Prédiction de l'age cérébral à partir de l'imagerie cérébrale anatomique

Léa Abriel, Annah Augier, Anne Coustans

February 22, 2024

1 Rationnel scientifique des méthodes de régularisation utilisées

Afin de prédire l'âge cérébral à partir des données qui nous ont été mises à disposition, nous avons décidé de partir sur une régression linéaire en testant des méthodes de régularisation : Lasso, Ridge et elastic net. Nous sommes dans une problématique d'apprentissage supervisé avec une variable quantitative et un nombre excessif de lignes par rapport aux variables. Ainsi, introduire une pénalité sur le vecteur de poids du modèle linéaire permet de réduire la dimensionnalité et le surapprentissage. La pénalité de Ridge, L2 amène les coefficients de la régression vers 0 et ainsi impose le vecteur de coefficient à être petit. La pénalité L1, de Lasso, elimine les variables qui n'ont pas d'intérêt dans la prédiction de la variable expliquée, mais fonctionne moins bien car peut supprimer trop de variables. La validation croisée est utilisée pour évaluer les performances du modèle, étant donné que le nombre d'observations est limité. Ainsi, nous avons premièrement testé la régularisation de Lasso. On cherche alors à minimiser :

$$\min_{\beta} \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{1}$$

Où β représente le vecteur des coefficients du modèle à estimer; N est le nombre d'observations dans l'ensemble de données; y_i est la valeur réelle de la i-ème observation; x_i est le vecteur des caractéristiques de la i-ème observation; λ est le paramètre de régularisation qui contrôle l'intensité de la pénalité L1 et que nous avons donc testé; p est le nombre de caractéristiques ou variables explicatives dans le modèle.
Le mélange entre régression ridge et Lasso nous donne l'elastic net. Nous cherchons à minimiser :

$$\min_{\beta} \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 \right) \tag{2}$$

Où β représente le vecteur des coefficients du modèle à estimer; N est le nombre d'observations dans l'ensemble de données; y_i est la valeur réelle de la i-ème observation; x_i est le vecteur des caractéristiques de la i-ème observation; λ_1 est le paramètre de régularisation pour la pénalité L1; λ_2 est le paramètre de régularisation pour la pénalité L2; p est le nombre de caractéristiques ou variables explicatives dans le modèle.

Nous avons également tenté d'augmenter les performances de notre prédiction en utilisant des méthodes ensemblistes telles que XGBoost et Gradient Boosting, sans grand succès sur le score RMSE.

2 Résultats en validation croisée

Résultats des modèles en validation croisée		
Type de modèle ou traitement	Paramètres utilisés	RMSE sur jeu de données test
LASSO avec la classe ROIsFeatureExtractor	$\lambda = 0.2$	RMSE=7.30
LASSO avec la classe ROIsFeatureExtractor	$\lambda = 0.15$	RMSE=7.31
ELASTIC NET avec la classe VBMFeatureExtractor	$\lambda_1 = 0.10,\ \lambda_1 ratio = 0.84$	RMSE=6.88
Suppression variables corrélées et ELASTIC NET avec la classe ROIsFeatureExtractor	$\lambda_1 = 0.10,\ \lambda_1 ratio = 0.84$	RMSE=7.89

3 Analyse de l'importance des prédicteurs

Nous choisissons alors de conserver le modèle Elastic Net avec la classe VBMFeatureExtractor qui obtient le meilleur score en terme de RMSE. Afin de mesurer l'importance des prédicteurs du modèle et d'en sortir les dix premiers, nous utilisons les valeurs de Shapley sur les variables dont le coefficient est non nul après la pénalité d'elastique net. Le modèle sélectionne 454 variables, sur lesquelles nous calculons la contribution de chaque prédicteur en tenant compte de toutes les interactions possibles entre prédicteurs. Les valeurs de Shapley sont calculées telles que :

$$\phi\_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \tag{3}$$

Où $\phi_i(v)$ est la valeur de Shapley du prédicteur i, calculée comme la somme pondérée des contributions marginales de i à toutes les combinaisons possibles de prédicteurs dans l'ensemble N, excluant i. La fonction de valeur v attribue un score à chaque sous-ensemble de prédicteurs, S, et |S|! (|N| - |S| - 1)! / |N|! est le poids de chaque contribution marginale, avec |S| et |N| représentant respectivement la taille de S et le nombre total de prédicteurs.

En calculant les valeurs de Shapley sur nos 454 prédicteurs, nous obtenons les 10 premiers, qui sont les suivants :

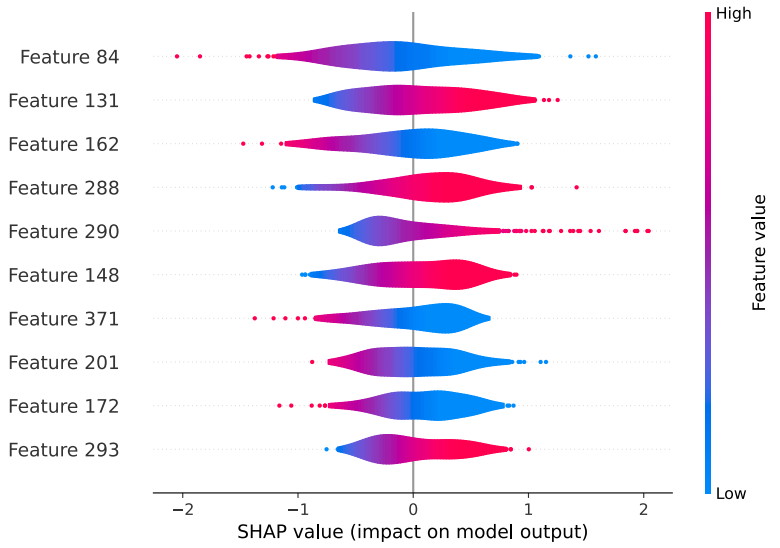

Figure 1: Valeurs de Shapley pour les 10 premiers prédicteurs

Ce graphique nous donne l'importance des dix premières variables avec la valeur de Shapley sur l'axe X et le nom des variables en Y. La couleur indique une forte valeur du prédicteur en rouge et un faible valeur en bleu. On peut interpréter par exemple l'impact de la feature 290 sur l'âge. On peut voir qu'une forte valeur de la feature aura un grand impact sur la variable à expliquer tandis qu'une valeur faible de cette feature aura un impact proche de zéro. Nous pouvons alors supposer que ces 10 premières variables sont les pixels qui se trouvent dans les zones d'intérêt sur le scanner et qui donnent une information sur l'âge du patient.