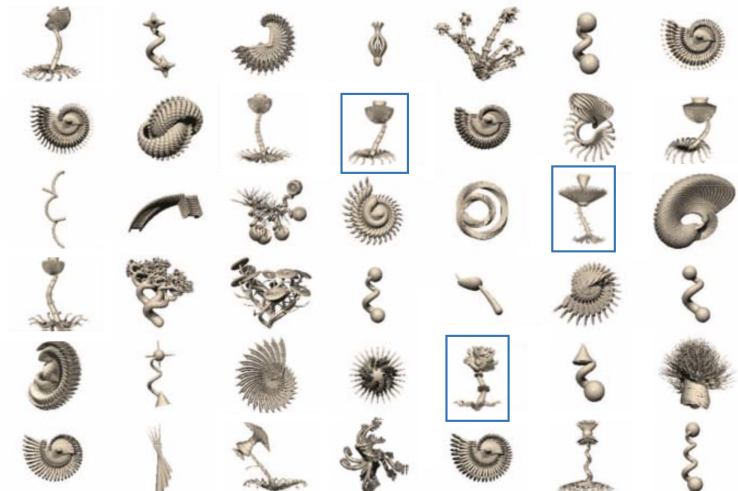
Voici un autre exemple de tufa (selon Tenenbaum et al. (2011))

Interprétation et classification d'images

Et encore un autre tufa (toujours selon Tenenbaum et al. (2011))

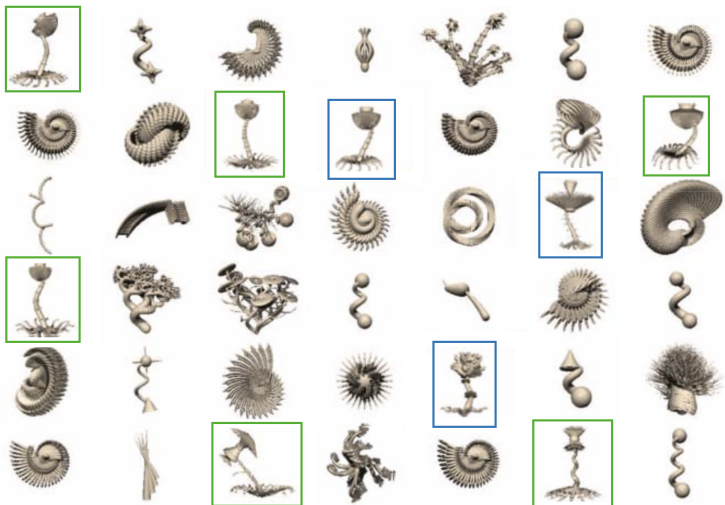
Interprétation et classification d'images

C'est bon ? où sont les tufas sur l'image suivante ?



Interprétation et classification d'images

Difficile ? pour pourtant, Tenenbaum et al. (2011) souligne qu'il y a un consensus fort entre les personnes qui ont participé à l'expérience pour les identifier...



Interprétation et classification d'images

N'est-ce pas ce qu'on essaye de faire quand on entraîne un algorithme pour identifier des dégâts des eaux sur des photos ?

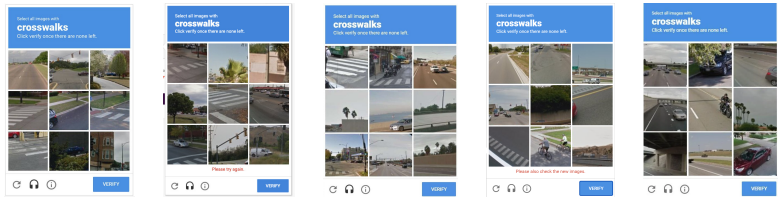



... ou pour identifier des passages pour piétons, pour traverser ?



Interprétation et classification d'images

Il est possible d'utiliser des humains pour créer davantage de données labélisées (utilisation de captcha par exemple),



Mais ici, il est aussi possible d'expliquer ce qu'est un passage piéton

Explicabilité ? I

Dans les années 80, la *Strategic Computing Initiative* du département de la défense américaine, lançait l'acronyme "EES (Explainable Expert Systems)", l'adjectif "explainable" donnant le nom "explainability", Swartout et al. (1991)

"*An explanation is an assignment of causal responsibility* ", Halpern and Pearl (2020a,b)

"*An explanation is an answer to a why-question* ", Lewis (1986), Dennett (1987) ou Lipton (1990)

L'explicabilité est "*la capacité, l'inclination ou l'aptitude à rendre clair ou compréhensible, ou à expliquer le sens d'un algorithme*", Beaudouin et al. (2020)

Explicabilité ? II

Selon les principes de Phillips et al. (2020),

- Explication : Un système fournit ou contient des preuves ou des raisons qui accompagnent les sorties et/ou les processus.
- Significatif : Un système fournit des explications qui sont compréhensibles pour le(s) consommateur(s) visé(s).
- Exactitude de l'explication : Une explication reflète correctement la raison pour laquelle le résultat est généré et/ou le processus du système.
- Limites des connaissances : Un système ne fonctionne que dans les conditions pour lesquelles il a été conçu et lorsqu'il atteint un niveau de confiance suffisant dans ses résultats.

On parle d'explications contrastives quand elles sont en réponse à des contrefactuels, Lipton (1990).

Explicabilité ? III

Souvent, on ne se demande pas *"pourquoi E s'est produit"* mais plutôt *"pourquoi E s'est produit, et pas F"* ?

Ou *"commment la prédiction aurait été si l'entrée **x** avait été différente"* ?

Si ma demande de prêt est rejetée, je ne veux pas forcément connaître l'ensemble des facteurs qui motivent ce refus, mais plutôt comprendre qui devrait changer (ou aurait du changer) pour obtenir le prêt. On veut le "contraste" entre ma demande, et une demande proche qui aurait été acceptée.

On parle d'explications sélectionnées quand on se sélectionne qu'une ou deux causes (potentiellement ce choix est influencé par des biais cognitifs). L'identification des causes est appelée attribution.

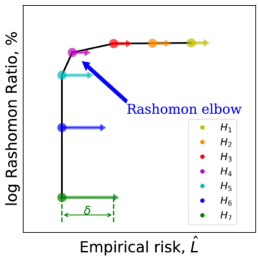
On parle d'explications sociales si elles ont lieu les d'une interaction ou d'une discussion. Elles sont alors souvent contextuelles

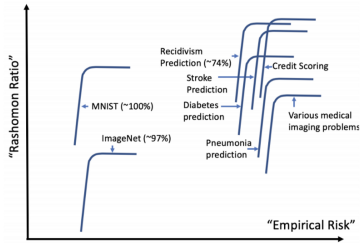
Explicabilité ? IV

(source Gauld (2021))

Explicabilité ? I

Effet Rashomon, Davis et al. (2015)





"Rashomon ratio" de Semenova et al. (2022) ou "Rashomon set", *"a set of models which all perform roughly equally well is called a Rashomon set"*, Fisher et al. (2019)

Explicabilité ? II

On parle parfois d'interprétabilité "post-hoc", Murdoch et al. (2019), car on tente de comprendre a posteriori, une fois le même construit, et calibré.

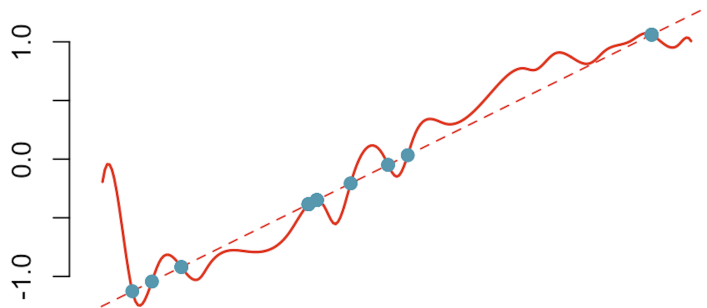
Lorsque nous sommes en mesure de choisir entre des modèles intelligibles et des modèles de boîte noire aussi performants les uns que les autres, la meilleure pratique consiste à choisir le modèle intelligible

"*When we're able to choose between equally-performing intelligible and black box models, the best practice is to choose the intelligible one*", Rudin (2019) (les informaticiens redécouvrent le principe de parcimonie).

Explicabilité ? III

Pluralitas non est ponenda sine necessitate
Entia non sunt multiplicanda praeter necessitatem,
Guillaume d'Ockham, XIVe siècle

"*la parcimonie est un principe consistant à n'utiliser que le minimum de causes élémentaires pour expliquer un phénomène*"



http://phdcomics.com

cf aussi overfit
(sur-apprentissage)

Régression et arbre (les "boites blanches") I

- Modèle de régression (régression linéaire et logistique)

Régression linéaire, $m(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$

Dans ce cas, $\beta_j = \left.\dfrac{\partial m(\boldsymbol{x})}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^*}$ (qui est ici constant, $\forall \boldsymbol{x}^*$)

Mais ce n'est plus le cas dans un modèle logistique

$$m(\boldsymbol{x}) = \frac{\exp[\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}]}{1 + \exp[\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}]}, \text{ interprété comme } \mathbb{P}[Y = 1 | \boldsymbol{X} = \boldsymbol{x}],$$

$$\left.\frac{\partial m(\boldsymbol{x})}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} = m(\boldsymbol{x}^*)[1 - m(\boldsymbol{x}^*)] \cdot \beta_j$$

Pour un modèle probit ($m(\boldsymbol{x}) = \Phi(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta})$),

$$\left.\frac{\partial m(\boldsymbol{x})}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} = \varphi(\beta_0 + \boldsymbol{x}^{*\top} \boldsymbol{\beta}) \cdot \beta_j$$

Régression et arbre (les “boites blanches”) II

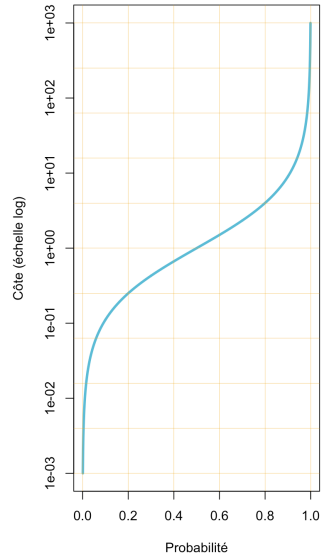
- Modèle de régression (régression linéaire et logistique)

Néanmoins, si on oublie la modélisation de la probabilité pour retenir une modélisation de la côte (odds),

$$c(\boldsymbol{x}) = \exp[\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}] \text{ interprété comme } \frac{\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]}{\mathbb{P}[Y = 0|\boldsymbol{X} = \boldsymbol{x}]},$$

Dans ce cas, $\beta_j = \left.\frac{\partial \log c(\boldsymbol{x})}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^*}$

(qui est ici constant, $\forall \boldsymbol{x}^*$)



Régression et arbre (les “boites blanches”) III

Un des avantages de la régression logistique (par rapport à de nombreuses approches boîtes noires) est qu’elle prédit une classe, mais renvoie aussi une probabilité.

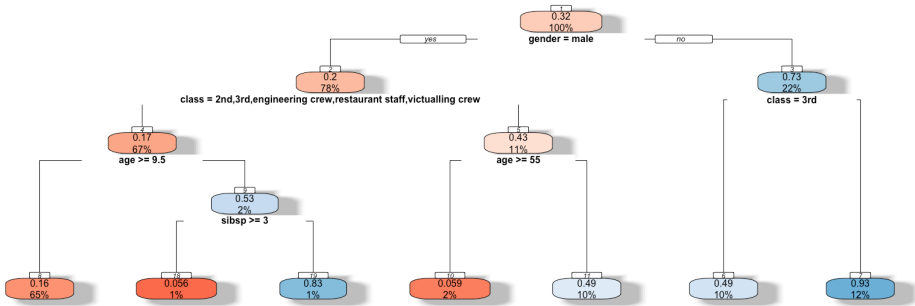
Expliquer un refus de crédit à une personne ayant une probabilité de défaut de 51% ou de 97% n’est pas pareil.

(dans les SVM, la distance à la droite de séparation est utilisé comme un score qui peut être ensuite interprété comme une probabilité - Platt scaling, Platt et al. (1999) ou isotonic regression Zadrozny and Elkan (2001, 2002) – voir aussi Niculescu-Mizil and Caruana (2005) “good probabilities")

Régression et arbre (les "boites blanches") IV

- Arbres (de régression et de classification) CART

En lien avec les "Rules-Based Explanations", Kuhn and Johnson (2013), Aasman (2019) ou van der Waa et al. (2021), "*Rule-based explanations are* if... then... *statements*" (décrit dans un arbre, pas trop profond)



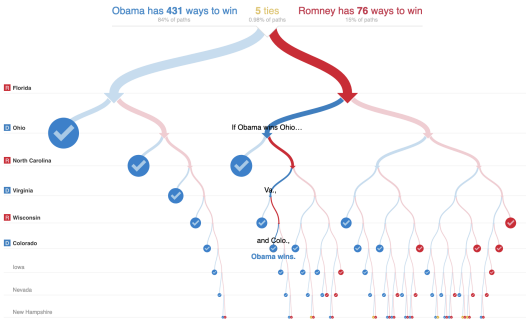
Régression et arbre (les “boites blanches”) V

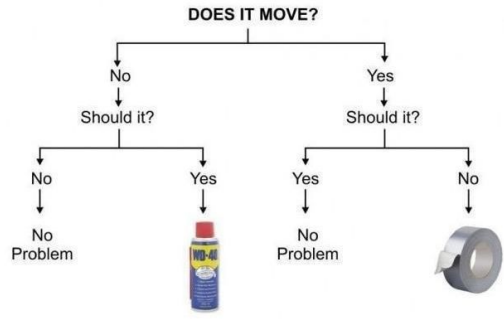
Les prédictions sont expliquées en décomposant le chemin de décision. On a de plus automatiquement à chaque noeud un contrefactuel.

L'arborescence des arbres présente une visualisation naturelle, et facile à comprendre (à condition de ne pas être trop profonds). Leur structure `si ... alors ...` est proche de la manière dont on raisonne (à condition de ne pas avoir trop de règles).

Régression et arbre (les “boites blanches”) VI

Bostock and Carter (2012) pour les élections américaines, ou l’arbre de décision des ingénieurs,

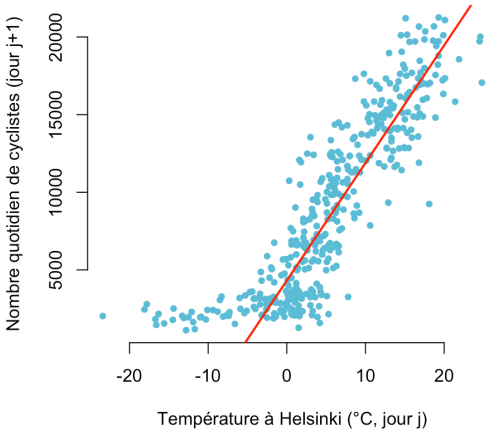


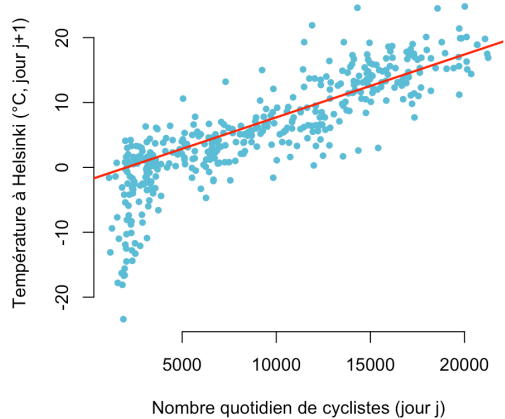


Problème de non-robustesse des arbres (voir aussi non-identifiabilité)

Régression et interprétation (1)

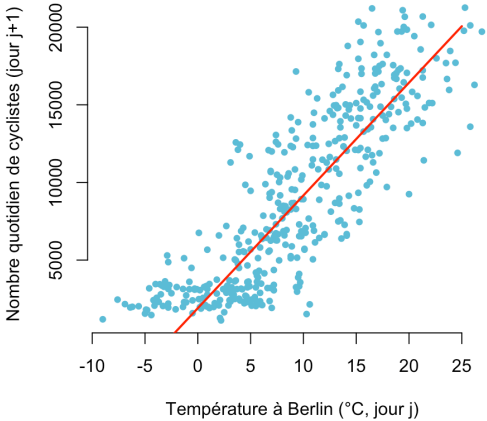
N_t : nombre de cyclistes, par jour, à Helsinki, et T_t : température moyenne journalière à Helsinki, $N_{t+1} = \alpha_0 + \alpha_1 T_t + \varepsilon_{t+1}$ ou $T_{t+1} = \beta_0 + \beta_1 N_t + \eta_{t+1}$?

