



Introduction à l'Apprentissage Statistique

Interprétabilité

M2 Actuariat – ISFA – 2021/2022

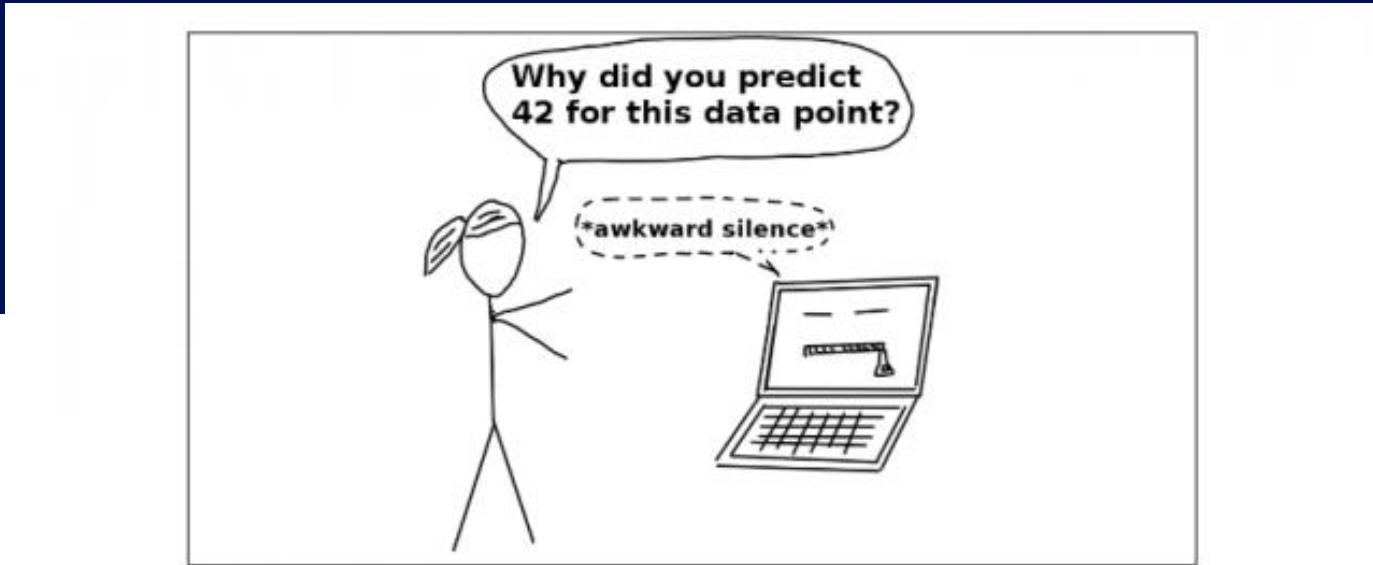
Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming



Interprétabilité



Pourquoi ce besoin d'interprétation ?

Parfois la prédiction seule n'est pas suffisante

- Problème incomplet dans la modélisation

Sensibilisation du public

- Augmenter la confiance dans les algorithmes

Détection de biais

- Transitivité des biais de la base d'entraînement au modèle

Robustesse du modèle

- Vérifier qu'il n'y a pas d'effets de seuil

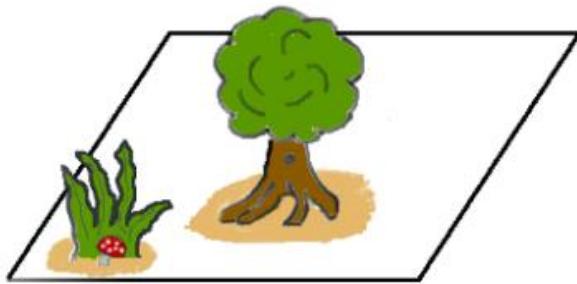
Interprétabilité

Que veut-on dire par « interpréter un modèle de machine learning » ?

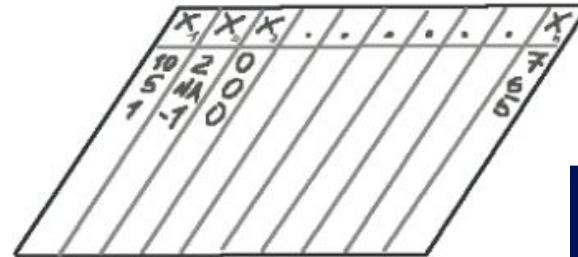
Pourquoi en avons-nous besoin ? Croire au modèle, le visualiser, trouver les relations causales ?

