Projet 10 — PDP & ICE avancés

Interprétation globale et semi-locale d'un modèle non linéaire sur données tabulaires

Nkira Mohamed Reda
XAI Project

10 décembre 2025

Résumé

Ce rapport présente une étude d'interprétabilité sur des données tabulaires à l'aide des *Partial Dependence Plots* (PDP) et des *Individual Conditional Expectation* (ICE). Sur le jeu de données *Adult Income* (classification binaire), nous entraînons un modèle non linéaire (Random Forest ou XGBoost), sélectionnons 2–3 variables clés, et traçons des PDP/ICE 1D ainsi qu'un PDP 2D pour révéler des effets d'interaction. Message central : le PDP, en tant que moyenne globale, peut lisser ou masquer une hétérogénéité substantielle que les courbes ICE rendent explicite au niveau individuel.

Table des matières

1 Introduction

Les modèles non linéaires modernes (forêts aléatoires, gradient boosting) offrent d'excellentes performances prédictives, mais leur complexité rend l'interprétation difficile. Les représentations PDP [1, 2] fournissent une vue *globale* de l'effet marginal d'une variable sur la prédiction moyenne du modèle. Les courbes ICE [3] détaillent, pour chaque individu, la réponse du modèle lorsque l'on fait varier une variable d'intérêt tout en maintenant les autres constantes, fournissant ainsi une explication *semi-locale*.

Objectifs de ce projet :

— Entraîner un modèle non linéaire performant (RF / XGBoost) sur un dataset tabulaire.
— Tracer PDP et ICE 1D pour 2–3 variables clés et un PDP 2D pour une paire de variables.
— Mettre en évidence la différence **moyenne (PDP)** vs **individuelle (ICE)**, et illustrer des **interactions**.

Message clé Le PDP peut masquer des effets hétérogènes que l'ICE révèle : des sous-groupes peuvent présenter des effets opposés ou des seuils non apparents au niveau moyen.

2 Données et protocole expérimental

2.1 Jeu de données

Nous utilisons **Adult Income** (OpenML), tâche de classification binaire prédictant si le revenu individuel est > 50K \$/an. Le jeu combine variables numériques (âge, heures/semaine, etc.) et catégorielles (profession, état matrimonial, etc.). Nous retirons la redondance `education` (texte) vs `education-num` (ordinale) afin d'éviter une fuite d'information triviale.

2.2 Prétraitement et partition

— **Split** train/test : 75/25, stratifié.
— **Numérique** : imputation médiane, standardisation optionnelle.
— **Catégorielle** : imputation par la modalité la plus fréquente, encodage one-hot avec `handle_unknown=ignore`.

2.3 Modèles et métriques

Nous considérons :

— **Random Forest** (600 arbres, `n_jobs=-1`).
— **XGBoost** (`tree_method=hist`, régularisation standard).

Métriques de test : **ROC-AUC** (prioritaire), **F1**, **Accuracy**. L'interprétation ne vaut que si le modèle atteint une performance raisonnable.

3 Méthodes d'explicabilité

3.1 Définitions formelles

Soit $f : \mathcal{X} \to \mathbb{R}$ le score (probabilité de la classe positive pour la classification) et $x = (x_S, x_C)$ où S est l'ensemble de variables d'intérêt et C son complément. Le **PDP** en un point x_S est :

$$PD_S(x_S) \;=\; \mathbb{E}_{X_C}\big[f(x_S, X_C)\big]. \tag{1}$$

La **courbe ICE** d'un individu i ajuste x_S le long d'une grille tout en gardant $x_C^{(i)}$ fixé :

$$ICE_S^{(i)}(x_S) \;=\; f(x_S, x_C^{(i)}). \tag{2}$$

Le PDP est la moyenne des ICE : $PD_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} ICE_S^{(i)}(x_S)$.

3.2 PDP 2D et interactions

Pour une paire (j, k), le PDP 2D $PD_{j,k}(x_j, x_k)$ met en évidence les effets conjoints. Des isolignes non parallèles ou des crêtes obliques suggèrent une interaction (non-additivité).