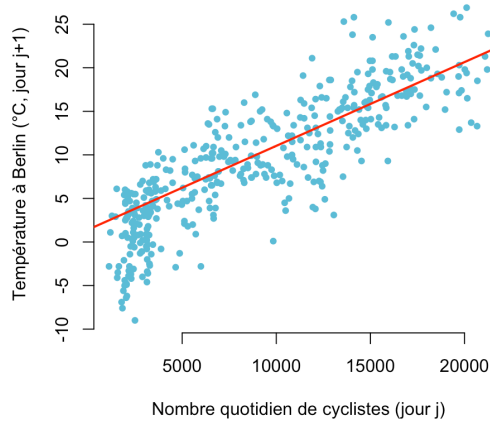




Régression et interprétation (2)

N_t : nombre de cyclistes, par jour, à Helsinki, et T'_t : température moyenne journalière à Berlin, $N_{t+1} = \alpha_0 + \alpha_1 T'_t + \varepsilon_{t+1}$ ou $T'_{t+1} = \beta_0 + \beta_1 N_t + \eta_{t+1}$?





Neural Networks (ou les "boites noires") I

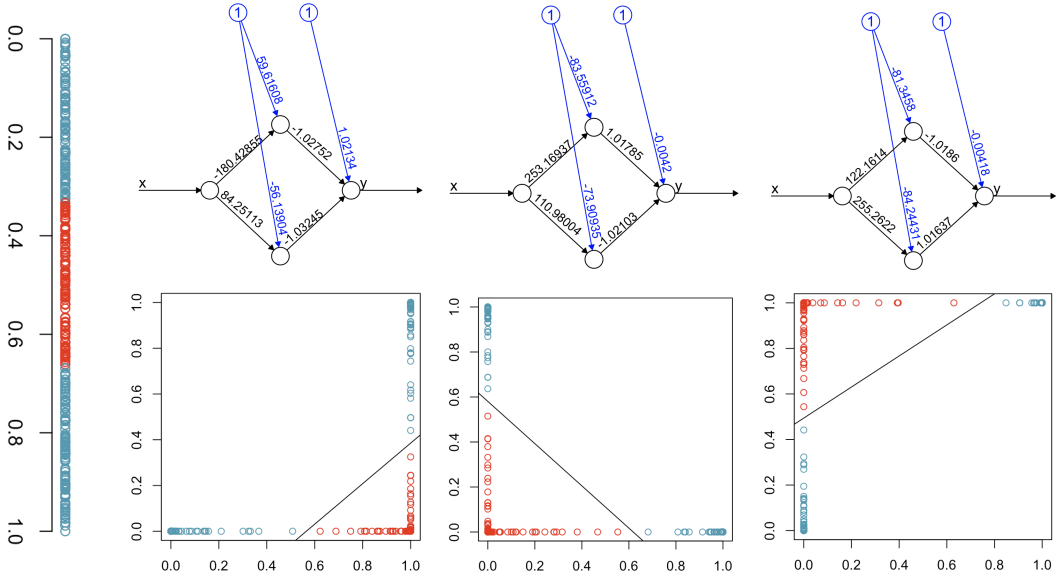
Belle and Papantonis (2021) oppose les "boîtes blanches" (transparentes), i.e. arbres et GLM aux "boîtes noires" (opaques), i.e. forêts, réseaux de neurones et SVM.

On parle parfois d'interprétabilité "post-hoc", Murdoch et al. (2019), car on tente de comprendre a posteriori, une fois le même construit, et calibré.

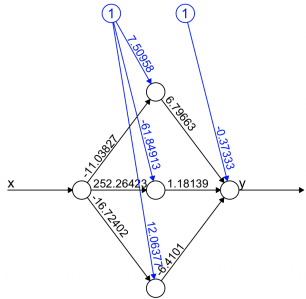
Problème de non-robustesse, Mangal et al. (2019), et de non-identifiabilité. Pour rappel, l'identiabilité d'un modèle signifie

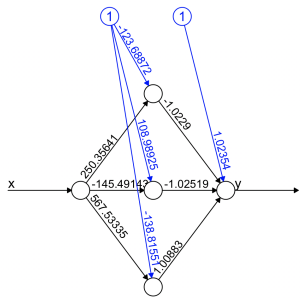
$$\theta_1 \neq \theta_2 \Longrightarrow m_{\theta_1} \neq m_{\theta_2} \text{ ou } m_{\theta_1} = m_{\theta_2} \Longrightarrow \theta_1 = \theta_2.$$

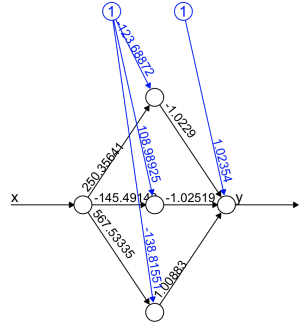
Interprétation et identifiabilité I

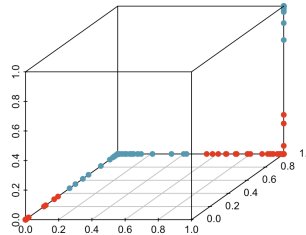


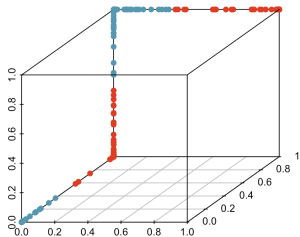
Interprétation et identifiabilité II

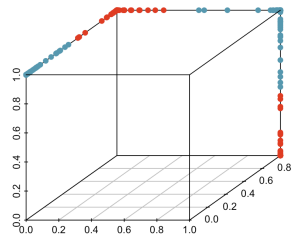












Comprendre sans prévoir, prévoir sans comprendre I

Les méthodes de simulations ont beaucoup à voir avec la théorie du chaos, "*when the present determines the future, but the approximate present does not approximately determine the future*" Edward Lorenz, cité par May (2004)

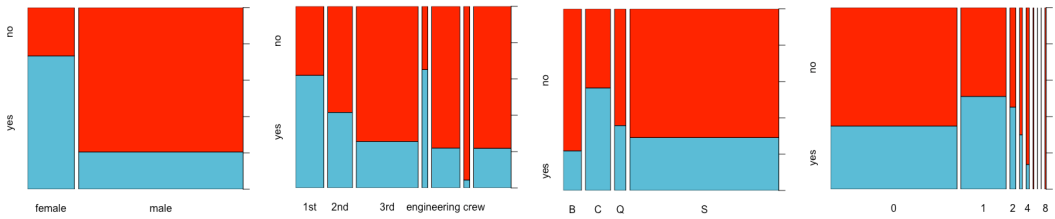
comprendre sans prévoir

Les algorithmes de Monte Carlo sont basés sur des modèles congruentiels ($x_n = f(x_{n-1})$),
... déterministes mais impossible à prévoir
Alors que les réseaux de neurones ...

prévoir sans comprendre

discuté par Thom (1991)
souvent observé pour des algorithmes neuronaux profonds

Mise en oeuvre pratique, les données Titanic I



On va estimer plusieurs modèles sur les données 2195 (2207) observations, $p = 8$, pour prédire la survie ($y \in \{\text{no}, \text{yes}\}$, taux de survie $\sim 32\%$)

- régression logistique (glm) avec lissage de l'âge (gam)
- arbre de classification (cart) et forêt aléatoire (rf)
- boosting (gbm) et support vecteur machine (svm)

On tente de comprendre la prévision pour deux personnes test, Kate et Leonardo,

Mise en oeuvre pratique, les données Titanic II

Les 6 modèles donnent non seulement une classification (survie ou pas) mais aussi un score, interprété comme une "probabilité de survie".

On va considérer deux personnes non-présentes dans la base d'apprentissage

- Kate (Miss. Winslet), femme, 17 ans, 1ère classe, 1 frère, 2 parents
- Leonardo (Mr. DiCaprio), homme, 20 ans, 3ème classe, ni frère, ni parents

	glm	gam	cart	rf	gbm	svm
Kate	93.4%	91.3%	92.7%	85.4%	84.9%	86.0%
Leonardo	12.5%	11.1%	15.6%	0.0%	7.2%	17.6%

- Pourquoi Kate a de (très) grandes chances de survie ?
- Pourquoi Leonardo a de (très) faibles chances de survie ?

Ceteris paribus vs. Mutatis mutandis I

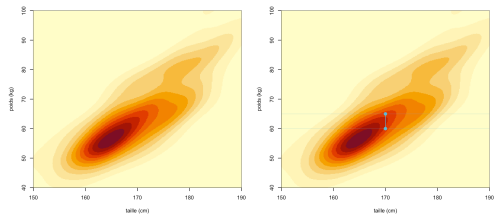
Considérons un modèle surpoids $= m(\text{poids}, \text{taille})$

Ceteris paribus

Ceteris paribus (*ceteris paribus sic stantibus*) est la locution latine se traduisant par *toutes choses étant égales par ailleurs*.

On pourrait regarder

$$m(\text{poids} = x + dx, \text{taille} = y) - m(\text{poids} = x, \text{taille} = y)$$



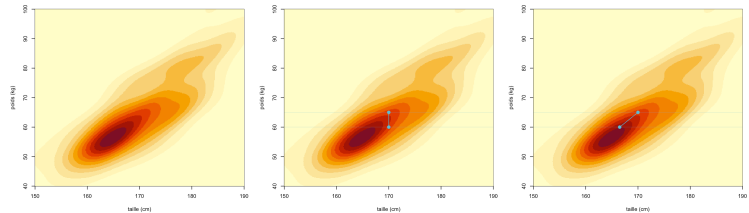
Ceteris paribus vs. Mutatis mutandis II

Mutatis mutandis

Mutatis mutandis est la locution latine se traduisant par *ce qui devait être changé ayant été changé* ou *une fois effectuées les modifications nécessaires*.

On devrait tenir compte du fait qu'un individu de poids différent serait aussi, probablement, de taille différente

$$m(\text{poids} = x + dx, \text{taille} = y + \epsilon) - m(\text{poids} = x, \text{taille} = y)$$



Ceteris paribus vs. Mutatis mutandis III

Modèle Gaussien et espérances conditionnelles

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

L'espérance conditionnelle vaut $\mathbb{E}_{X_1}[X_2|x_1^*] = \mu_2 + \frac{r\sigma_2}{\sigma_1}(x_1^* - \mu_1)$

Et on notera $\mathbb{E}_{X_1^\perp}[X_2|x_1^*] = \mathbb{E}[X_2] = \mu_2$.

$$\mathbb{E}_{X_1^\perp}[h(X_1, X_2)|x_1^*] \approx \frac{1}{n}\sum_{i=1}^{n} h(x_1^*, x_{i,2})$$

$$\mathbb{E}_{X_1}[h(X_1, X_2)|x_1^*] \approx \frac{1}{\|\mathcal{V}_\epsilon(x_1^*)\|}\sum_{i \in \mathcal{V}_\epsilon(x_1^*)} h(x_1^*, x_{i,2}), \text{ où } \mathcal{V}_\epsilon(x_1^*) = \left\{ i : |x_{i,1} - x_1^*| \leq \epsilon \right\}$$

Ceteris paribus vs. Mutatis mutandis IV

Soit $(X_1, X_2, \varepsilon)^\top$ un vecteur Gaussien,

$$\begin{pmatrix} X_1 \\ X_2 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 & 0 \\ r\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right)$$

Posons $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

$$\mathbb{E}_{\boldsymbol{X}}[Y|\boldsymbol{x}^*] = \mathbb{E}_{\boldsymbol{X}}[Y|x_1^*, x_2^*] = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^*$$

$$\mathbb{E}_{X_1}[Y|x_1^*] = \beta_0 + \beta_1 x_1^* + \beta_2 \left(\mu_2 + \frac{r\sigma_2}{\sigma_1}(x_1^* - \mu_1) \right) \text{ mutatis mutandis}$$

$$\mathbb{E}_{X_1^\perp}[Y|x_1^*] = \beta_0 + \beta_1 x_1^* + \beta_2 \mu_2 \text{ ceteris paribus}$$

de telle sorte que $\mathbb{E}_{X_1}[Y|x_1^*] = \mathbb{E}_{X_1^\perp}[Y|x_1^*] + \beta_2 \frac{r\sigma_2}{\sigma_1}(x_1^* - \mu_1)$

$$\mathbb{E}[Y] = \beta_0 + \beta_1 \mu_1 + \beta_2 \mu_2$$

Variable Importance I

Saltelli et al. (2008), Helton and Davis (2002), Azen and Budescu (2003) ou Rifkin and Klautau (2004) pour les arbres et les forêts aléatoires. Fisher et al. (2019) a proposé

VI_j ou "permutation VI_j"

Étant donnée une fonction de perte ℓ,

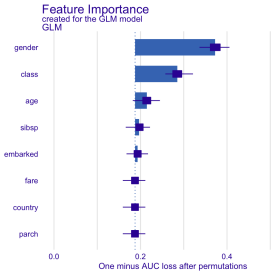
$$VI_j = \mathbb{E}\big[\ell(Y, m(\boldsymbol{X}_{-j}, X_j))\big] - \mathbb{E}\big[\ell(Y, m(\boldsymbol{X}_{-j}, X_j^{\perp}))\big]$$

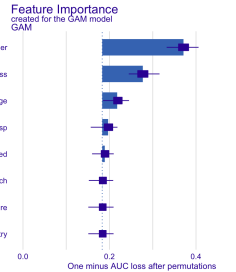
et la version empirique est

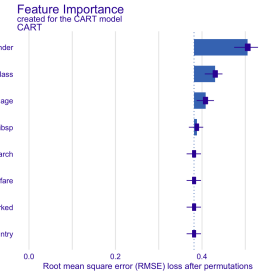
$$\widehat{VI}_j = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_{i,-j}, x_{i,j})) - \ell(y_i, m(\boldsymbol{x}_{i,-j}, \tilde{x}_{i,j}))$$

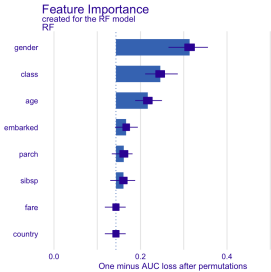
pour une permutation $\tilde{\boldsymbol{x}}_j$ de $\boldsymbol{x}_j$.