Pour les modèles black box, les notions de transparence et d'interprétabilité sont très similaires. Différentes stratégies sont possibles

- Expliquer le modèle (sa structure, son estimation)
- Expliquer la réponse (niveau global ou niveau local)
- Inspecter le modèle en interne
- Proposer une solution transparente



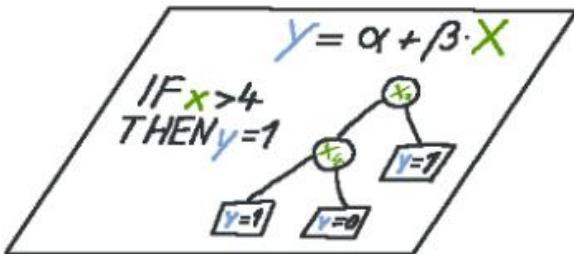
Collecte des données du monde



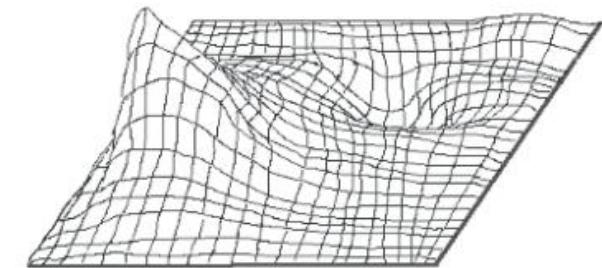
Apprentissage par modèles complexes



Transmettre l'information à un humain
(avec des images c'est mieux ☺)



Extraction d'un modèle simple



Littérature

Question finalement assez récente

- Lakkaraju *et al.*, 2019, Faithful and Customizable Explanations of Black Box Models
- Molnar, 2019, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable
- Guidotti *et al.*, 2018, A Survey of Methods for Explaining Black Box Models
- Gilpin *et al.*, 2018, Explaining Explanations: An Overview of Interpretability of Machine Learning
- Mailliart, 2021, Quelques Méthodes d'Explicabilité pour les Modèles d'Apprentissage Statistique en Actuariat

Model linéaire

Si on considère le modèle de régression suivant

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

avec l'économétrie classique, si on fixe x_2 on a

$$\beta_1 = \frac{\partial y}{\partial x_1}$$

ou encore, en notant m l'estimateur

$$\beta_1 = \frac{\partial m(x)}{\partial x_1}$$

CART

La visualisation est assez évidente par la mise en forme et la lecture de l'arbre (merci *rpart.plot*), mais l'ordre des nœuds peut être trompeur

Une autre idée est d'estimer l'importance de chaque covariable pour la prédiction de la variable d'intérêt

- Chaque nœud diminue l'hétérogénéité
- Chaque nœud est un split sur une unique variable
- On peut en déduire la diminution totale induite par une covariable en parcourant l'arbre en entier
- Comparaison des **Feature Importances**



Model Agnostic Methods

Permutation Feature Importance

Estimer le changement en terme de pouvoir de prédiction de l'estimateur si une covariable est remélangée.

1. Estimer le modèle \hat{f} et l'erreur de prédiction $e_0 = L(y, \hat{f}(x))$
2. Pour chaque covariable $j \in \{1, \dots, p\}$
 1. On génère, par permutation aléatoire, des données altérées \tilde{x}_j en mélangeant les valeurs de la covariable j et en gardant les autres inchangées
 2. On estime l'erreur de prédiction $e_j = L(y, \hat{f}(\tilde{x}_j))$
 3. On définit la feature importance par $FI_j = e_j - e_0$

Remarque : pour plus de robustesse, on peut **itérer l'étape 2.1 et 2.2** et définir la feature importance comme la moyenne des feature importances de chaque mélange.

Partial Dependence Plot

Monter graphiquement l'effet marginal d'une ou deux covariable(s) sur la variable réponse.

