Mesures d'importance relative par décomposition de la performance de modèles de régression

Marouane Il Idrissi, Bertrand Iooss, Vincent Chabridon

Mesures d'importance relative par décomposition de la performance de modèles de régression

Marouane Il Idrissi [1,2,3], Bertrand Iooss [1,2,3], Vincent Chabridon [1,3]

[1] *EDF R&D, 6 Quai Watier, 78400 Chatou, France; marouane.il-idrissi@edf.fr*
[2] *Institut de Mathématiques de Toulouse, 31062, Toulouse, France*
[3] *SINCLAIR AI Laboratory, Saclay, France*

Résumé. En apprentissage statistique supervisé, les mesures d'importance relative ont pour but de quantifier de manière interprétable l'importance des covariables sur la sortie du modèle d'apprentissage, notamment en présence de dépendance entre ces covariables. Dans ce papier, deux mesures particulières (les valeurs de Shapley et les valeurs proportionnelles) sont étudiées. Ces mesures sont inspirées de deux solutions d'allocations issues de la théorie des jeux. Leurs liens avec d'autres mesures connues en régression linéaire (LMG et PMVD) sont présentés. Après une première illustration de leur formulation analytique dans le cas linéaire gaussien à deux variables, leur estimation pratique, dans un contexte de régression logistique, sur un jeu de données public de prévision des feux de fôret (Algerian Forest Fires) est proposée et discutée.

Mots-clés. Mesures d'importance, régression linéaire, régression logistique, Shapley.

Abstract. In the context of supervised statistical learning, the goal of relative importance measures is to quantify, in an interpretable manner, the importance of each input in the model output, even in the context of input dependency. In the present paper, two particular measures (Shapley values and Proportional values) are studied. These measures arise from conceptual games and allocation strategies in game theory. Here, their links with usual importance measures for linear regression (LMG and PMVD) are presented. After a first illustrative analytical derivation in a two-dimensional linear Gaussian case, their practical estimation using a logistic regression model applied to a public dataset (Algerian Forest Fires) is proposed and discussed.

Keywords. Importance measures, linear regression, logistic regression, Shapley.

1 Introduction

En apprentissage statistique, la quantification de l'importance relative vise à produire des méthodes permettant d'identifier et de mesurer l'importance des covariables par le biais de mesures interprétables. Nous nous attachons ici au modèle linéaire $Y = \beta_0 + X^\top \beta$, où $Y \in \mathbb{R}$, $X = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ est le vecteur aléatoire des covariables du modèle, $\beta_0 \in \mathbb{R}$ et $\beta \in \mathbb{R}^d$. On note $\mathbb{V}(X_i) = \sigma_i^2$ et $\mathrm{Cov}(X_i, X_j) = \sigma_i \sigma_j \rho_{ij}$ pour $i \in D$, $D = \{1, \ldots, d\}$. La *décomposition de la variance* de Y fournit la métrique CVD (*covariance decomposition*), qui s'écrit $\mathrm{CVD}_i = \beta_i \sigma_i \sum_{j=1}^{d} \beta_j \sigma_j \rho_{ij}$, et qui permet d'assigner une valeur d'importance

à chaque X_i, $i \in D$ (Feldman 2005). Si les covariables sont indépendantes, la part de variance de Y expliquée par chacune d'entre elles est donnée par $\beta_i^2\sigma_i^2/\mathbb{V}(Y)$. Cependant, dans le cas de covariables corrélées, cette mesure CVD peut être négative ce qui rend son interprétation, en tant que part de variance, sujette à caution.

La notion d'allocation de jeux coopératifs statistiques (Feldman 2005) permet de relier les domaines de la théorie des jeux et de l'apprentissage statistique, afin de construire des mesures d'importance interprétables. Pour un modèle total paramétrique emboîtable $\Theta(X, \beta)$ (construit avec toutes les covariables), une mesure de performance positive et faiblement monotone μ_Θ peut lui être associé et être évaluée pour chaque sous-modèle restreint aux covariables d'indices $S \subset D$, et de performance $\mu_\Theta(S)$. Une famille d'allocations pour le jeu coopératif statistique (D, μ_Θ), dites à *ordres aléatoires*, peut être définie pour chaque covariable $i \in D$ (Weber 1988):