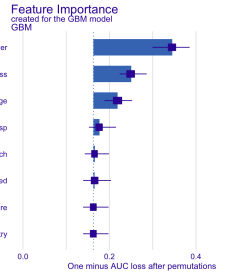
Variable Importance II

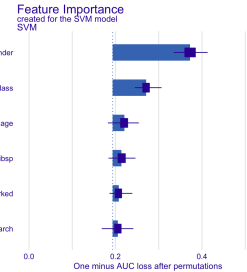












Variable Importance III

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.0$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance IV

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.1$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance V

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.2$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance VI

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.3$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance VII

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.4$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance VIII

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.5$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance IX

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.6$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance X

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.7$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance XI

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.8$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance XII

On simule un modèle linéaire ($n = 1000$)

$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.9$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire

Variable Importance XIII

- ✚ approche “agnostique”, indépendante du modèle utilisé
  relativement simple à lire
- ⛔ explication seulement globale

ICE (Ceteris-paribus) I

Mais au lieu de mesure globale, on va considérer maintenant des mesures locales. Goldstein et al. (2015) a parlé de ICE *individual conditional expectation*

$z \mapsto m_{\boldsymbol{x}^*,j}(z)$

Ceteris paribus (*ceteris paribus sic stantibus*) est la locution latine se traduisant par *toutes choses étant égales par ailleurs*, on considère la fonction sur $\mathcal{X}_j$

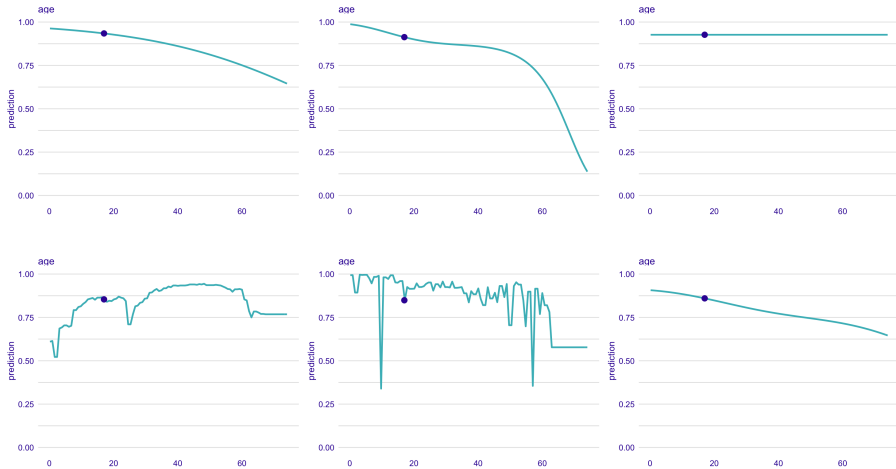
$$z \mapsto m_{\boldsymbol{x}^*,j}(z) = m(\boldsymbol{x}^*_{-j}, z) = m(x_1^*, \cdots, x_{j-1}^*, z, x_{j+1}^*, \cdots, x_p^*)$$

en un point $\boldsymbol{x}^* \in \mathcal{X}$.

	glm	gam	cart	rf	gbm	svm
Kate	93.4%	91.3%	92.7%	85.4%	84.9%	86.0%
Leonardo	12.5%	11.1%	15.6%	0.0%	7.2%	17.6%

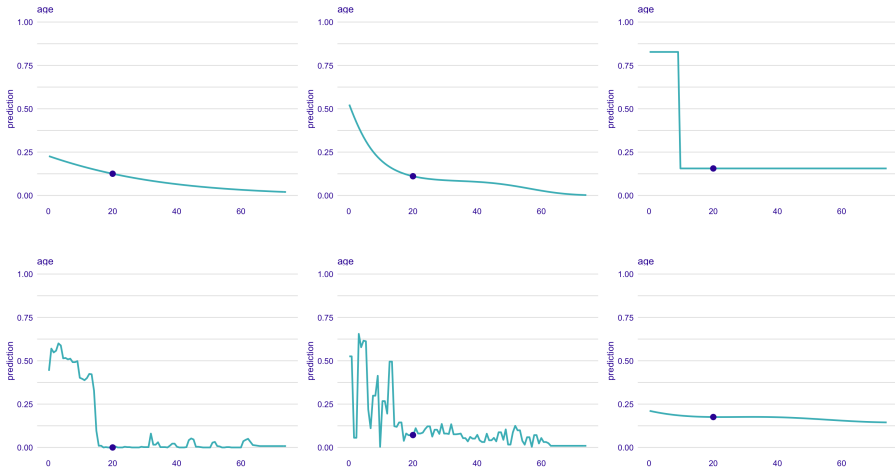
ICE (Ceteris-paribus) II

$\boldsymbol{x}^* =$ Kate, $j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



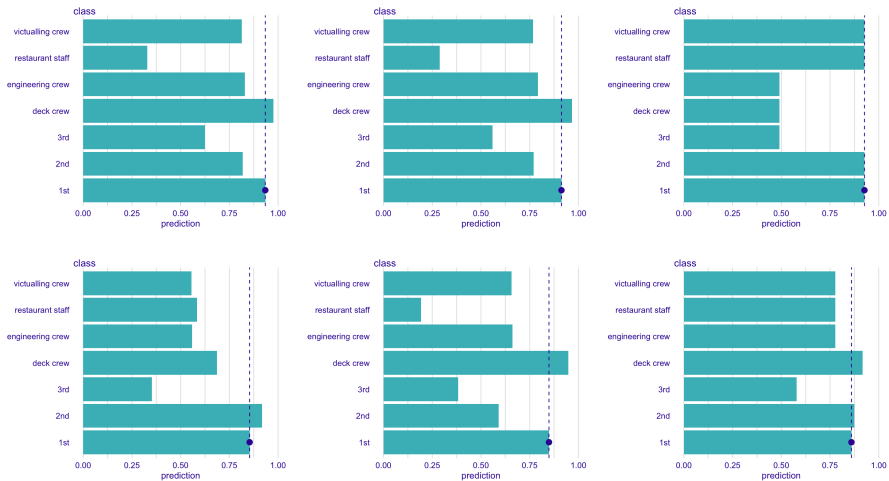
ICE (Ceteris-paribus) III

$\boldsymbol{x}^*$ = Leonardo, j = âge, $m \in \{\text{glm, gam, cart, rf, gbm, svm}\}$



ICE (Ceteris-paribus) IV

$\boldsymbol{x}^* = \text{Kate}$, $j = \text{classe}$, $m \in \{\text{glm, gam, cart, rf, gbm, svm}\}$



ICE (Ceteris-paribus) V

$\boldsymbol{x}^* =$ Leonardo, $j =$ classe, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$

ICE (Ceteris-paribus) VI

On peut alors regarder la fonction de déviation $z \mapsto \delta m_{\boldsymbol{x}^*,j}(z) = m_{\boldsymbol{x}^*,j}(z) - m_{\boldsymbol{x}^*,j}(x_j^*)$

$dm_j^{cp}(\boldsymbol{x}^*)$

La déviation moyenne absolue de la $j$ème variable, en $\boldsymbol{x}^*$, est $dm_j(\boldsymbol{x}^*)$,

$$dm_j(\boldsymbol{x}^*) = \mathbb{E}\big[|\delta m_{\boldsymbol{x}^*,j}(X_j)|\big] = \mathbb{E}\big[|m(\boldsymbol{x}^*_{-j}, X_j) - m(\boldsymbol{x}^*_{-j}, x_j^*)|\big]$$

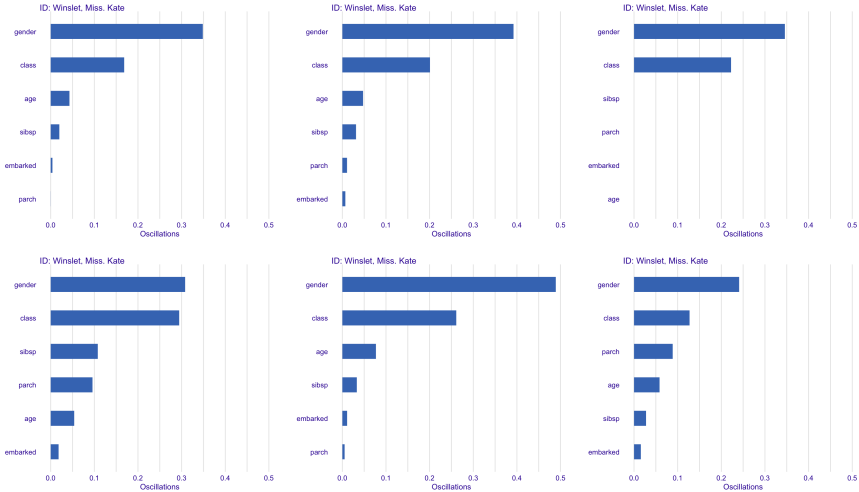
$\widehat{dm}_j^{cp}(\boldsymbol{x}^*)$

La déviation moyenne absolue empirique de la $j$ème variable, en $\boldsymbol{x}^*$, est

$$\widehat{dm}_j(\boldsymbol{x}^*) = \frac{1}{n}\sum_{i=1}^{n} |m(\boldsymbol{x}^*_{-j}, x_{i,j}) - m(\boldsymbol{x}^*_{-j}, x_j^*)|$$

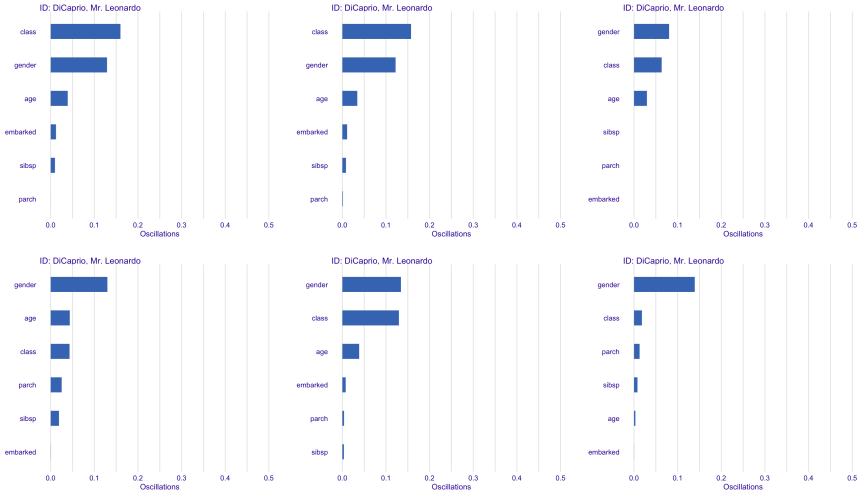
ICE (Ceteris-paribus) VII

$\boldsymbol{x}^* =$ Kate, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



ICE (Ceteris-paribus) VIII

$\boldsymbol{x}^* =$ Leonardo, $m \in \{\text{glm, gam, cart, rf, gbm, svm}\}$



Break-down I

Pour un modèle linéaire

$$\widehat{m}(\boldsymbol{x}^*) = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}^* = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_j^* = \overline{y} + \sum_{j=1}^{p} \underbrace{\widehat{\beta}_j \big(x_j^* - \overline{x}_j\big)}_{=v_j(\boldsymbol{x}^*)}$$

où $v_j(\boldsymbol{x}^*)$ est la contribution de la variable j dans la prédiction pour $\boldsymbol{x}^*$.

Plus généralement, Robnik-Šikonja and Kononenko (1997), (2003) and (2008), la contribution de la $j$ème variable, en $\boldsymbol{x}^*$, est

$$v_j(\boldsymbol{x}^*) = m(x_1^*, \cdots, , x_{j-1}^*, x_j^*, x_{j+1}^*, \cdots, x_p^*) - \mathbb{E}_{\boldsymbol{X}_{-j}^\perp}[m(x_1^*, \cdots, x_{j-1}^*, X_j, x_{j+1}^*, \cdots, x_p^*)]$$

de telle sorte que

$$m(\boldsymbol{x}^*) = \mathbb{E}\big[m(\boldsymbol{X})\big] + \sum_{j=1}^{p} v_j(\boldsymbol{x}^*)$$

Break-down II

donc pour un modèle linéaire $v_j(\boldsymbol{x}^*) = \beta_j\big(x_j^* - \mathbb{E}_{\boldsymbol{X}_{-j}^\perp}[X_j]\big)$ et $\widehat{v}_j(\boldsymbol{x}^*) = \widehat{\beta}_j\big(x_j^* - \overline{x}_j\big)$.

Mais plus généralement $v_j(\boldsymbol{x}^*) = m(\boldsymbol{x}^*) - \mathbb{E}_{\boldsymbol{X}_{-j}^\perp}[m(\boldsymbol{x}_{-j}^*, X_j))]$, où on peut aussi écrire $m(\boldsymbol{x}^*)$ sous la forme $\mathbb{E}_{\boldsymbol{X}}[m(\boldsymbol{x}^*)]$, i.e.

$$v_j(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}_{\boldsymbol{X}}\big[m(\boldsymbol{X})\big|x_1^*, \cdots, x_p^*\big] - \mathbb{E}_{\boldsymbol{X}_{-j}^\perp}\big[m(\boldsymbol{X})\big|x_1^*, \cdots, x_{j-1}^*, x_{j+1}^*, \cdots, x_p^*\big] \\ \mathbb{E}_{\boldsymbol{X}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*\big] - \mathbb{E}_{\boldsymbol{X}_{-j}^\perp}\big[m(\boldsymbol{X})\big|\boldsymbol{x}_{-j}^*\big] \end{cases}$$

$\gamma_j^{bd}(\boldsymbol{x}^*)$

La contribution de la $j$ème variable, en $\boldsymbol{x}^*$, est

$$\gamma_j^{bd}(\boldsymbol{x}^*) = v_j(\boldsymbol{x}^*) = \mathbb{E}_{\boldsymbol{X}}\big[m(\boldsymbol{X})\big|\boldsymbol{x}^*\big] - \mathbb{E}_{\boldsymbol{X}_{-j}^\perp}\big[m(\boldsymbol{X})\big|\boldsymbol{x}_{-j}^*\big]$$

Break-down III

"*In other words, the contribution of the j-th variable is the difference between the expected value of the model's prediction conditional on setting the values of the first j variables equal to their values in $\boldsymbol{x}^*$ and the expected value conditional on setting the values of the first $j-1$ variables equal to their values in $\boldsymbol{x}^*$*", Biecek and Burzykowski (2021).

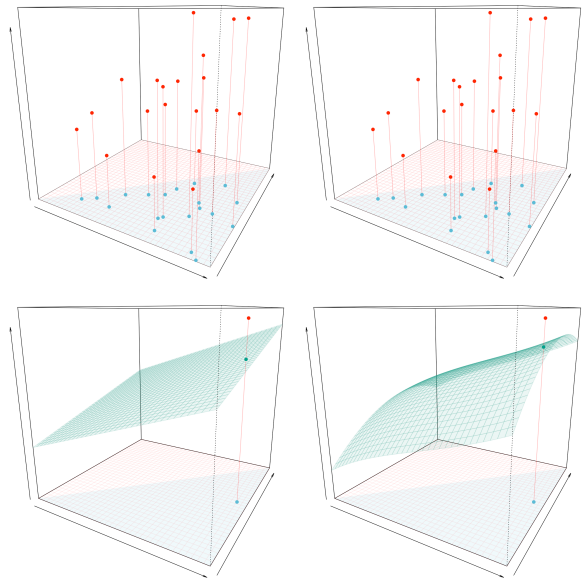
- ⊕ approche "agnostique", indépendante du modèle utilisé
  approche intuitive et naturelle pour les modèles linéaires
  pas trop complexe, numériquement
- ⊖ ceteris paribus, pas de prise en compte d'intéractions entre les variables $\boldsymbol{x}$

Break-down IV

On a 20 points, et 2 modèles
avec 2 variables explicatives x_1, x_2

	m_{L}	m_{NL}
constante	2.008	2.043
x_1^* (0.9447)	0.322	0.326
x_2^* (0.5045)	0.184	0.336
$\widehat{y}$	2.514	2.705

avec $\overline{x}_1 = 0.669$ et $\overline{x}_2 = 0.289$,
et $\widehat{y} = m(\boldsymbol{x}^*)$

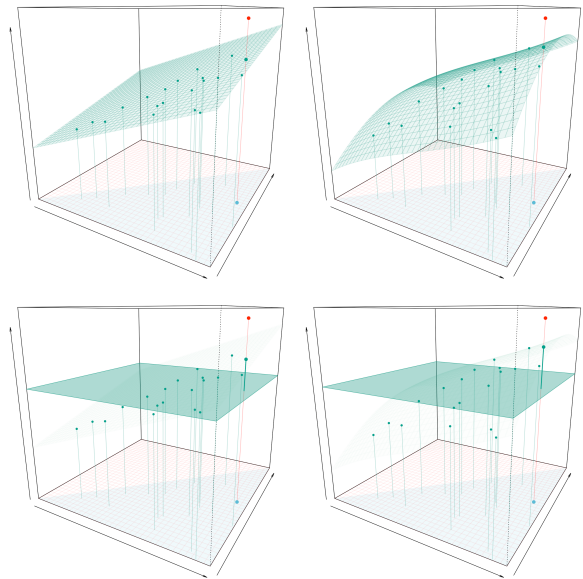


Break-down V

pour le terme constant,

$$\mathbb{E}\big[m(\boldsymbol{X})\big] \approx \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i)$$

	m_{L}	m_{NL}
constante	2.008	2.043



Break-down VI

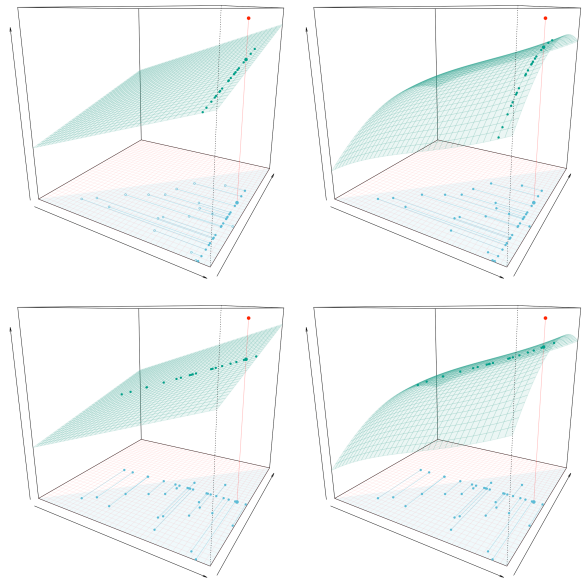
Pour les contributions

$$\mathbb{E}_{X_2^\perp}\left[m(X_1, x_2^*)\right] \approx \frac{1}{n}\sum_{i=1}^{n} m(x_{1,i}, x_2^*)$$