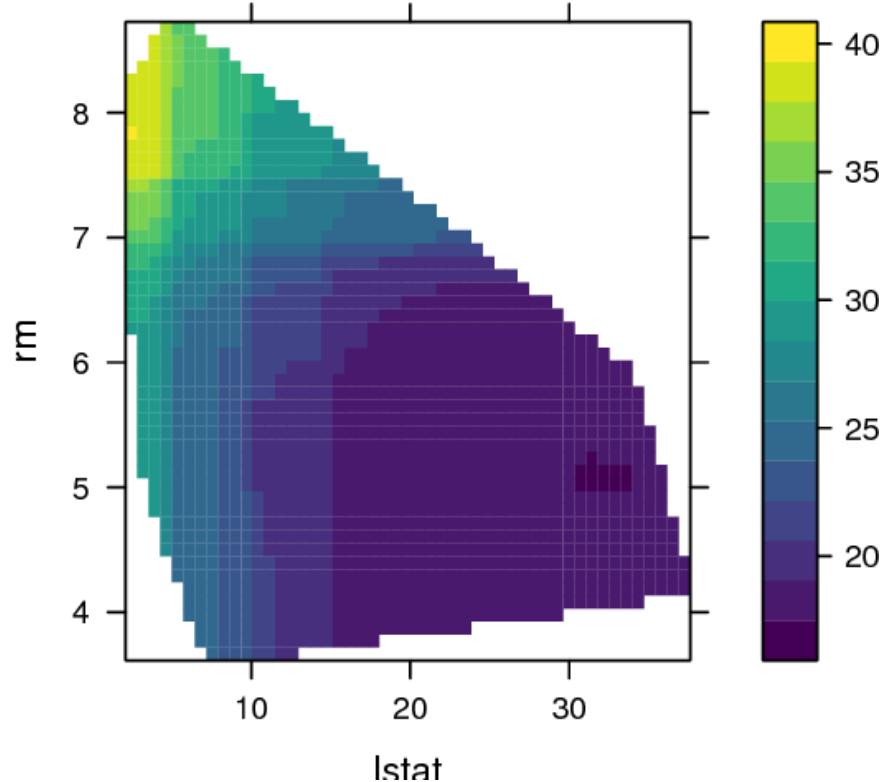
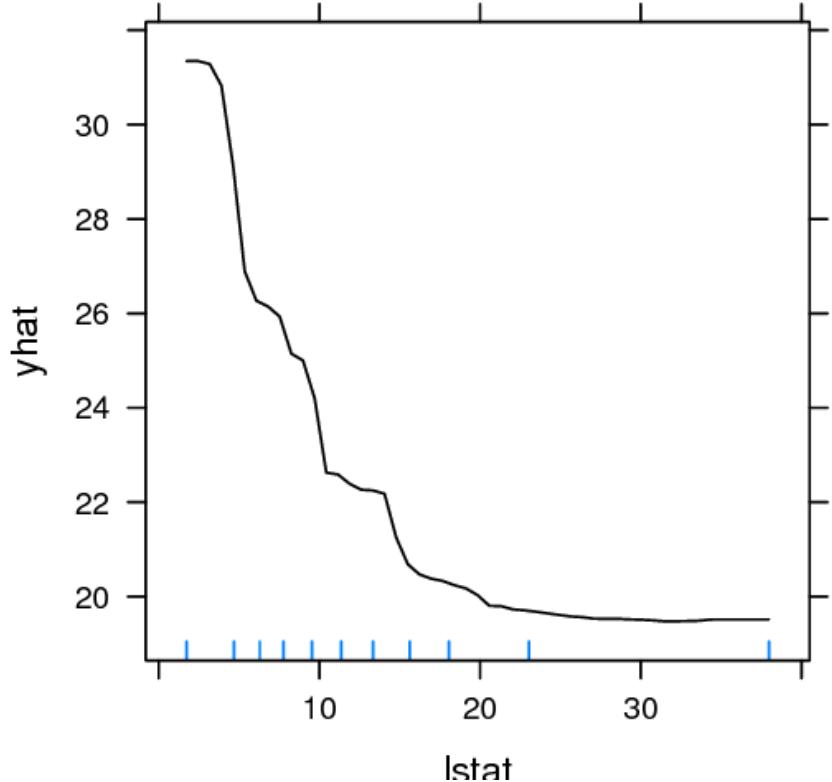
Fonction de dépendance partielle (régression) d'une covariable j par rapport à l'ensemble des autres covariables $C = \{1, \dots, p\} \setminus \{j\}$

$$\hat{f}_j(x_j) = \mathbb{E}_{X_C}[\hat{f}(x_j, X_C)] = \int \hat{f}(x_j, X_C) d\mathbb{P}(X_C)$$