$$\phi_i = \mathbb{E}_p\Big[\mu_\Theta\big(D \setminus S_{i-1}^r\big) - \mu_\Theta\big(D \setminus S_i^r\big)\Big] = \sum_{r \in \mathcal{R}(D)} p(r)\Big(\mu_\Theta\big(D \setminus S_{r(i)-1}^r\big) - \mu_\Theta\big(D \setminus S_{r(i)}^r\big)\Big) \quad (1)$$

où p désigne une fonction de masse de probabilité définie sur $\mathcal{R}(D)$ (l'ensemble des permutations de D), $S_i^r = \{r_j\}_{j=1}^i$ dénote l'ensemble des i$^{\text{èmes}}$ premières composantes de r ($r = (r_1, \ldots, r_d) \in \mathcal{R}(D)$), et $r(i)$ est la position de l'indice i dans r. Cette famille d'allocations permet de redistribuer la performance du modèle total à chacune de ses covariables, facilitant leur interprétation.

Dans le cadre du modèle linéaire, quatre critères permettent de définir une mesure d'importance relative admissible (Johnson and Lebreton 2004; Feldman 2005; Grömping 2007) : la *positivité* ($\forall i \in D$, $\phi_i \geq 0$), l'*exclusion* (soit $\Theta(X, \beta)$, si $\beta_i = 0$, alors $\phi_i = 0$), l'*inclusion* (soit $\Theta(X, \beta)$, si $\beta_i \neq 0$, alors $\phi_i > 0$), la *contribution totale* ($\sum_{i=1}^n \phi_i = \mu_\Theta(D)$). Pour un choix spécifique de p, les allocations définies en Eq. (1) permettent de produire des allocations candidates à mesurer l'importance relative, sous couvert du respect de ces critères.

Notre objectif est d'étudier deux cas particuliers d'allocations, inspirés de résultats généraux en théorie des jeux : les valeurs de Shapley et les valeurs proportionnelles. La Section 2 vise à définir ces mesures d'importance relative dans le cadre de jeux coopératifs statistiques et d'étudier leur admissibilité, ainsi que d'illustrer leur comportement analytique dans le cas linéaire gaussien. La Section 3 est dédiée à un cas d'utilisation sur le jeu de données Algerian Forest Fires, dans un but de classification binaire par modèle de régression logistique.

2 Valeurs de Shapley et valeurs proportionnelles

Les valeurs de Shapley (Shapley 1951) constituent une solution d'allocation pour jeux coopératifs. Elles sont l'unique solution d'une définition axiomatique garantissant le respect de quatre propriétés désirées en théorie des jeux. Dans le contexte des allocations

par modèle à ordres aléatoires (Eq. (1)), les valeurs de Shapley du jeu coopératif statistique (D, μ_Θ) sont équivalentes à choisir p comme étant uniforme sur $\mathcal{R}(D)$ (i.e., chaque ordre est équi-vraisemblable). Ainsi, $\forall r \in \mathcal{R}(D)$, $p(r) = 1/d!$, ce qui amène à la mesure d'importance relative Sh_i, $\forall i \in D$ (Eq. (1) avec $p(r) = 1/d!$). Cependant, Feldman (2005) montre que cette mesure d'importance relative particulière n'est pas admissible au sens des critères énoncés plus haut. En effet, elle ne respecte pas le critère d'*exclusion* : une covariable qui n'est pas présente dans le modèle total (i.e., son paramètre est égal à zéro) peut recevoir une part de performance si elle est corrélée avec une ou plusieurs covariables qui sont présentes dans le modèle (i.e., dont les paramètres sont différents de zéro). Cet effet a été mis en évidence en analyse de sensibilité (Iooss and Prieur 2019).