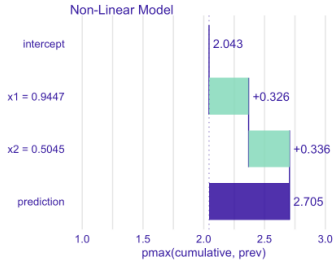
$$\mathbb{E}_{X_1^\perp}\left[m(x_1^*, X_2)\right] \approx \frac{1}{n}\sum_{i=1}^{n} m(x_1^*, x_{2,i})$$

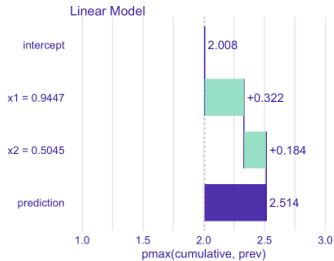
	m_{L}	m_{NL}
x_1^* (0.9447)	0.322	0.326

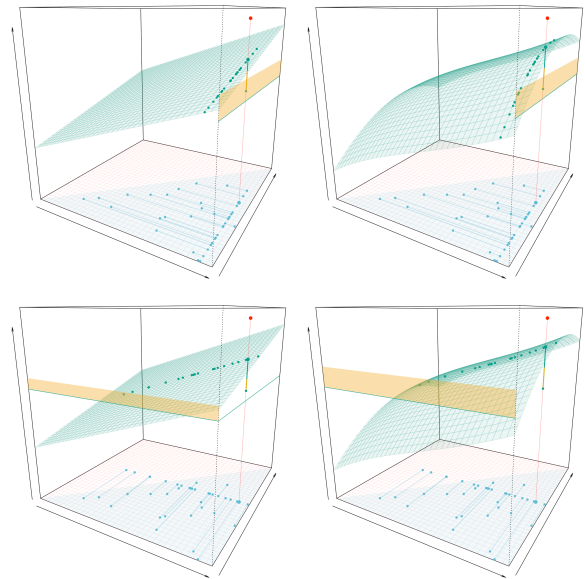
	m_{L}	m_{NL}
x_2^* (0.5045)	0.184	0.336



Break-down VII

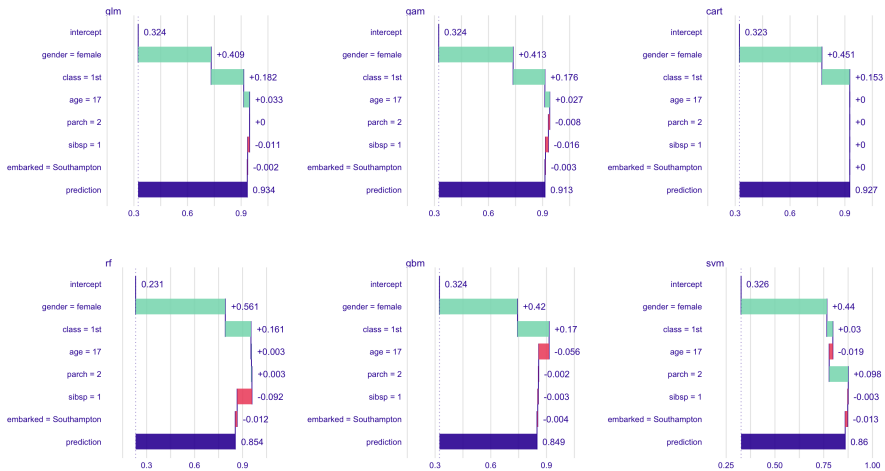






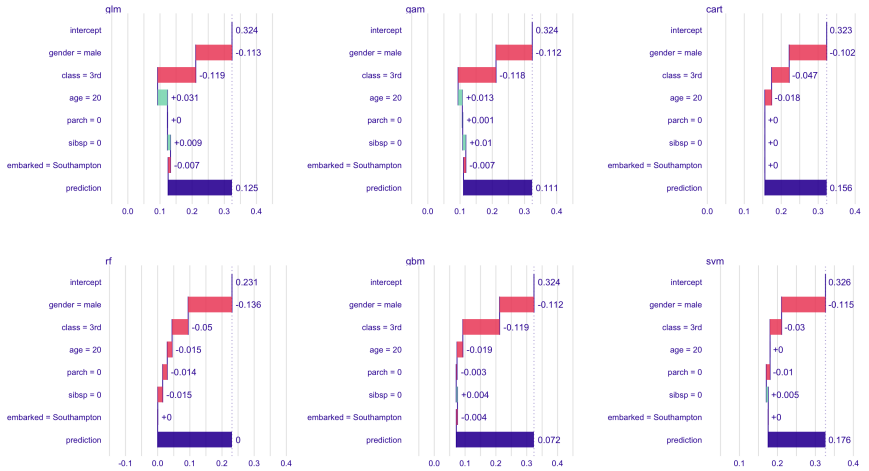
Break-down VIII

$\boldsymbol{x}^* =$ Kate, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Break-down IX

$\boldsymbol{x}^*$ = Leonardo, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Break-down X

On simule un modèle linéaire ($n = 1000$)

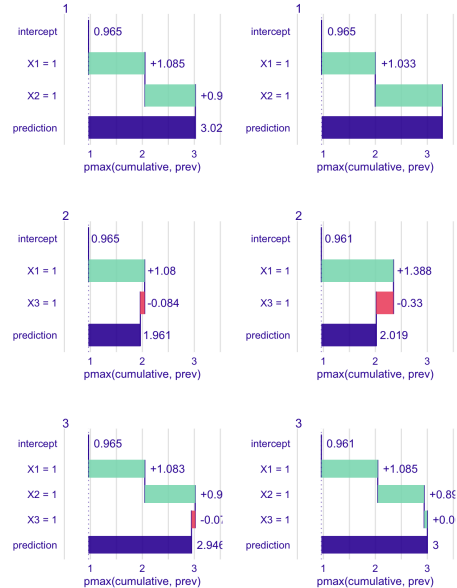
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.0$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XI

On simule un modèle linéaire ($n = 1000$)

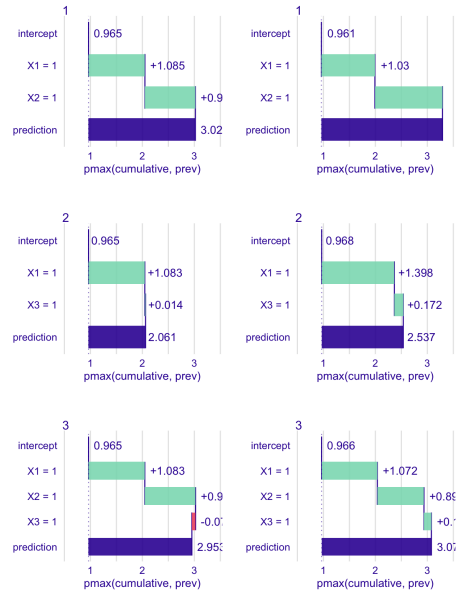
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.1$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XII

On simule un modèle linéaire ($n = 1000$)

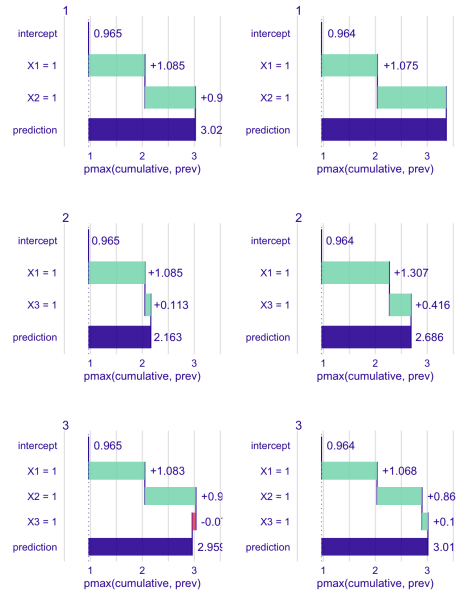
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.2$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XIII

On simule un modèle linéaire ($n = 1000$)

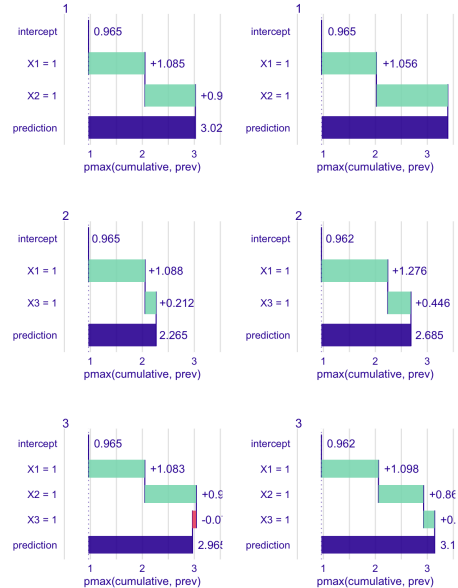
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.3$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XIV

On simule un modèle linéaire ($n = 1000$)

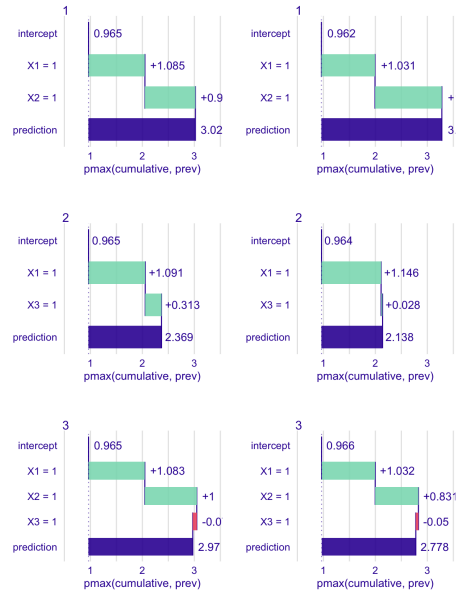
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.4$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XV

On simule un modèle linéaire ($n = 1000$)

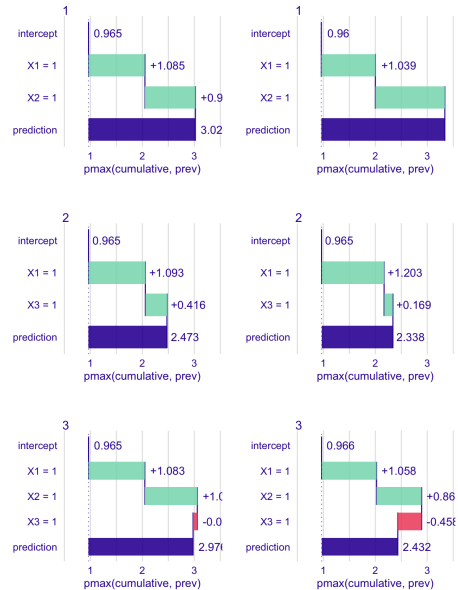
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.5$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XVI

On simule un modèle linéaire ($n = 1000$)

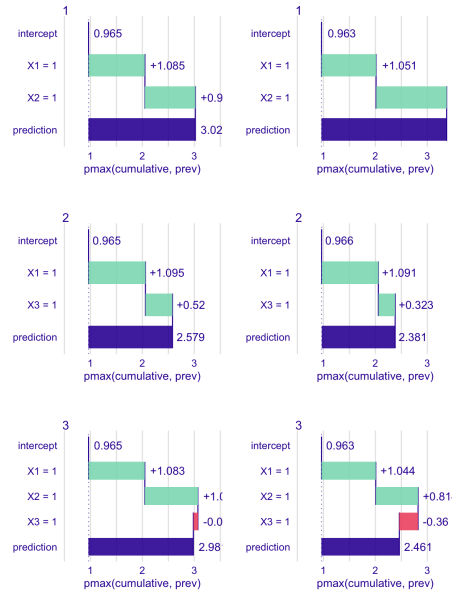
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.6$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XVII

On simule un modèle linéaire ($n = 1000$)

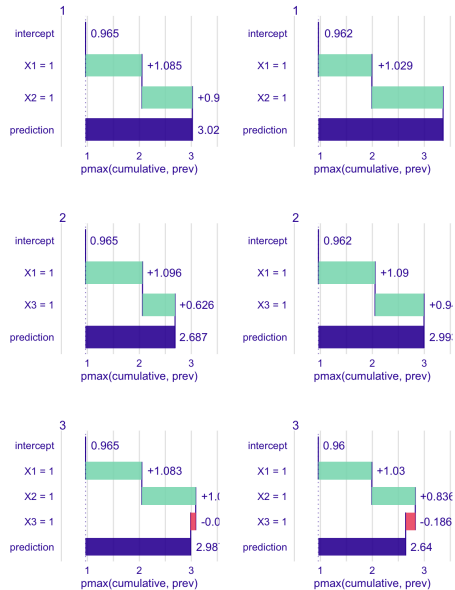
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.7$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XVIII

On simule un modèle linéaire ($n = 1000$)

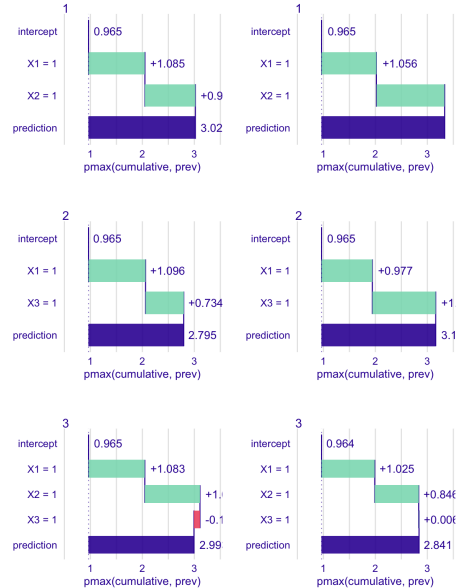
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.8$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XIX

On simule un modèle linéaire ($n = 1000$)

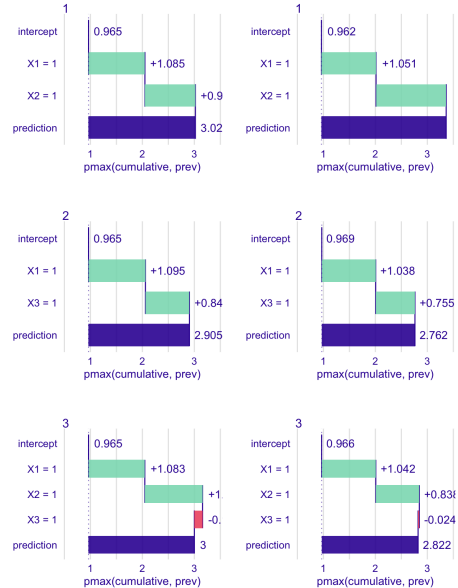
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|cc|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.9$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Break-down XX

On peut réécrire la contribution de la $j$ème variable, en $\boldsymbol{x}^*$,

$$v_j(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}_{\boldsymbol{X}}\left[m(\boldsymbol{X})\middle|x_1^*, \cdots, x_p^*\right] - \mathbb{E}_{\boldsymbol{X}_{-j}^{\perp}}\left[m(\boldsymbol{X})\middle|x_1^*, \cdots, x_{j-1}^*, x_{j+1}^*, \cdots, x_p^*\right] \\ \mathbb{E}_{\boldsymbol{X}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*\right] - \mathbb{E}_{\boldsymbol{X}_{-j}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{-j}^*\right] \end{cases}$$

$\Delta_{j|S}(\boldsymbol{x}^*)$

La contribution de la $j$ème variable conditionnelle à un groupe de variables $S \subset \{1, \cdots, p\} \backslash \{j\}$, en $\boldsymbol{x}^*$, est

$$\Delta_{j|S}(\boldsymbol{x}^*) = \begin{cases} \mathbb{E}_{\boldsymbol{X}_{\bar{S}}^{\perp}, X_j^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_S^*, x_j^*\right] - \mathbb{E}_{\boldsymbol{X}_{\bar{S}}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_S^*\right] \\ \mathbb{E}_{\boldsymbol{X}_{\bar{S}\cup\{j\}}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_{S\cup\{j\}}^*\right] - \mathbb{E}_{\boldsymbol{X}_{\bar{S}}^{\perp}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}_S^*\right] \end{cases}$$

Break-down XXI

Aussi, pour $S \subset \{1, \cdots, p\} \backslash \{j\}$

$$\Delta_{j|S}(\boldsymbol{x}^*) = \mathbb{E}_{\boldsymbol{X}_{\overline{S \cup \{j\}}}} \left[m(\boldsymbol{X}) \middle| \boldsymbol{x}^*_{S \cup \{j\}} \right] - \mathbb{E}_{\boldsymbol{X}_{\overline{S}}} \left[m(\boldsymbol{X}) \middle| \boldsymbol{x}^*_{S} \right]$$

de telle sorte que $v_j(\boldsymbol{x}^*) = \Delta_{j|\{1,2,\cdots,p\} \backslash \{j\}} = \Delta_{j|-j}$.