Pratiquement, la fonction peut facilement être estimée à partir des valeurs $x_c^{(i)}$ prises par les covariables sur la base d'entraînement

$$\hat{f}_j(x_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}\left(x_j, x_c^{(i)}\right)$$

Partial Dependence Plot



PDP – Feature Importance

Une courbe PDP plate indique que la covariable n'est pas importante

Pour des variables continues, on estime l'écart type standard par rapport à la courbe PDP moyenne

$$I(x_j) = \sqrt{\frac{1}{K-1} \sum_{k=1}^K \left(\hat{f}_j(x_j^{(k)}) - \frac{1}{K} \sum_{l=1}^K \hat{f}_j(x_j^{(l)}) \right)^2}$$

Avec $\{x_j^{(k)}\}_{k=1,\dots,K}$ les K valeurs uniques prises par la covariables j

Extension aux variables catégorielles

$$I(x_j) = \left(\max_k \{\hat{f}_j(x_j^{(k)})\} - \min_k \{\hat{f}_j(x_j^{(k)})\} \right) / 4$$



SHAP

Théorie des jeux

Soit un jeu de coalition de N joueurs et une fonction caractéristique ν qui définit le gain collectif d'un sous-échantillon de joueurs $S \subseteq \{1, \dots, N\}$.

Question : quelle est la contribution du joueur i à la coalition ?

Contribution marginale du joueur i dans la coalition S

$$\Delta_\nu(i, S) = \nu(S \cup i) - \nu(S)$$

La Shapley value du joueur i est la moyenne de sa contribution marginale pour toutes les coalitions de joueurs possibles

$$\phi_\nu(i) = \sum_{S \subseteq \{1, \dots, N\} \setminus \{i\}} \frac{|S|! (N - |S| - 1)!}{N!} \times \Delta_\nu(i, S)$$

SHapley Additive exPlanations

Idée initiale : partir de covariables x et d'un modèle f complexes

$$x \rightarrow f(x)$$

Simplifier **localement** la relation

- \mathbf{z} une discréétisation des covariables x
- g une fonction linéaire

$$f(x) \approx g(\mathbf{z}) = \phi_0 + \sum_{i=1}^p \phi_i \mathbf{z}_i$$

Parallèle de la valeur Shapley entre théorie des jeux et machine learning

- les joueurs sont les modalité des variables explicatives
- le gain à répartir est la différence entre la prévision et la moyenne des prévisions

Estimation de SHAP

Tester toutes les coalitions peut vite devenir impossible

- 6 features : 64 coalitions
- 16 features : ~33 millions de coalitions
- et ce n'est qu'une mesure locale

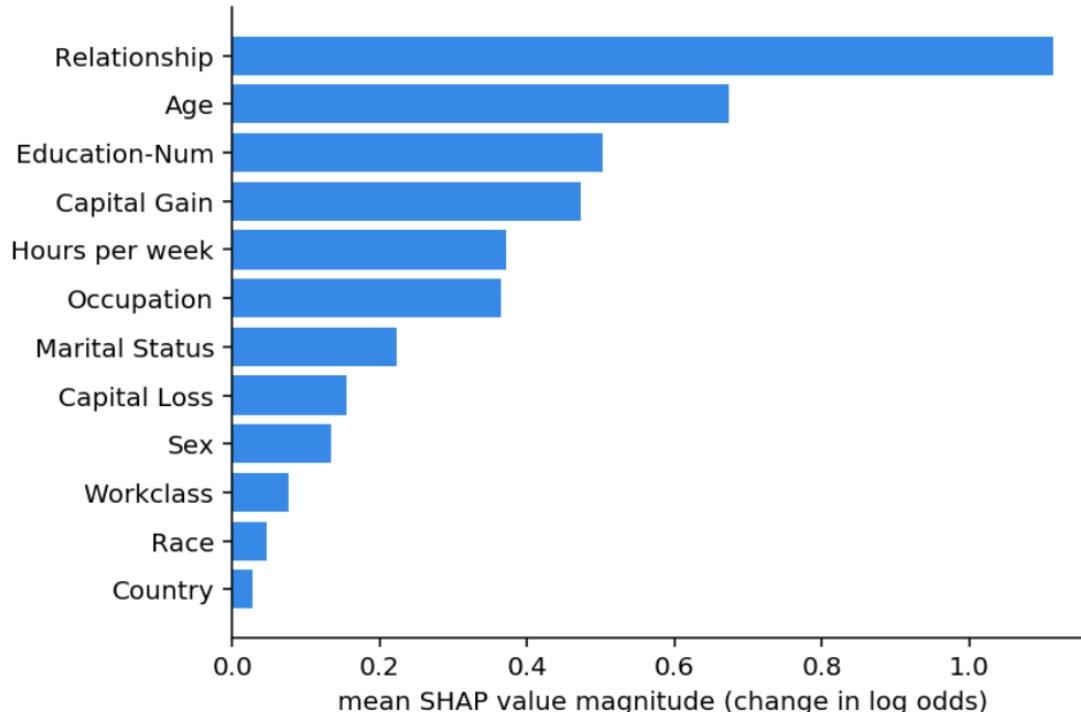
Développement d'algorithmes pour approximer ces valeurs