De plus, le choix d'un a priori uniforme peut être remis en question en remarquant que μ_Θ peut contenir de l'information sur l'importance relative. En effet, si pour une permutation $r = (r_1, \ldots, r_d)$, les contributions séquentielles $M_i(r) = \mu_\Theta\big(D \setminus S^r_{i-1}\big) - \mu_\Theta\big(D \setminus S^r_i\big)$ sont croissantes (i.e., $M_1(r) < M_2(r) < \cdots < M_d(r)$), alors il est probable que les éléments de r soient rangés par ordre croissant d'importance relative. En étendant les *valeurs proportionnelles* (issues des jeux coopératifs, voir Ortmann (2000)) aux jeux coopératifs statistiques, Feldman (2005) introduit la mesure PMD (*proportional marginal decomposition*) par le biais d'une autre définition de p :

$$p(r) = \frac{L(r)}{\sum_{m \in \mathcal{R}(D)} L(m)}, \quad \text{avec} \quad L(r) = \left(\prod_{S \in \{S^r_i\}_{i=1}^d} \big(\mu_\Theta(D) - \mu_\Theta(D \setminus S)\big) \right)^{-1}.$$

L'allocation sur (D, μ_Θ) ainsi définie est donnée par :

$$\mathrm{PMD}_i = \left(\sum_{m \in \mathcal{R}(D)} L(m) \right)^{-1} \sum_{r \in \mathcal{R}(D)} L(r) \Big(\mu_\Theta\big(D \setminus S^r_{r(i)-1}\big) - \mu_\Theta\big(D \setminus S^r_{r(i)}\big) \Big).$$

L'existence et l'unicité de cette fonction de masse sont garanties par une définition axiomatique qui stipule notamment que si $\mu_\Theta(D \setminus \{i\}) = \mu_\Theta(D)$ (i.e., la covariable X_i n'a aucun effet sur la performance du modèle total), alors $\mathrm{PMD}_i = 0$. La valeur de l'allocation est définie sur le jeu coopératif statistique $(D \setminus \{i\}, \mu_\Theta)$, et X_i reçoit une part nulle. Ceci permet de garantir qu'une covariable non-présente dans le modèle total ne reçoive aucune contribution, malgré le fait qu'elle puisse être corrélée aux variables présentes. Cette mesure d'importance relative est *admissible* (cf. Section 1).

Appliquées au modèle de régression linéaire, avec le coefficient de détermination R^2 comme mesure de performance, les valeurs de Shapley du jeu coopératif statistique (D, R^2) sont connus comme étant les indices LMG (Lindeman, Merenda, and Gold 1980). Les valeurs proportionnelles de (D, R^2), quant à elles, sont connues sous le nom de PMVD (*proportional marginal variance decomposition*).

Pour un modèle linéaire à deux covariables $Y = \beta_1 X_1 + \beta_2 X_2$, avec $(X_1, X_2)^\top$ un vecteur gaussien centré vérifiant pour $i = 1, 2$, $\mathbb{V}(X_i) = \sigma_i^2$ et $\mathrm{Cov}(X_1, X_2) = \sigma_1 \sigma_2 \rho$,

associé au coefficient de détermination R^2 comme mesure de performance, les mesures LMG et PMVD sont données analytiquement par :

$$\begin{array}{rclcrcl} \mathrm{LMG}_1 & = & \frac{1}{\mathbb{V}(Y)}\left(\beta_1^2\sigma_1^2 + \beta_1\beta_2\sigma_1\sigma_2\rho + \frac{\rho^2}{2}(\beta_2^2\sigma_2^2 - \beta_1^2\sigma_1^2)\right) ; & \mathrm{PMVD}_1 & = & \frac{\beta_1^2\sigma_1^2}{\beta_1^2\sigma_1^2+\beta_2^2\sigma_2^2}; \\ \mathrm{LMG}_2 & = & \frac{1}{\mathbb{V}(Y)}\left(\beta_2^2\sigma_2^2 + \beta_1\beta_2\sigma_1\sigma_2\rho + \frac{\rho^2}{2}(\beta_1^2\sigma_1^2 - \beta_2^2\sigma_2^2)\right) ; & \mathrm{PMVD}_2 & = & \frac{\beta_2^2\sigma_2^2}{\beta_1^2\sigma_1^2+\beta_2^2\sigma_2^2} \end{array}$$

vérifiant la relation $\mathrm{LMG}_1 + \mathrm{LMG}_2 = \mathrm{PMVD}_1 + \mathrm{PMVD}_2 = R^2(\{1,2\}) = 1$. Dans ce cas précis, les PMVD ne dépendent pas de ρ (ce comportement ne se généralise pas pour $d > 2$), alors que la mesure LMG semble partager la part de variance due à la corrélation équitablement entre X_1 et X_2. De plus, si $\beta_1^2\sigma_1^2 = \beta_2^2\sigma_2^2$, les quatres indices associés aux covariables se confondent. Dans le cas où $\rho = 0$, les deux mesures se comportent de la même façon. Lorsque l'un des deux paramètres $\beta_i = 0$, alors $\mathrm{PMVD}_i = 0$ tandis que LMG_i peut être non-nul, pour des valeurs de ρ non-nulles : ce comportement est contraire au critère d'exclusion.