On peut aussi définir $\Delta_{K|S}(\boldsymbol{x}^*)$, ou $\Delta_{i,j|S}(\boldsymbol{x}^*)$ (interractions).

Shapley Additive Explanations (SHAP) for Average Attributions I

$\forall S \subseteq \{1, \ldots, p\}$, on dispose d'une fonction $\mathcal{V}(S)$,

On cherche des contributions $\phi_j(\mathcal{V})$ – on notera $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)$ – qui vérifient

- efficiency: $\sum_{j=1}^{p} \phi_j(\mathcal{V}) = \mathcal{V}(\{1, \ldots, p\})$
- symmetry: si $\mathcal{V}(S \cup \{j\}) = \mathcal{V}(S \cup \{k\})$ $\forall S \subseteq \{1, \ldots, p\} \setminus \{j, k\}$, alors $\phi_j = \phi_k$
- dummy: si $\mathcal{V}(S \cup \{j\}) = \mathcal{V}(S)$ $\forall S \subseteq \{1, \ldots, p\}$, alors $\phi_j = 0$
- additivity: si $\mathcal{V}^{(1)}$ et $\mathcal{V}^{(2)}$ ont pour décomposition $\boldsymbol{\phi}^{(1)}$ et $\boldsymbol{\phi}^{(2)}$, alors $\mathcal{V}^{(1)} + \mathcal{V}^{(2)}$ a pour décomposition $\phi^{(1)} + \phi^{(2)}$

Shapley Additive Explanations (SHAP) for Average Attributions II

Shapley (1953) a montré que la seule contribution qui vérifie ces conditions est

$$\phi_j(\mathcal{V}) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(\mathcal{V}\left(S \cup \{j\}\right) - \mathcal{V}(S)\right)$$

$$\phi_j(\mathcal{V}) = \frac{1}{p} \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \binom{p-1}{|S|}^{-1} \left(\mathcal{V}(S \cup \{j\}) - \mathcal{V}(S)\right)$$

Shapley Additive Explanations (SHAP) for Average Attributions III

On va utiliser ici $\mathcal{V}(S) = \mathbb{E}_{\boldsymbol{X}_{\bar{S}}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_S\right]$

$\gamma_j^{shap}(\boldsymbol{x}^*)$

La contribution de la $j$ème variable, en $\boldsymbol{x}^*$, est

$$\gamma_j^{shap}(\boldsymbol{x}^*) = \frac{1}{p}\sum_{S\subseteq\{1,\ldots,p\}\setminus\{j\}}\binom{p-1}{|S|}^{-1}\Delta_{j|S}(\boldsymbol{x}^*)$$

- local accuracy: $\sum_{j=1}^{p}\gamma_j^{shap}(\boldsymbol{x}^*) = m(\boldsymbol{x}^*) - \mathbb{E}\left[m(\boldsymbol{X})\right]$
- symmétrie: si j et k sont interchangeables, $\gamma_j^{shap}(\boldsymbol{x}^*) = \gamma_k^{shap}(\boldsymbol{x}^*)$
- dummy: si X_j ne contribue jamais, $\gamma_j^{shap}(\boldsymbol{x}^*) = 0$

Shapley Additive Explanations (SHAP) for Average Attributions IV

Si $p = 2$, $\gamma_1^{shap}(\boldsymbol{x}^*) = \Delta_{1|2}(\boldsymbol{x}^*) = \gamma_1^{bd}(\boldsymbol{x}^*)$

Si $p \gg 2$, les calculs peuvent vite devenir lourds. Štrumbelj and Kononenko (2014) ont proposé une méthode par simulations.
À partir de $\boldsymbol{x}^*$ et d'un individu $\boldsymbol{x}_i$, on construit

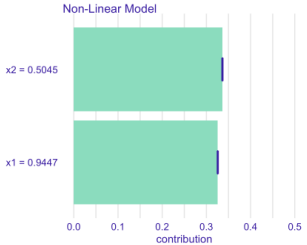
$$\tilde{x}_{i,j'} = \begin{cases} x_{j'}^* \text{ avec probabilité } 1/2 \\ x_{i,j'} \text{ avec probabilité } 1/2 \end{cases} \quad \text{et} \quad \begin{cases} \boldsymbol{x}_i^{*+} = (\tilde{x}_{i,1}, \cdots, x_j^*, \cdots, \tilde{x}_{i,p}) \\ \boldsymbol{x}_i^{*-} = (\tilde{x}_{i,1}, \cdots, x_{i,j}, \cdots, \tilde{x}_{i,p}) \end{cases}$$

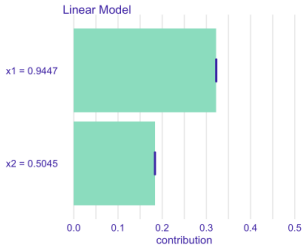
et on note que $\gamma_j^{shap}(\boldsymbol{x}^*) \approx m(\boldsymbol{x}_i^{*+}) - m(\boldsymbol{x}_i^{*-})$, et donc

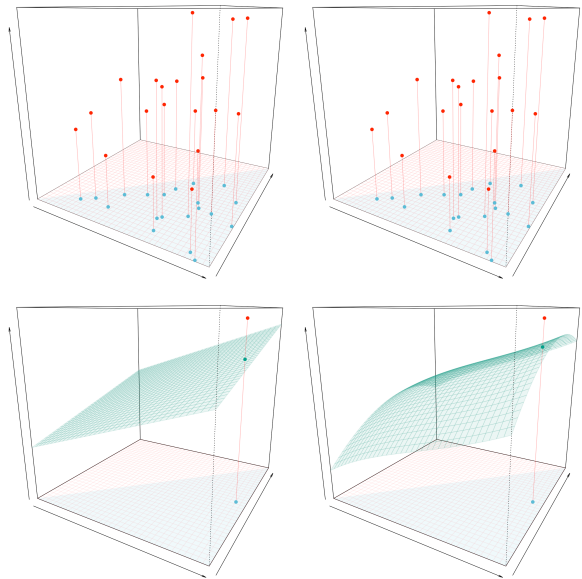
$$\widehat{\gamma}_j^{shap}(\boldsymbol{x}^*) = \frac{1}{s} \sum_{i \in \{1, \cdots, n\}} m(\boldsymbol{x}_i^{*+}) - m(\boldsymbol{x}_i^{*-})$$

(on tire à chaque étape un individu i dans la base d'apprentissage, s fois).

Shapley Additive Explanations (SHAP) for Average Attributions V







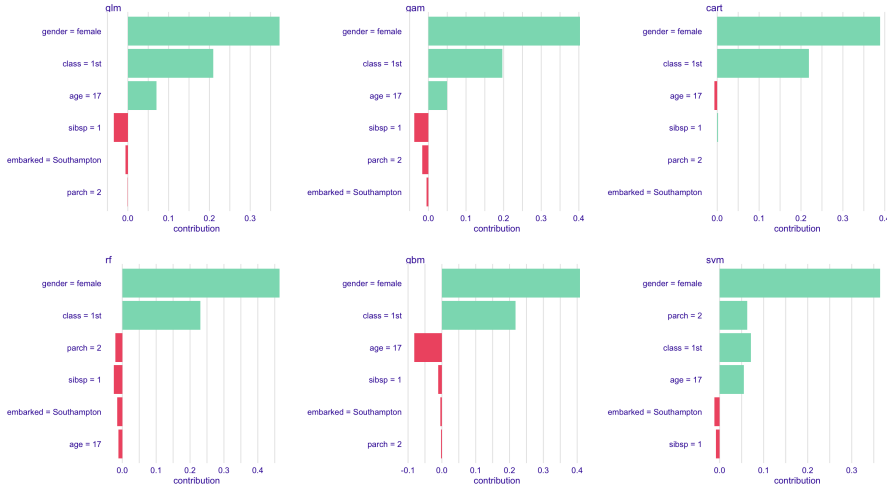
Shapley Additive Explanations (SHAP) for Average Attributions VI

Pour les données Titanic, pour Leonardo (break-down vs. Shapley contributions)

	glm	gam	cart	rf	glm	gam	cart	rf
intercept	0.324	0.324	0.323	0.231	0.324	0.324	0.323	0.231
age = 20	0.031	0.013	-0.018	-0.015	0.045	0.025	-0.011	-0.018
class = 3rd	-0.119	-0.118	-0.047	-0.050	-0.134	-0.132	-0.069	-0.070
embarked = S	-0.007	-0.007	0.000	0.000	-0.007	-0.008	0.000	-0.013
gender = male	-0.113	-0.112	-0.102	-0.136	-0.115	-0.113	-0.089	-0.120
parch = 0	0.000	0.001	0.000	-0.014	0.000	0.002	0.000	-0.010
sibsp = 0	0.009	0.010	0.000	-0.015	0.012	0.013	0.003	-0.001
prediction	0.125	0.111	0.156	0.000	0.125	0.111	0.156	0.000

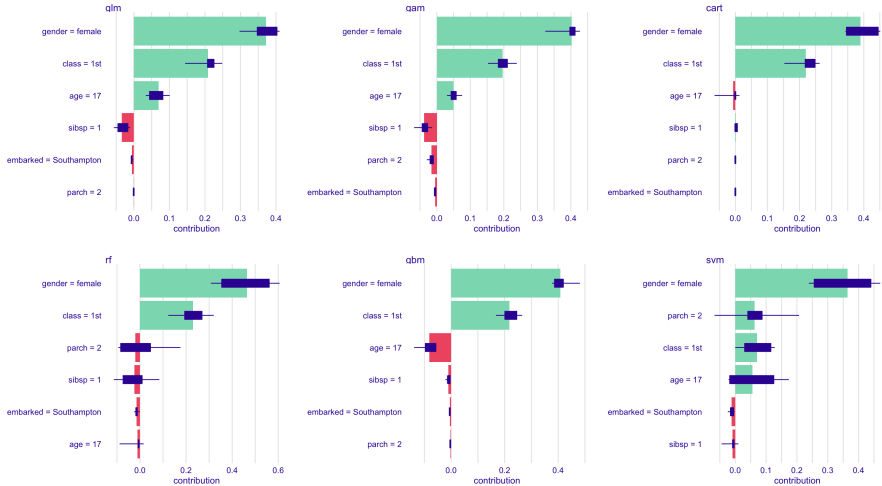
Shapley Additive Explanations (SHAP) for Average Attributions VII

$\boldsymbol{x}^* =$ Kate, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Shapley Additive Explanations (SHAP) for Average Attributions VIII

$\boldsymbol{x}^* =$ Kate, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$ (with confidence boxplot)

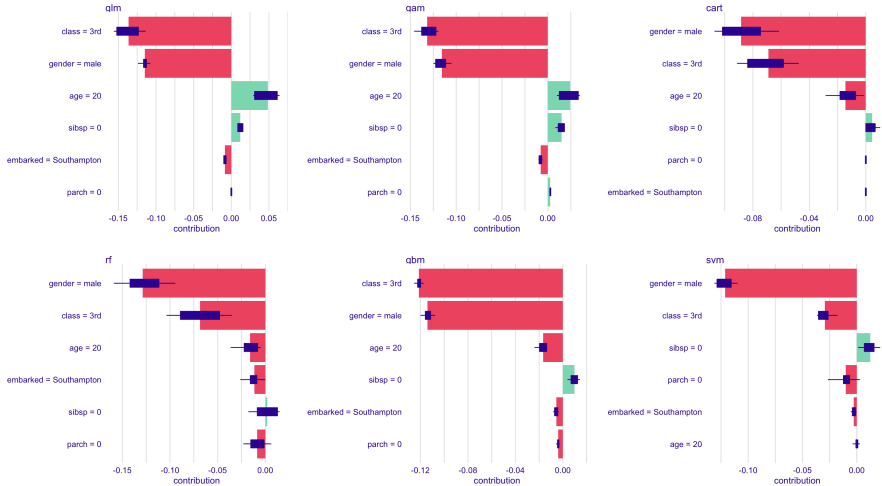


Shapley Additive Explanations (SHAP) for Average Attributions IX

$\boldsymbol{x}^*$ = Leonardo, $m \in \{\text{glm, gam, cart, rf, gbm, svm}\}$

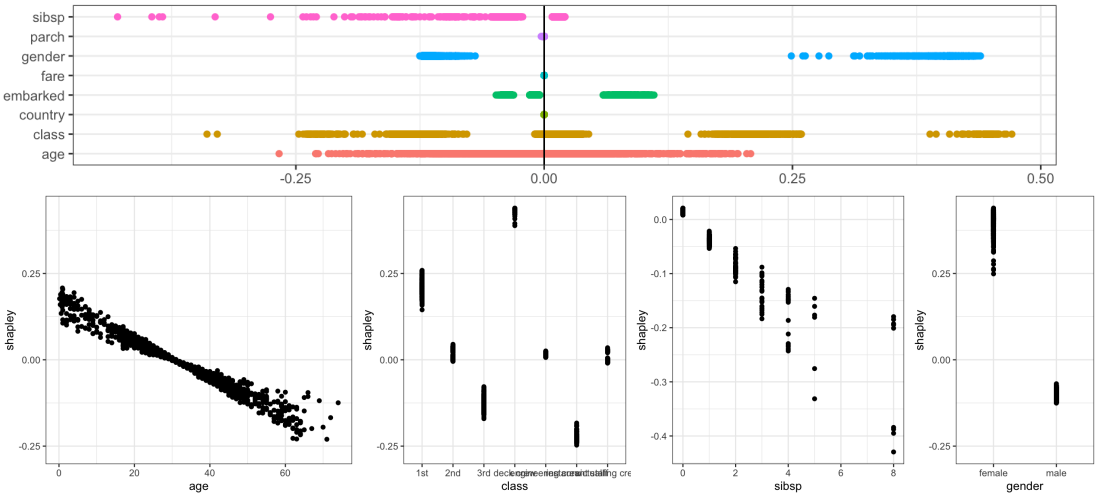
Shapley Additive Explanations (SHAP) for Average Attributions X

$\boldsymbol{x}^*$ = Leonardo, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$ (with confidence boxplot)



Shapley Additive Explanations (SHAP) for Average Attributions XI

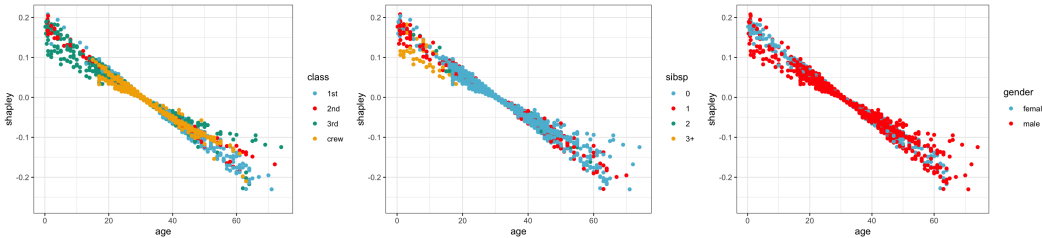
Sur l'ensemble de la base d'apprentissage,



Shapley Additive Explanations (SHAP) for Average Attributions XII

Ces graphiques $\{(x_{i,j}, \gamma_j^{shap}(\boldsymbol{x}_i))\}$ sont dit "Shapley Dependence Plots".