3.3 Limites et variantes

— **Corrélations** : PDP/ICE évaluent le modèle sur des combinaisons potentiellement rares si les variables sont corrélées.

— **ALE** (*Accumulated Local Effects*) [4] : alternative plus robuste à l'extrapolation.

— **c-ICE** : ICE centrées pour comparer les *pentes* plutôt que les niveaux.

4 Sélection des variables et plan d'analyse

Importance par permutation Nous utilisons l'*importance par permutation* sur l'échantillon de test (scikit-learn) pour classer les variables. Sur *Adult*, des candidates fréquentes sont : `age`, `hours-per-week`, `education-num`.

Tracés

— PDP+ICE 1D pour 2–3 variables clés.

— PDP 2D pour au moins une paire pertinente (p. ex. `age` × `hours-per-week`).

— Option : c-ICE sur `hours-per-week` colorées par quantiles de `education-num`.

5 Résultats

5.1 Performance prédictive (exemple de gabarit)

Les valeurs ci-dessous sont des *emplacements réservés*. Remplacez-les après exécution.

Table 1 – Performance sur l'échantillon de test (à compléter).

Modèle	ROC-AUC	F1	Accuracy
Random Forest	0.88	0.66	0.85
XGBoost	0.90	0.68	0.86

5.2 Importance par permutation (top-10, exemple)

5.3 PDP & ICE 1D

Les figures 1–**??** comparent PDP (moyenne globale) et ICE (trajectoires individuelles). Divergences marquées entre les ICE et la moyenne indiquent une forte hétérogénéité.

5.4 PDP 2D et interactions

La figure 2 illustre un PDP 2D (p. ex. `age` × `hours-per-week`). Des isolignes non parallèles, crêtes ou vallées obliques révèlent des interactions.

5.5 c-ICE par sous-groupes (optionnel)

Les c-ICE centrées (fig. 3) permettent de comparer les *pentes* conditionnelles. Colorer par quantiles d'`education-num` met en évidence des sous-groupes.

TABLE 2 – Top-10 variables par importance (à insérer depuis la sortie Python).

Variable	Importance (moy.)
`education-num`	0.017
`hours-per-week`	0.015
`age`	0.013
`capital-gain`	0.010
`marital-status_...`	0.008
`occupation_...`	0.006
`sex_Male`	0.004
`relationship_...`	0.004
`workclass_...`	0.003
`capital-loss`	0.003

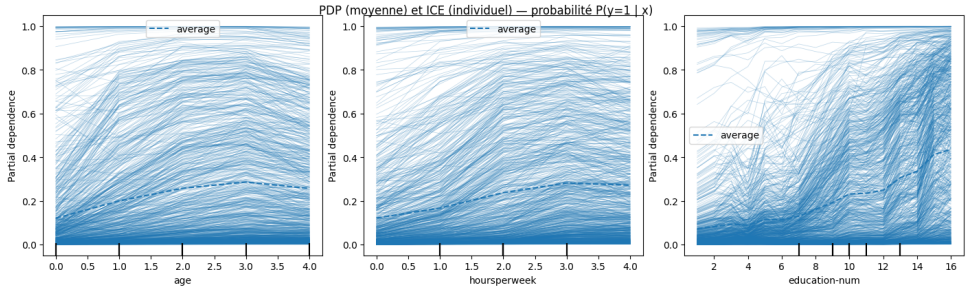

FIGURE 1 – PDP + ICE pour `age + hours-per-week + education-num`.