- Kernel SHAP : fonctionne pour tous les modèles
- Tree SHAP : optimisation pour les modèles tree-based
- Deep SHAP : optimisation pour les réseaux de neurones
- etc.

Vision globale (1/3)

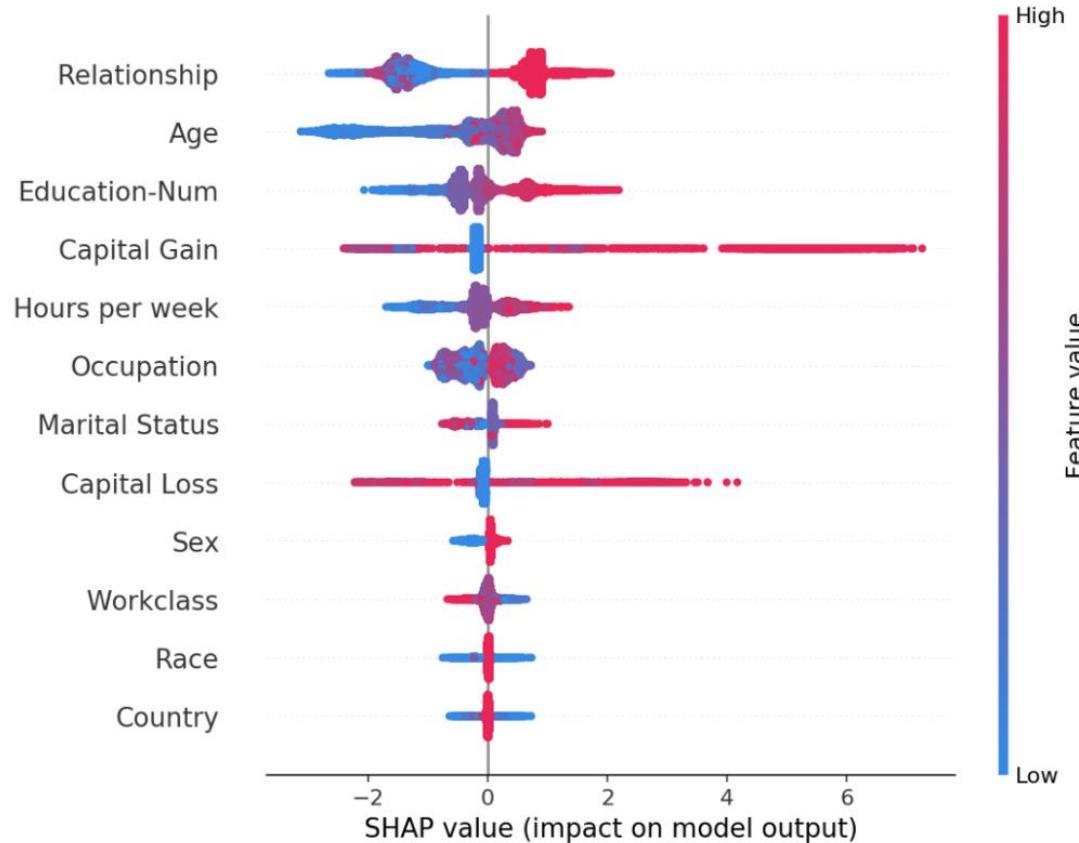
La valeur shapley n'est valable que pour un point donné, c'est une vision locale qui ne permet pas de prendre en compte l'ensemble du modèle

1. Moyenne de la valeur absolue des SHAP (importance feature)



Vision globale (2/3)

2. Densité des valeurs SHAP



Vision globale (3/3)

3. Dépendance des valeurs SHAP en fonction de la feature