On peut aussi rajouter une couleur pour $x_{i,k}$ et on parle de Shapley interaction plot



Shapley Additive Explanations (SHAP) for Average Attributions XIII

On simule un modèle linéaire ($n = 1000$)

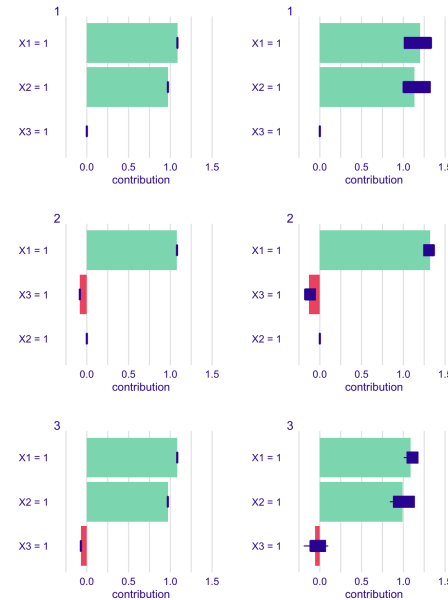
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.0$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XIV

On simule un modèle linéaire ($n = 1000$)

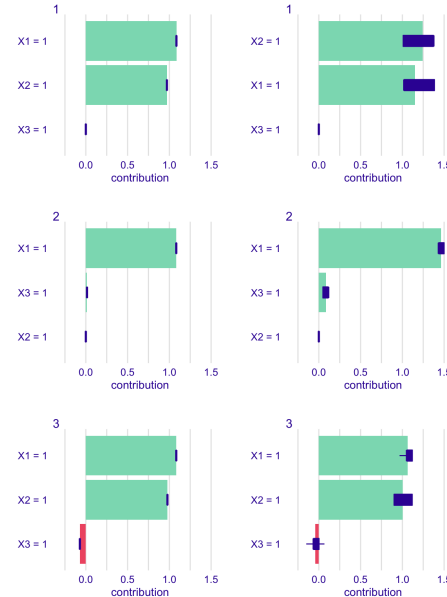
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.1$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XV

On simule un modèle linéaire ($n = 1000$)

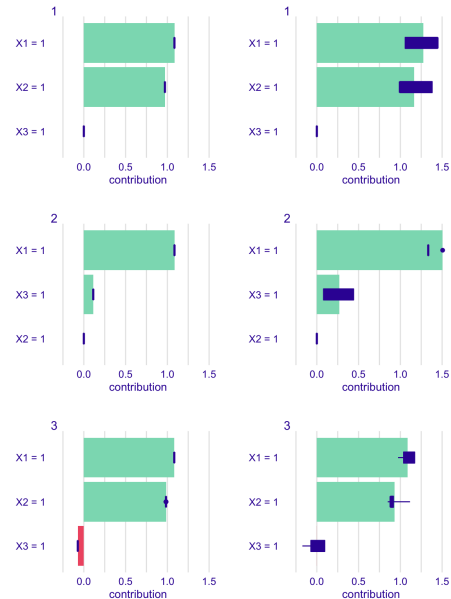
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.2$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XVI

On simule un modèle linéaire ($n = 1000$)

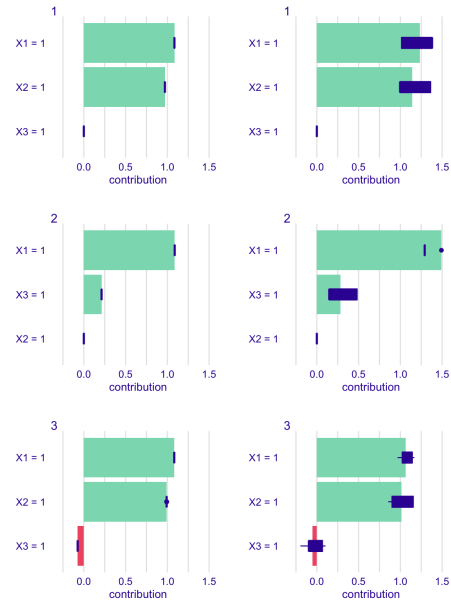
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.3$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XVII

On simule un modèle linéaire ($n = 1000$)

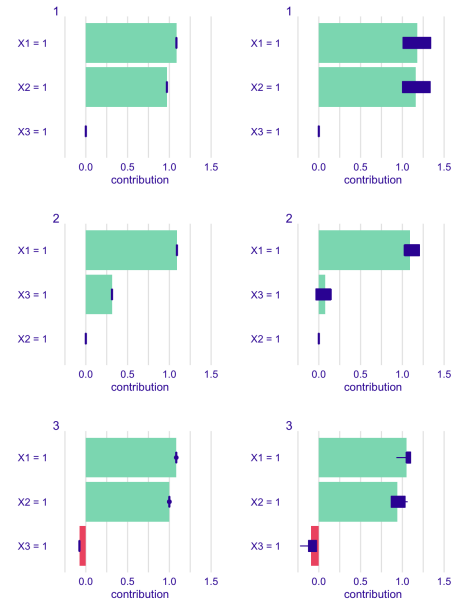
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.4$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XVIII

On simule un modèle linéaire ($n = 1000$)

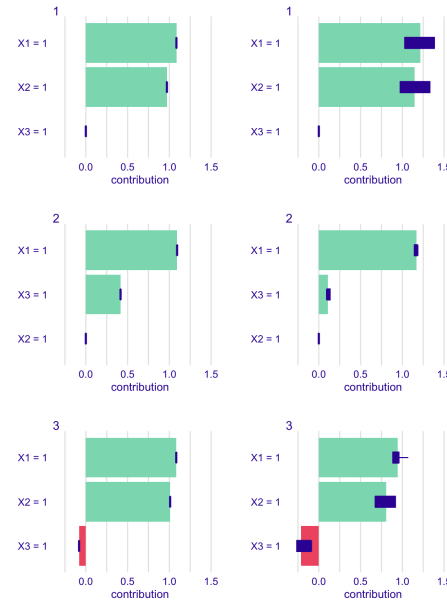
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.5$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XIX

On simule un modèle linéaire ($n = 1000$)

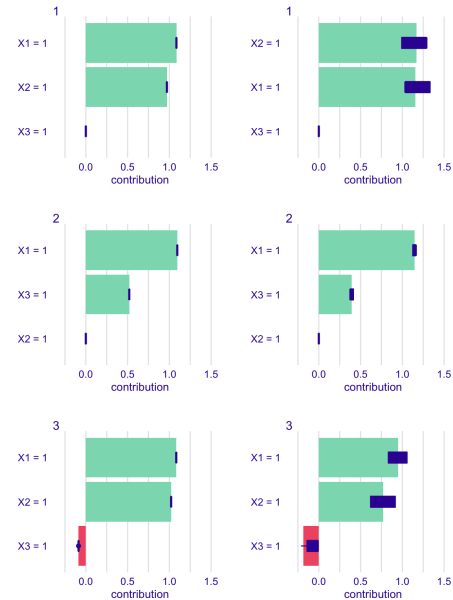
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.6$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XX

On simule un modèle linéaire ($n = 1000$)

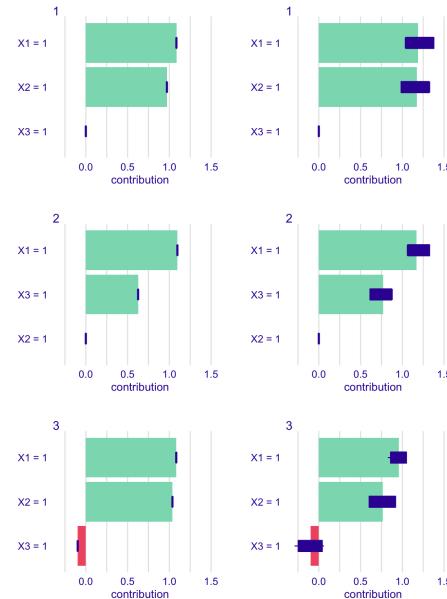
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.7$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XXI

On simule un modèle linéaire ($n = 1000$)

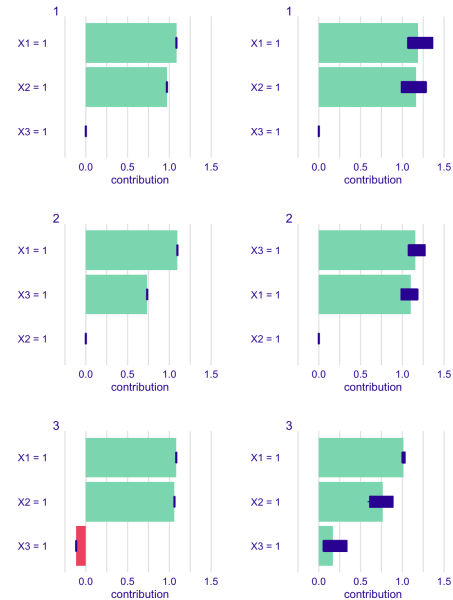
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.8$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XXII

On simule un modèle linéaire ($n = 1000$)

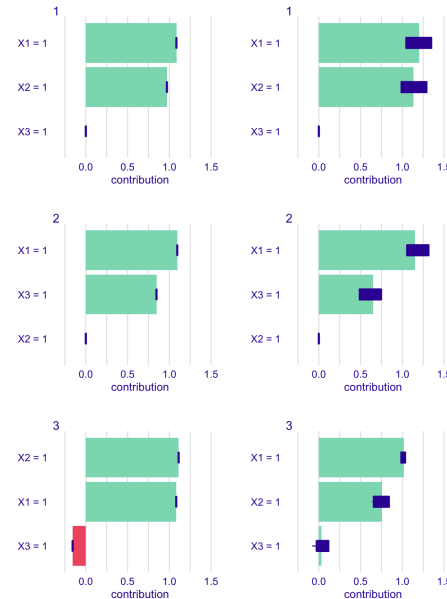
$$y = 1 + x_1 + x_2 + \varepsilon$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) \right)$$

avec $r = 0.9$, on estime trois modèles

$$\begin{cases} (1) & y \sim x_1 + x_2 \\ (2) & y \sim x_1 + x_3 \\ (3) & y \sim x_1 + x_2 + x_3 \end{cases}$$

modèle linéaire et forêt aléatoire, $\boldsymbol{x}^* = (1\ 1\ 1)^\top$



Shapley Additive Explanations (SHAP) for Average Attributions XXIII

Au lieu d'une vision locale (en $\boldsymbol{x}^*$) on peut avoir une vision globale

Štrumbelj and Kononenko (2014) puis Lundberg and Lee (2017) ont proposé d'utiliser cette décomposition pour expliquer la contribution de chaque variable.

γ_j^{shap}

La contribution de la $j$ème variable,

$$\gamma_j^{shap} = \frac{1}{n} \sum_{i=1}^{n} \gamma_j^{shap}(\boldsymbol{x}_i)$$

Shapley Additive Explanations (SHAP) for Average Attributions XXIV

$\gamma_{i,j}^{shap}(\boldsymbol{x}^*)$

La contribution d'interaction de Shapley, entre les variables i et j, en $\boldsymbol{x}^*$, est

$$\gamma_{i,j}(\boldsymbol{x}^*) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{i,j\}} \frac{|S|!\,(p-|S|-2)!}{2\ p!} \Delta_{i,j|S}(\boldsymbol{x}^*)$$

où

$$\begin{aligned} \Delta_{i,j|S}(\boldsymbol{x}^*) &= \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{i,j\}}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_{S\cup\{i,j\}}\right] - \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{j\}}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_{S\cup\{j\}}\right] \\ &- \mathbb{E}_{\boldsymbol{X}^{\perp}_{S\cup\{i\}}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_{S\cup\{i\}}\right] + \mathbb{E}_{\boldsymbol{X}^{\perp}_{S}}\left[m(\boldsymbol{X})\middle|\boldsymbol{x}^*_{S}\right] \end{aligned}$$

Shapley Additive Explanations (SHAP) for Average Attributions XXV

- ➕ approche “agnostique”, indépendante du modèle
  garanties théoriques solides, local et global, si les contributions sont additives
- ➖ peut être complexe à calculer, numériquement (donc lent)
  algorithmes rapides sur quelques modèles spécifiques (arbres)
  problème des “unrealistic input”, Kumar et al. (2020)

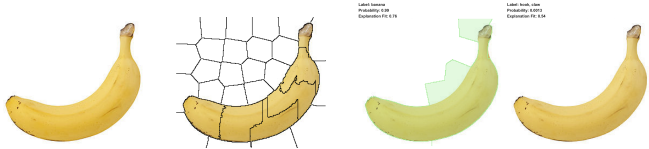
LIME I

Étant donné un modèle m sur $\mathcal{X}$, on cherche (Ribeiro et al. (2016))

$$\underset{m_e \in \mathcal{E}}{\operatorname{argmin}} \left\{ \ell_{\boldsymbol{x}^*}(m, m_e) \right\} + \mathcal{P}(m_e)$$

où

- $\mathcal{E}$ est un sous-ensemble de fonctions $\mathcal{X} \to \mathbb{R}$ "explicables"
  ou $\tilde{\mathcal{X}} \to \mathbb{R}$, où $\tilde{\mathcal{X}}$ est un espace plus simple (*space for interpretable representation*)
- $\ell_{\boldsymbol{x}^*}$ est une distance entre deux modèles au voisinage de $\boldsymbol{x}^*$
- $\mathcal{P}$ est une fonction de pénalité, croissante en la complexité du modèle



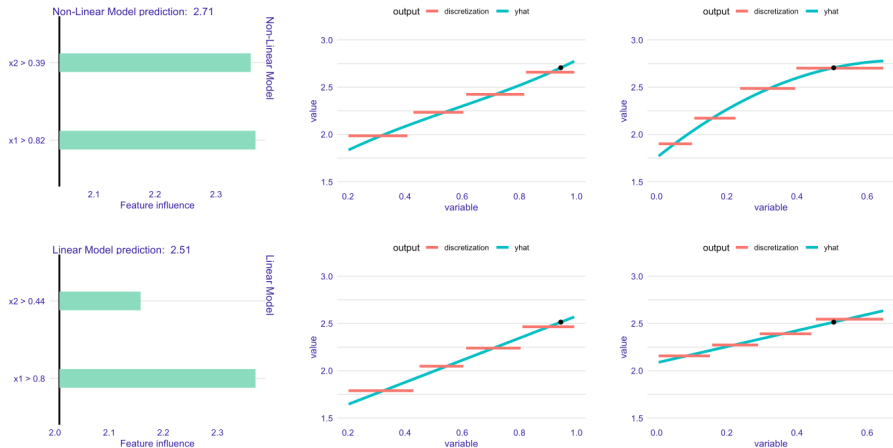
LIME II

Classiquement, on peut utiliser un modèle linéaire (avec LASSO) ou des arbres.
On parle d'explicabilité par substitution (surrogate models) locale

- ⊕ approche "agnostique", indépendante du modèle
  la représentation est (souvent) interprétable (transformation de $\mathcal{X}$ en $\tilde{\mathcal{X}}$)
  largement adoptée pour l'analyse des textes et des images (grâce à la représentation interprétable des données)
- ⊖ difficile de définir un voisinage robuste, Alvarez-Melis and Jaakkola (2018)

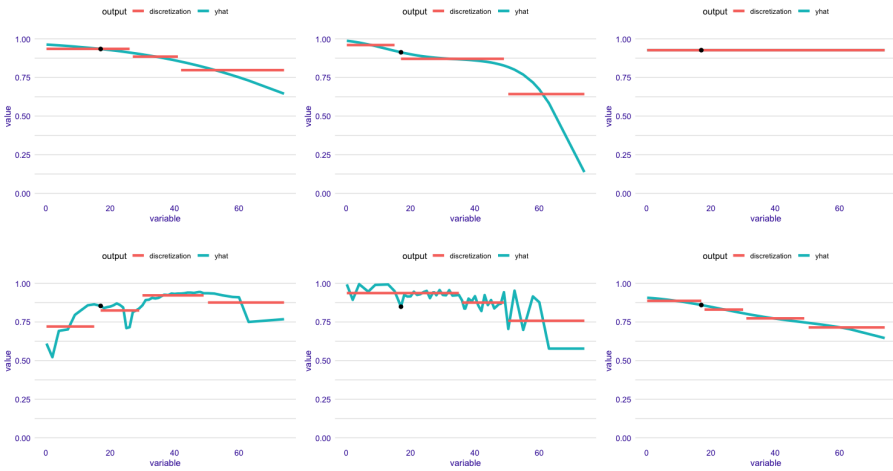
LIME III

Sur les données $(x_{i,1}, x_{i,2})$



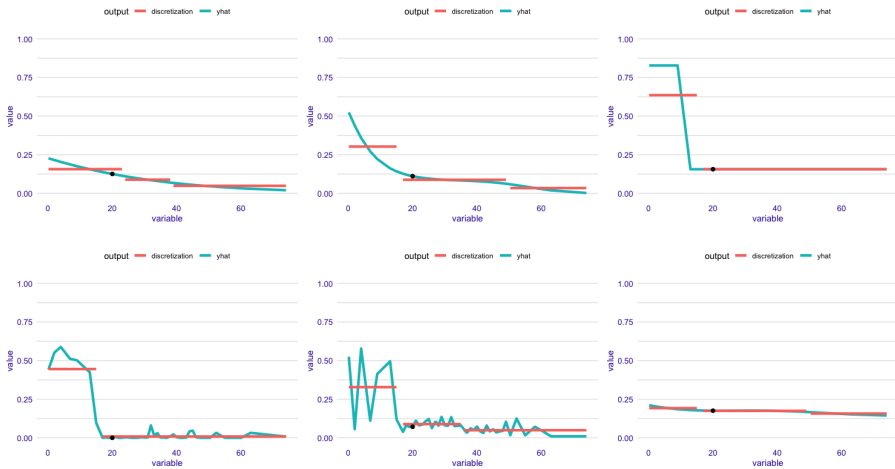
LIME IV

$\boldsymbol{x}^* =$ Kate, $j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