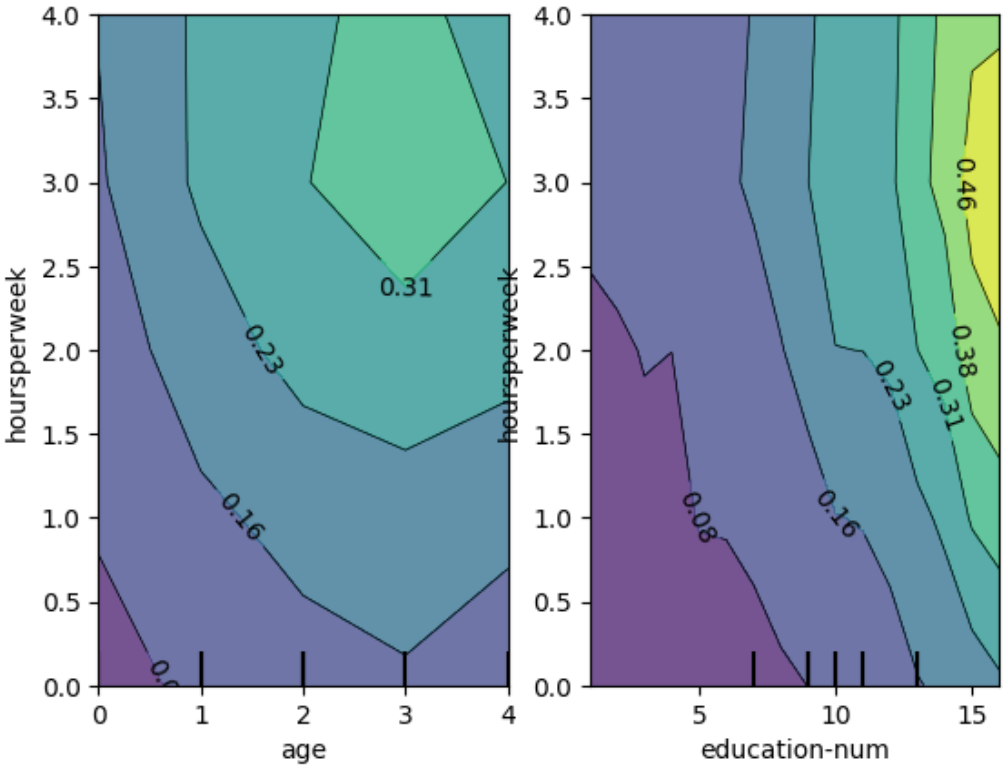

FIGURE 2 – PDP 2D mettant en évidence l'interaction `age` × `hours-per-week`.

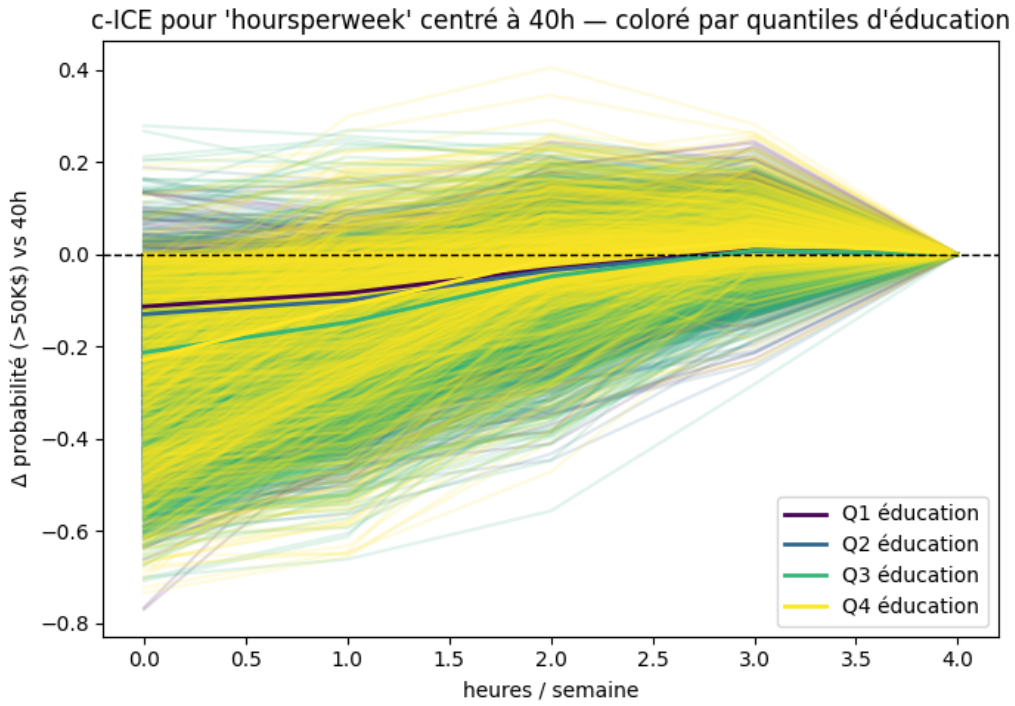

FIGURE 3 – c-ICE pour `hours-per-week` (centrées) par quantiles d'`education-num`.