Dans les cas à plus de deux covariables, l'étude analytique de ces mesures proposée par Grömping (2007) permet de conclure que la mesure LMG aura tendance à répartir les effets dus à la corrélation équitablement entre les covariables, indépendamment de leur importance dans le modèle, tandis que la mesure PMVD aura tendance à attribuer ces effets aux covariables les plus importantes.

3 Prévision des feux de forêt

Dans cette section, nous étendons l'utilisation des mesures d'importance précédentes au cadre du modèle linéaire généralisé pour la régression logistique et les appliquons à un jeu de données public.

Le jeu de données Algerian Forest Fires (Abid and Izeboudjen 2020) contient $n = 244$ observations journalières enregistrées dans deux régions d'Algérie (Bejaia et Sidi Bel-abbes) entre les mois de juin et septembre 2012. Les 8 covariables sont Temp (température maximale en degrés Celsius), RH (humidité relative en %), Ws (vitesse du vent en km/h), Rain (pluviométrie totale en mm), FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code) et ISI (Initial Spread Index). Comme illustré en Figure 1, ces covariables peuvent être très corrélées les unes aux autres. Nous cherchons à modéliser la probabilité qu'un feu de forêt ait eu lieu par régression logistique et nous prenons comme mesure de performance le coefficient de détermination généralisé, qui dans ce cas vaut :

$$R^2(S) = 1 - \frac{\text{déviance du sous-modèle d'indices dans } S}{\text{déviance du modèle nul}}.$$

Estimé sur le jeu de données, nous obtenons $\widehat{R^2} \simeq 0.803$ et un coefficient de prédictivité Q^2 (R^2 en prédiction), calculé par validation croisée égal à $\widehat{Q^2} \simeq 0.79$. Les mesures de multicolinéarité VIF (*variance inflation factor*)[1] ont également été calculées, en plus

1. Fonction vif() du package car sous R.

o

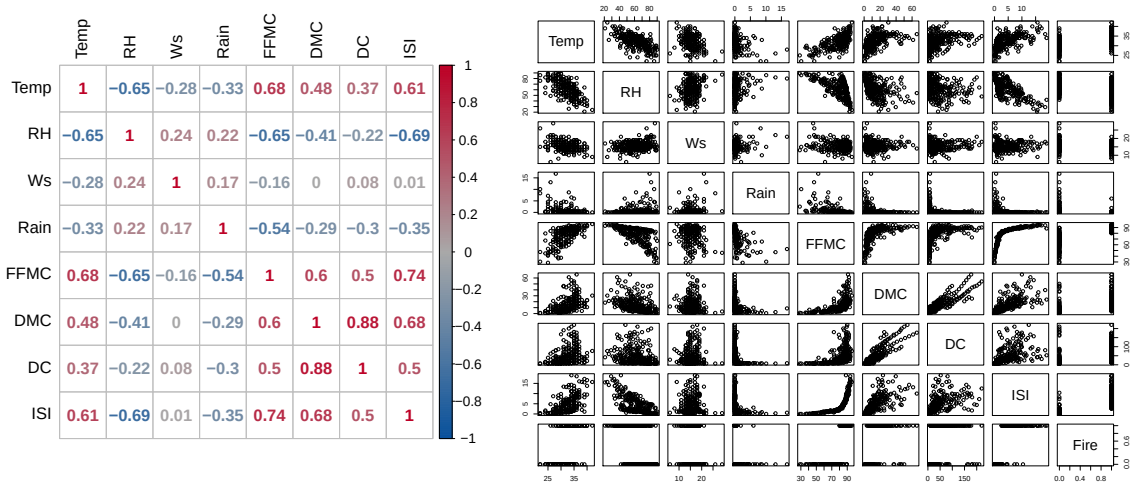

Figure 1: Matrice des corrélations (gauche) et nuages de points croisés (droite) des covariables du jeu de données Algerian Forest Fires.