LIME V

$\boldsymbol{x}^* =$ Leonardo, $j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Local-diagnostic plots I

A partir de $\boldsymbol{x}^*$, on cherche les plus proches voisins
On compare alors

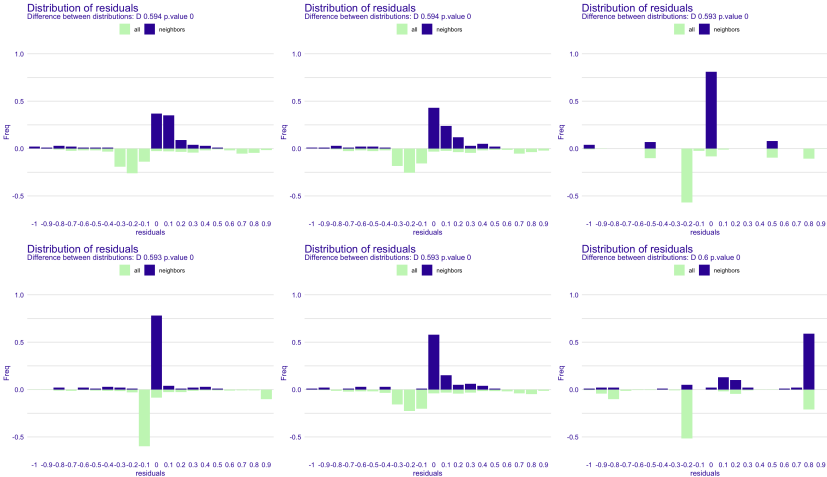
- la distribution globale (■) des résidus
- la distribution des résidus des voisins de $\boldsymbol{x}^*$ (■)

pour la mesure de similarité de Gower (Gower (1971)), i.e.

$$d_G(\boldsymbol{x}_i, \boldsymbol{x}^*) = \frac{1}{p}\sum_{j=1}^{p} d_j(x_{i,j}, x_j^*), \text{ où } d_j(x_{i,j}, x_j^*) = \begin{cases} \dfrac{|x_{i,j} - x_j^*|}{\max\{x_j\} - \min\{x_j\}}, & \text{si } j \text{ continue} \\ \mathbf{1}(x_{i,j} \neq x_j^*), & \text{si } j \text{ qualitative} \end{cases}$$

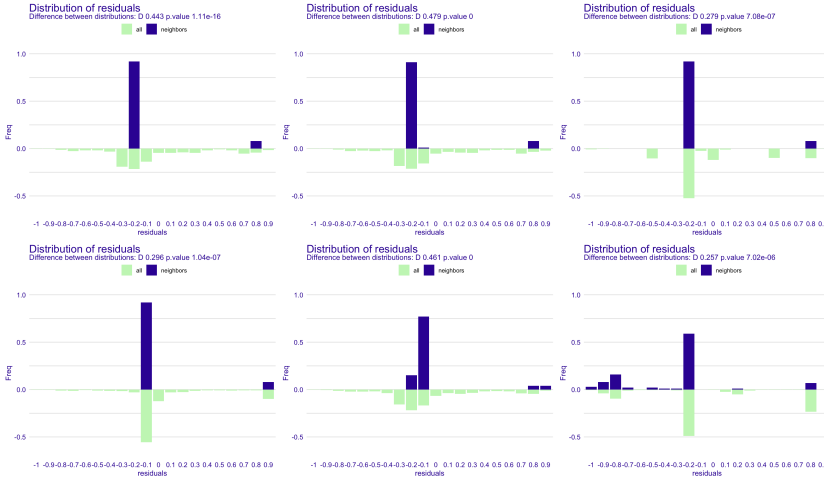
Local-diagnostic plots II

$\boldsymbol{x}^* =$ Kate, $m \in \{\text{glm, gam, cart, rf, gbm, svm}\}$



Local-diagnostic plots III

$\boldsymbol{x}^* =$ Leonardo, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Local-stability plots I

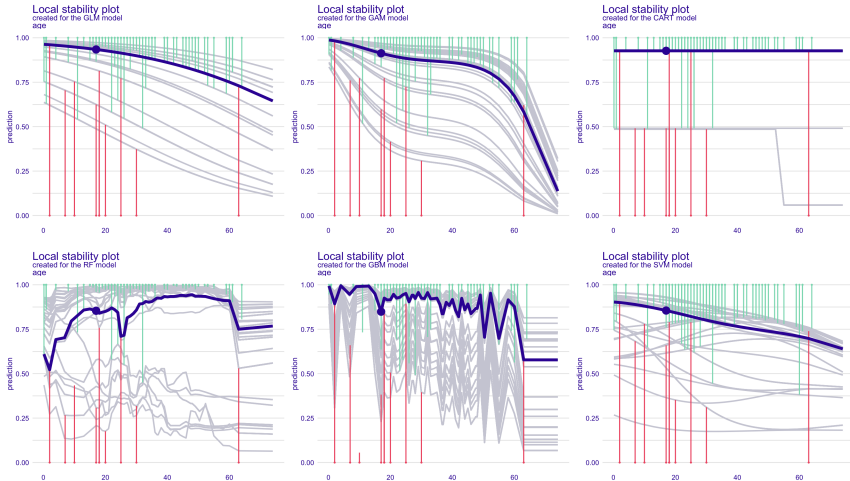
- ➕ outils classique en régression linéaire
- ➖ difficile à exploiter

- courbe — profile ceteris paribus de $\boldsymbol{x}^*$
- courbe — profiles ceteris paribus des voisins de $\boldsymbol{x}^*$
- courbes — et —, résidus positifs et négatifs des voisins de $\boldsymbol{x}^*$

pour la mesure de similarité de Gower (Gower (1971))

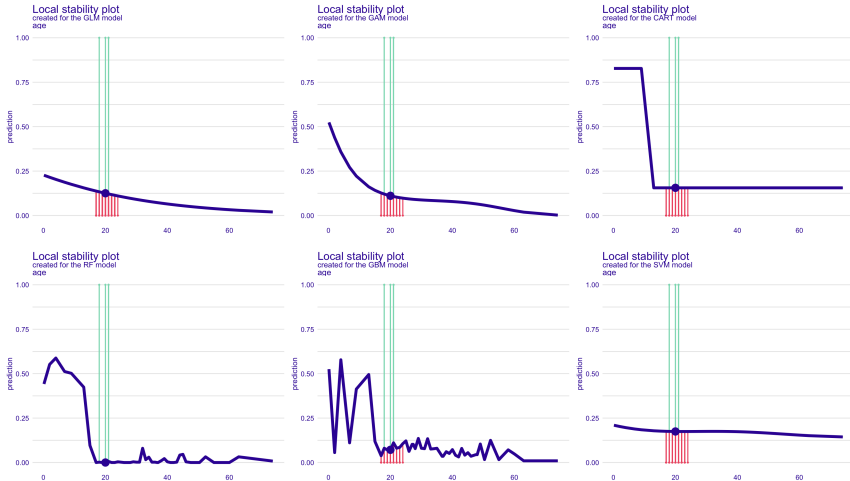
Local-stability plots II

$\boldsymbol{x}^* =$ Kate, $j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



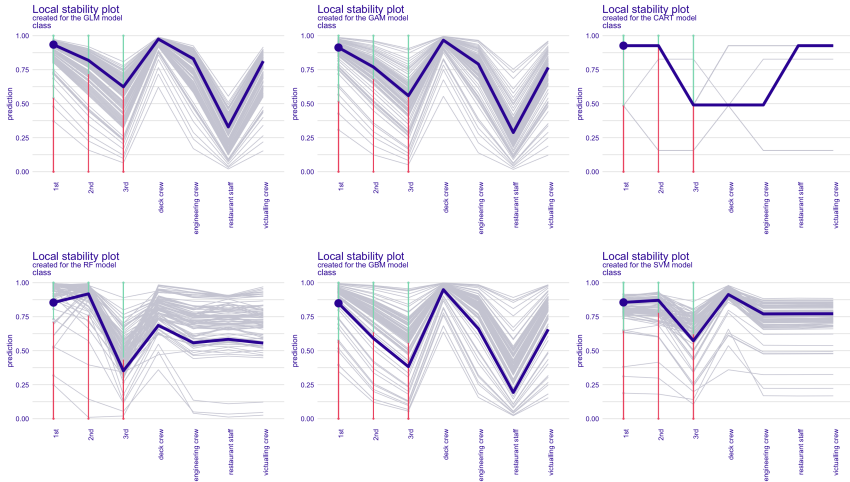
Local-stability plots III

$\boldsymbol{x}^* =$ Leonardo, $j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



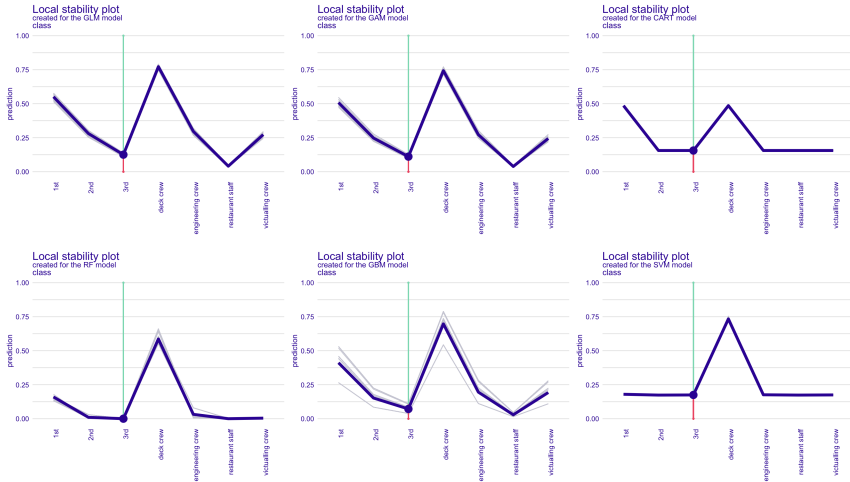
Local-stability plots IV

$\boldsymbol{x}^* =$ Kate, $j =$ classe, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Local-stability plots V

$\boldsymbol{x}^* =$ Leonardo, $j =$ classe, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



Partial Dependence Plot I

Friedman (2001), dans le contexte du gradient boosting

Partial Dependence Plot $p_j(x_j^*)$ et $\widehat{p}_j(x_j^*)$

Le Partial Dependence Plot de la variable j est la fonction $\mathcal{X}_j \to \mathbb{R}$

$$p_j(x_j^*) = \mathbb{E}_{\boldsymbol{X}_j^\perp}\left[m(\boldsymbol{X})|x_j^*\right]$$

et sa version empirique est

$$\widehat{p}_j(x_j^*) = \frac{1}{n}\sum_{i=1}^{n} m(x_j^*, \boldsymbol{x}_{i,-j}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{m_{\boldsymbol{x}_i,j}(x_j^*)}_{\text{ceteris paribus}}$$