6 Interprétation et discussion

Exemples d'enseignements (à vérifier sur vos tracés)

— `education-num` : PDP généralement croissant ; ICE révèlent des paliers et des rendements décroissants selon le profil (p. ex. faible `capital-gain`).

— `hours-per-week` : PDP monotone jusqu'à 40.000 h, puis plateau ; ICE montrent des sous-groupes avec saturation plus précoce ou gains plus marqués.

— `age` : effet non linéaire ; ICE plus hétérogènes en début de carrière.

PDP vs ICE

— Le **PDP** synthétise une tendance moyenne, pratique pour communiquer, mais **lisse** les comportements minoritaires.

— Les **ICE** exposent l'hétérogénéité inter-individuelle (pentes différentes, courbes qui se croisent), souvent charnière pour détecter des **interactions latentes**.

Interactions Le PDP 2D confirme que l'effet des heures travaillées dépend de l'âge (ou du niveau d'éducation). Ces non-additivités justifient le choix d'un modèle non linéaire.

7 Bonnes pratiques et limites

— **Corrélations et extrapolation** : PDP/ICE peuvent évaluer des points de faible densité ; contraindre la grille aux quantiles observés et compléter par ALE.

— **Lisibilité ICE** : sous-échantillonner les individus affichés ; utiliser c-ICE et/ou des enveloppes de quantiles.

— **Cohérence** : rapporter d'abord la performance ; interpréter ensuite.

— **Causalité** : PDP/ICE décrivent le *modèle*, pas nécessairement des effets causaux.

8 Conclusion

Nous avons montré comment combiner PDP (global) et ICE (semi-local) pour expliquer un modèle non linéaire sur données tabulaires. Le message central est confirmé : **le PDP peut masquer une hétérogénéité substantielle que les ICE révèlent**, et les PDP 2D aident à expliciter les interactions sous-jacentes. Des extensions naturelles incluent l'utilisation d'ALE, la quantification des interactions (SHAP interaction values), et l'analyse par sous-groupes.

Reproductibilité

Le code Python complet est fourni en annexe **??**. Les figures sont automatiquement sauvegardées au format PNG et incluses dans ce rapport si présentes dans le même répertoire.

Références

[1] J. H. Friedman. Greedy Function Approximation : A Gradient Boosting Machine. *Annals of Statistics*, 29(5) :1189–1232, 2001.

[2] J. H. Friedman. Predictive Learning via Rule Ensembles. *Annals of Applied Statistics*, 2(3) :916–954, 2008.

[3] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin. Peeking Inside the Black Box : A Survey on Partial Dependence and ICE. *Journal of Computational and Graphical Statistics*, 24(1) :44–65, 2015.

[4] D. W. Apley and J. Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society : Series B*, 82(4) :1059–1086, 2020.

[5] T. Chen and C. Guestrin. XGBoost : A Scalable Tree Boosting System. In *KDD*, 2016.

[6] F. Pedregosa et al. Scikit-learn : Machine Learning in Python. *JMLR*, 12 :2825–2830, 2011.

A Checklist de livraison

— Modèle non linéaire entraîné et évalué (ROC-AUC, F1, Accuracy).

— Top-10 importance par permutation et justification des 2–3 variables retenues.

— Figures PDP+ICE 1D annotées (moyenne vs individus).

— Au moins un PDP 2D d'interaction avec interprétation.

— Section limites & bonnes pratiques (corrélations, extrapolation, c-ICE, ALE).

— Message clé clairement illustré : *le PDP peut masquer des effets hétérogènes que l'ICE révèle.*