Covariables	Temp	RH	Ws	Rain	FFMC	DMC	DC	ISI	Total
VIF	1.36	1.90	1.72	1.44	7.08	8.04	6.24	5.04	-
Sh (%)	4.5	3.7	0.4	5.5	33.3	6.2	3.2	23.5	80.3
PMD (%)	0.4	0	0	0.7	69.7	6.4	0	3.1	80.3

Table 1: Mesure de colinéarité et mesures d'importance relative du modèle de régression logistique sur les données Algerian Forest Fire.

des mesures d'importance par valeurs de Shapley et par valeurs proportionnelles[2]. Les résultats se trouvent en Table 1.

Les covariables FFMC, DMC, DC et ISI présentent de fortes valeurs de VIF (i.e., supérieures à 5), ce qui fait écho aux valeurs élevées des corrélations déjà identifiées en Figure 1. La mesure Sh semble favoriser les covariables FFMC et ISI, avec des part respectives de 33.3% et 23.5% de la performance totale. Les parts de performance des autres covariables oscillent entre 3.7% et 6.2%, sauf pour la vitesse du vent, avec une part inférieur. La mesure PMD, quant à elle, attribue une part de près de 69.7% de la performance à la covariable FFMC, avec des contributions à la performance de 6.4% et 3.1% respectivement aux covariables DMC et ISI. Les autres covariables reçoivent moins de 1% de performance. Ceci peut-être expliqué par les fortes corrélations dont les effets sur la performance sont principalement attribués aux covariables ayant un paramètre

2. Fonctions lmg() et emvd() du package sensitivity sous R.

estimé élevé (en valeur absolue). La mesure PMD permet donc de détecter les covariables les plus importantes de manière plus prononcée que la mesure Sh qui aura tendance à lisser l'importance par répartition des effets de corrélation. La covariable RH en est un exemple parlant : étant fortement corrélée linéairement avec FFMC (-0.65) et ISI (-0.69), les valeurs de Shapley auront tendance à lui accorder de l'importance (3.7% de la performance), alors que les valeurs proportionnelles indiquent que son importance est nulle.

En conclusion, nous pouvons donc retenir que, dans le contexte de la régression logistique (resp. linéaire), et en choisissant le R^2 comme critère de performance, les mesures d'importance PMD (resp. PMVD) présentent l'intérêt de remplir les quatre critères d'admissibilité d'une mesure d'importance interprétable contrairement à ses homologues que sont les valeurs de Shapley (resp. LMG).

Nous remercions Nicolas Bousquet (EDF R&D) grâce à qui ce travail a pu être réalisé.

Bibliographie

Abid, F., and N. Izeboudjen. 2020. "Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm." In *Advanced Intelligent Systems for Sustainable Development,* 363–370. Springer International Publishing.

Feldman, B. 2005. "Relative Importance and Value." *SSRN Electronic Journal.*

Grömping, U. 2007. "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition." *The American Statistician* 61 (2): 139–147.

Iooss, B., and C. Prieur. 2019. "Shapley Effects For Sensitivity Analysis With Correlated Inputs: Comparisons With Sobol' Indices, Numerical Estimation And Applications." *International Journal for Uncertainty Quantification* 9 (5): 493–514.

Johnson, J. W., and J. M. Lebreton. 2004. "History and Use of Relative Importance Indices in Organizational Research." *Organizational Research Methods* 7:238–257.

Lindeman, R. H., P. F. Merenda, and R. Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis.* Scott, Foresman.

Ortmann, K. M. 2000. "The proportional value for positive cooperative games." *Mathematical Methods of Operations Research* 51 (2): 235–248.

Shapley, L. S. 1951. *Notes on the n-Person Game – II: The Value of an n-Person Game.* Research Memorandum ATI 210720. RAND Corporation.

Weber, R. J. 1988. "Probabilistic values for games." In *The Shapley Value,* 1st ed., 101–120. Cambridge University Press.