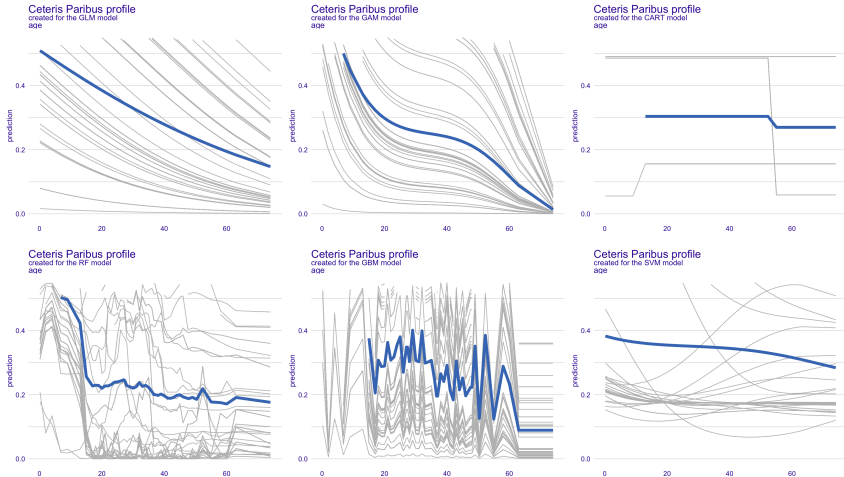
Partial Dependence Plot II

- ⊕ intuitif, facile à comprendre
- ⊖ hérite de l'approche ceteris paribus

  e.g. suppose l'indépendance entre les variables, pas d'effets croisés

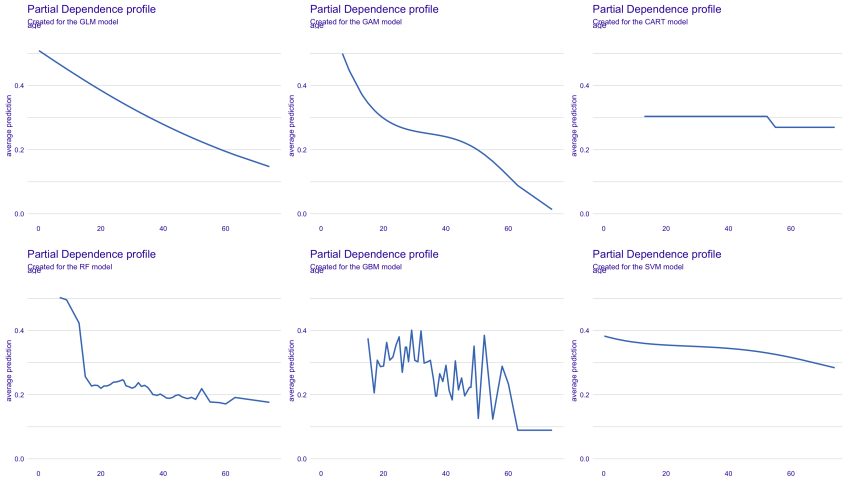
Partial Dependence Plot III

$j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



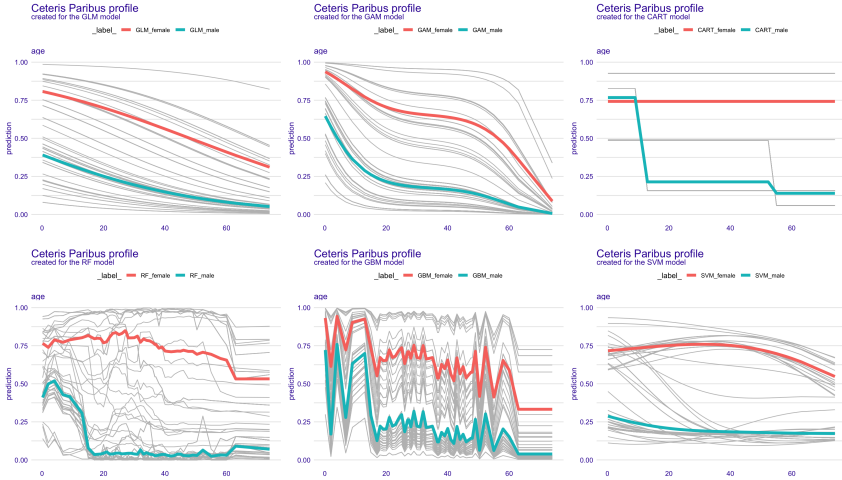
Partial Dependence Plot IV

$j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



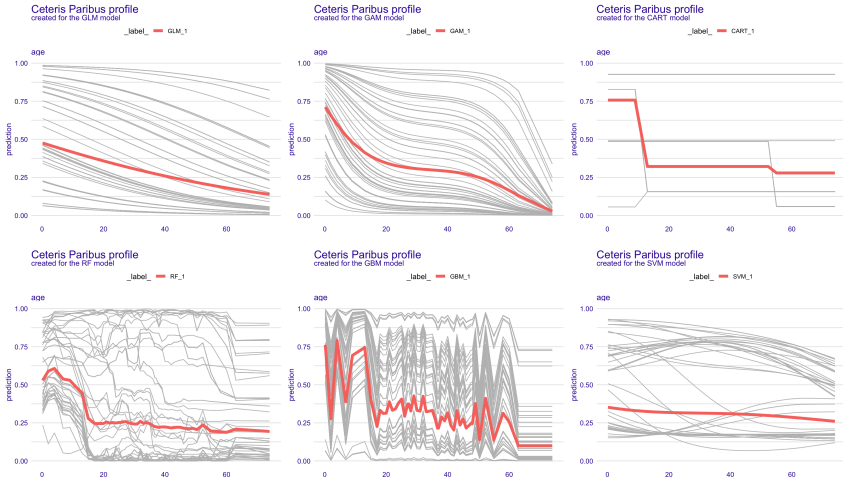
Partial Dependence Plot V

On peut regrouper les courbes par classes, $j =$ âge



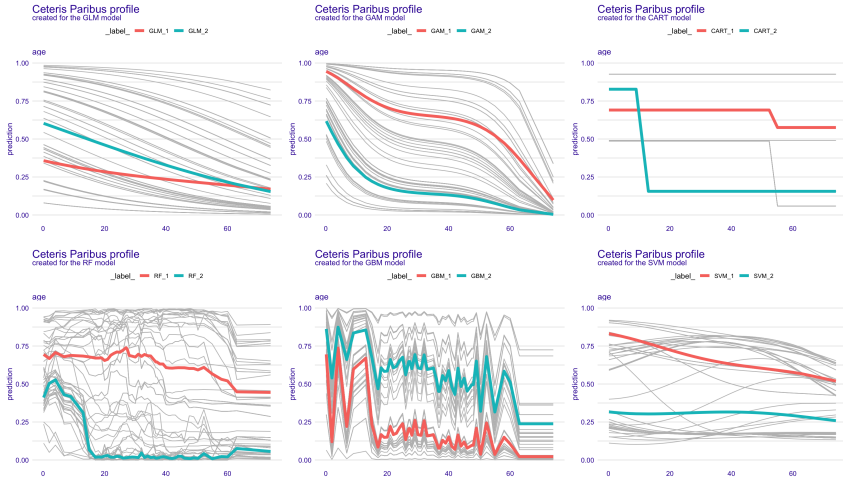
Partial Dependence Plot VI

on peut aussi construire des classes non-supervisées ($k = 1$), $j =$ âge



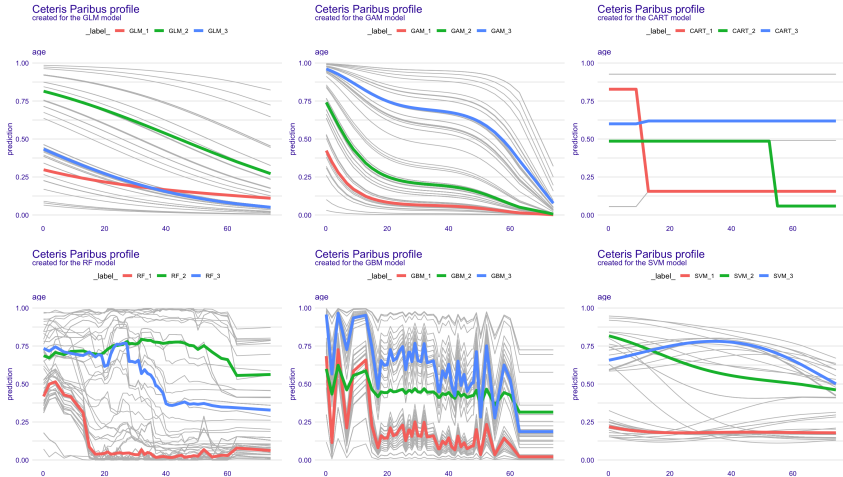
Partial Dependence Plot VII

on peut aussi construire des classes non-supervisées ($k = 2$), $j =$ âge



Partial Dependence Plot VIII

on peut aussi construire des classes non-supervisées ($k = 3$), $j =$ âge



Local Dependence I

$p_j(x_j^*) = \mathbb{E}_{X_j^\perp}\left[m(\boldsymbol{X})|x_j^*\right]$ ne tenait pas compte du faire que $(\boldsymbol{X}_{-j}|X_j) \overset{\mathcal{L}}{\neq} \boldsymbol{X}_{-j}$

Apley and Zhu (2020) propose

Local Dependence Plot $\ell_j(x_j^*)$ et $\widehat{\ell}_j(x_j^*)$

$$\ell_j(x_j^*) = \mathbb{E}_{X_j}\left[m(\boldsymbol{X})|x_j^*\right]$$

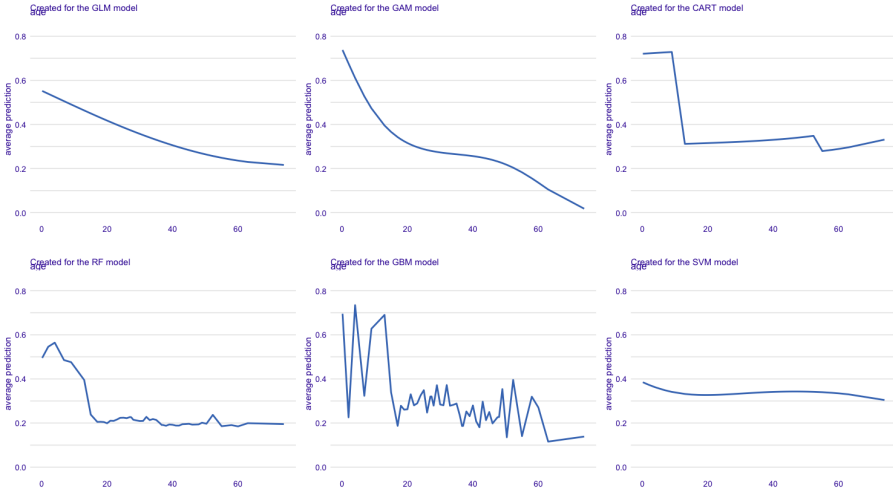
$$\widehat{\ell}_j(x_j^*) = \frac{1}{\text{card}(V(x_j^*))} \sum_{i \in V(x_j^*)} m(x_j^*, \boldsymbol{x}_{i,-j}) \text{ où } V(x_j^*) = \left\{i : d(x_{i,j}, x_j^*) \leq \epsilon\right\}$$

$$\widehat{\ell}_j(x_j^*) = \frac{1}{\sum_i \omega_i(x_j^*)} \sum_{i=1}^{n} \omega_i(x_j^*) m(x_j^*, \boldsymbol{x}_{i,-j}) \text{ où } \omega_i(x_j^*) = K_h(x_j^* - x_{i,j})$$

pour une version lissée par noyau.

Local Dependence II

$j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



ALE (Accumulated Local Effects) I

Apley and Zhu (2020) propose

Accumlated Local $a_j(x_j^*)$

$$a_j(x_j^*) = \int_{-\infty}^{x_j^*} \mathbb{E}_{X_j}\left[\frac{\partial m(x_j, \boldsymbol{X}_{-j})}{\partial x_j}\Big| x_j\right] dx_j$$

$\frac{\partial m(x_j, \boldsymbol{X}_{-j})}{\partial x_j}$ décrit le changement local du modèle m du à x_j (ceteris paribus),

$$m(x_j + dx_j, \boldsymbol{x}_{-j}) - m(x_j, \boldsymbol{x}_{-j}) \approx \frac{\partial m(x_j, \boldsymbol{X}_{-j})}{\partial x_j} dx_j$$

et on regarde la valeur moyenne locale.

ALE (Accumulated Local Effects) II

Apley and Zhu (2020) propose

Accumlated Local $\widehat{a}_j(x_j^*)$

$$\widehat{a}_j(x_j^*) = \alpha + \sum_{u=1}^{k_j^*} \frac{1}{n_u} \sum_{u:x_{i,j} \in (a_{u-1}, a_u]} [m(a_k, \boldsymbol{x}_{i,-j}) - m(a_{k-1}, \boldsymbol{x}_{i,-j})]$$

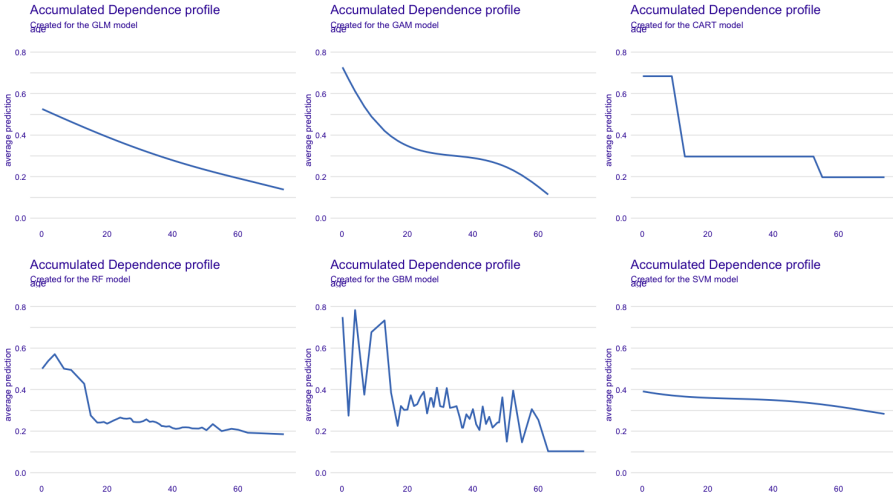
(où α est une constante de normalisation car $\mathbb{E}[a_j(X_j)] = 0$)

Apley and Zhu (2020) propose une discrétisation de X_j (partition $\{(a_{k-1}, a_k]\}$), ou une version lissée par noyau

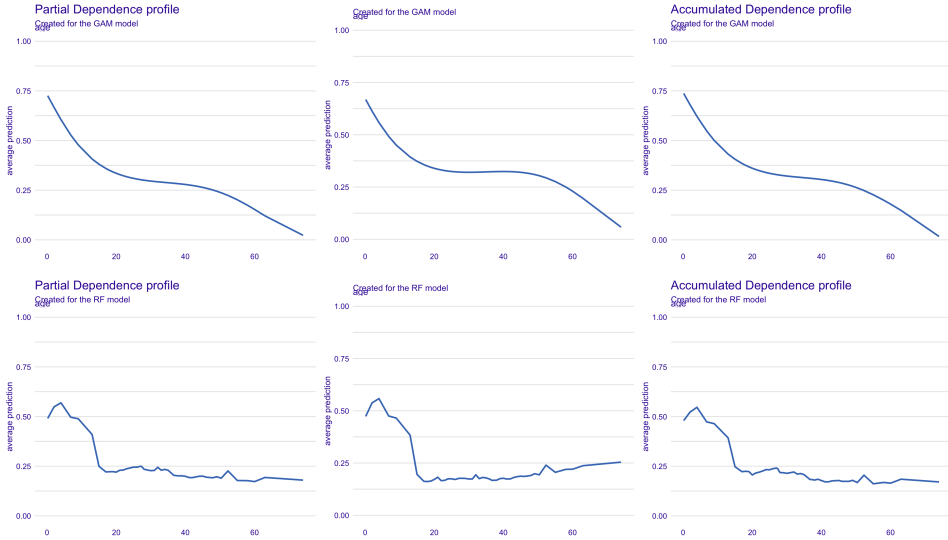
- ⊕ peut être utilisé sur des variables corrélées, Apley and Zhu (2020)
- ⊖ peut être difficulté à calculer

ALE (Accumulated Local Effects) III

$j =$ âge, $m \in \{$glm, gam, cart, rf, gbm, svm$\}$



PDP, LD et ALE, une comparaison



Parmi les choses qui existent par ailleurs...

Les informaticiens ont remis au goût du jour beaucoup de choses anciennes, et d'autres techniques existaient aussi avant.
Par exemple, en statistique, la fonction d'influence, introduites par Hampel (1968).
Soit $\boldsymbol{X}$ de loi F, et considérons une fonction de F, $T(F)$ (par exemple) $T(F) = \mathbb{E}(\boldsymbol{X})$ avec $\boldsymbol{X} \sim F$. *"La fonction d'influence de la statistique T quantifie la sensibilité lorsqu'une proportion infinitésimale des données est corrompue"*.
Étant donnée F, notons pour tout $\varepsilon \in [0, 1]$ $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\delta_{\boldsymbol{x}}$, et notons

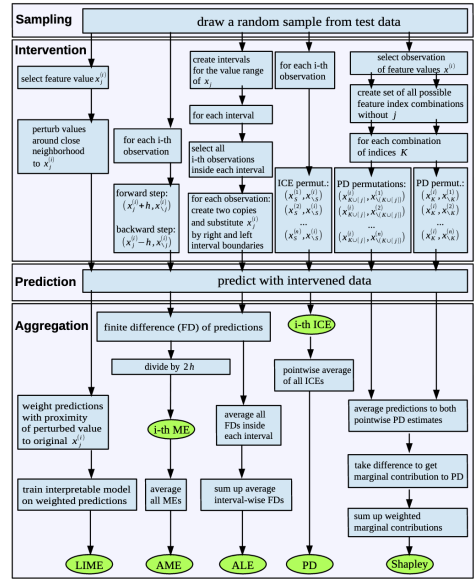
$$\mathrm{IF}(T, F, \boldsymbol{x}) = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}$$

Cette fonction a été utilisée par Koh and Liang (2017) pour interpréter des algorithmes boîtes noires.

Résumé

- proche de questions classiques en statistique (robustesse, influence, etc)
- ceteris paribus $\mathbb{E}_{X_1^{\perp}}[Y|x_1^*]$
- mutatis mutandis $\mathbb{E}_{X_1}[Y|x_1^*]$
- boîtes blanches trompeuses
- besoin de mieux comprendre les métriques
- boîtes noires utiles pour construire des features interprétables
- définitions dépendant de la finalité

(schéma Scholbeck et al. (2019))



References I

Aasman, J. (2019). Creating explainable ai with rules. *Forbes*, May 17.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv*, 1806.08049.

Apley, D. W. and Zhu, J. (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):1059–1086.

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

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., and Parekh, J. (2020). Flexible and context-specific ai explainability: A multidisciplinary approach. *arXiv preprint arXiv:2003.07703*.

Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, page 39.

Biecek, P. and Burzykowski, T. (2021). *Explanatory model analysis: Explore, explain and examine predictive models*. Chapman and Hall/CRC.

Bostock, M. and Carter, S. (2012). 512 paths to the white house. *New York Times*, November 2nd.

References II

Conan-Doyle, A. (1887). *A study in scarlet*. Ward Lock & Co's.

Davis, B., Anderson, R., and Walls, J. (2015). *Rashomon effects: Kurosawa, Rashomon and their legacies*. Routledge.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Friedman, J. (2001). Full-Text. *Annals of Statistics*, 29 (5):1189–1232.

Gauld, T. (2021). Two ways to explain scientific research. *New Scientist*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

References III

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Goyal, Y., Feder, A., Shalit, U., and Kim, B. (2019). Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Halpern, J. Y. and Pearl, J. (2020a). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*.

Halpern, J. Y. and Pearl, J. (2020b). Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. University of California, Berkeley.

Helton, J. C. and Davis, F. (2002). Illustration of sampling-based methods for uncertainty and sensitivity analysis. *Risk analysis*, 22(3):591–622.

Kahneman, D. and Tversky, A. (1981). The simulation heuristic. Technical report, Stanford Univ CA Dept of Psychology.

References IV

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR.

Lewis, D. (1986). Causal explanation. In *Philosophical Papers Vol. II*. Oxford University Press.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

Lombrozo, T. (2012). Explanation and abductive inference.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Mangal, R., Nori, A. V., and Orso, A. (2019). Robustness of neural networks: A probabilistic and practical approach. *arXiv*, 1902.05983.

Marx, C., Calmon, F., and Ustun, B. (2020). Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR.

References V

May, R. M. (2004). Simple mathematical models with very complicated dynamics. In *The Theory of Chaotic Attractors*, pages 85–93. Springer.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., and Przybocki, M. A. (2020). Four principles of explainable artificial intelligence. *Gaithersburg, Maryland*.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Rahimi, A. (2017). Machine learning has become alchemy. In *Thirsty-first Conference on Neural Information Processing Systems*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *arXiv*, 1602.04938.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141.

References VI

Robnik-Šikonja, M. and Bohanec, M. (2018). Perturbation-based explanations of prediction models. In *Human and machine learning*, pages 159–175. Springer.

Robnik-Šikonja, M. and Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *Machine learning: Proceedings of the fourteenth international conference (ICML97)*, volume 5, pages 296–304.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1):23–69.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., and Casalicchio, G. (2019). Sampling, intervention, prediction, aggregation: a generalized framework for model-agnostic interpretations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 205–216. Springer.

References VII

Semenova, L., Rudin, C., and Parr, R. (2022). On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Swartout, W., Paris, C., and Moore, J. (1991). Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, 6(3):58–64.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Thom, R. (1991). *Prédire n'est pas expliquer*. Flammarion.

Ustun, B. and Rudin, C. (2014). Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*.

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616. Citeseer.

References VIII

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.