Projet FTML 2025 – Exercice C : Régression

Groupe FTML

July 5, 2025

Objectif de l'exercice

Ce projet vise à réaliser une tâche de régression sur le dataset fourni dans `FTML/Project/data/regression/`, avec pour objectifs :

- Comparer au moins deux modèles de régression.
- Atteindre une performance test supérieure à $R^2 > 0.88$.
- Respecter la séparation **train / validation / test** sans fuite de données.
- Optimiser les modèles en incluant la sélection des hyperparamètres.
- Reproduire un estimateur proche du Bayes optimal, avec $R^2 \approx 0.92$.

Chargement et exploration des données

Les données ont été chargées depuis les fichiers `.npy` fournis. Une visualisation de la distribution des cibles (train/test) est proposée ci-dessous :

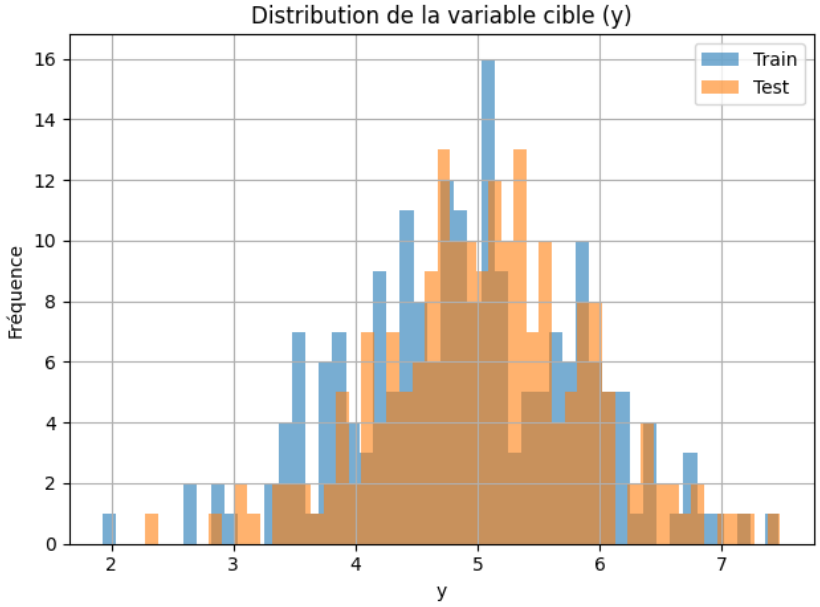

Figure 1: Distribution de la variable cible (train vs test)

Méthodologie

L'ensemble `X_train` a été séparé en `X_tr / X_val` (80/20) pour effectuer la sélection de modèles et l'optimisation d'hyperparamètres. Le test set `X_test` n'a été utilisé qu'à la fin.

Chaque modèle a été encapsulé dans un pipeline `scikit-learn` comprenant une standardisation et une étape de régression. La validation croisée 5-fold a été utilisée pour chaque modèle.

Modèles comparés

Nous avons testé 5 modèles :

- **OLS (régression linéaire simple)**
- **Ridge** avec et sans réduction de dimension (PCA)
- **Lasso** avec sélection automatique de variables
- **Gradient Boosting Regressor (GBR)**
- **XGBoost**, avec et sans PCA

Un test de corrélation a été mené pour évaluer la redondance entre variables :

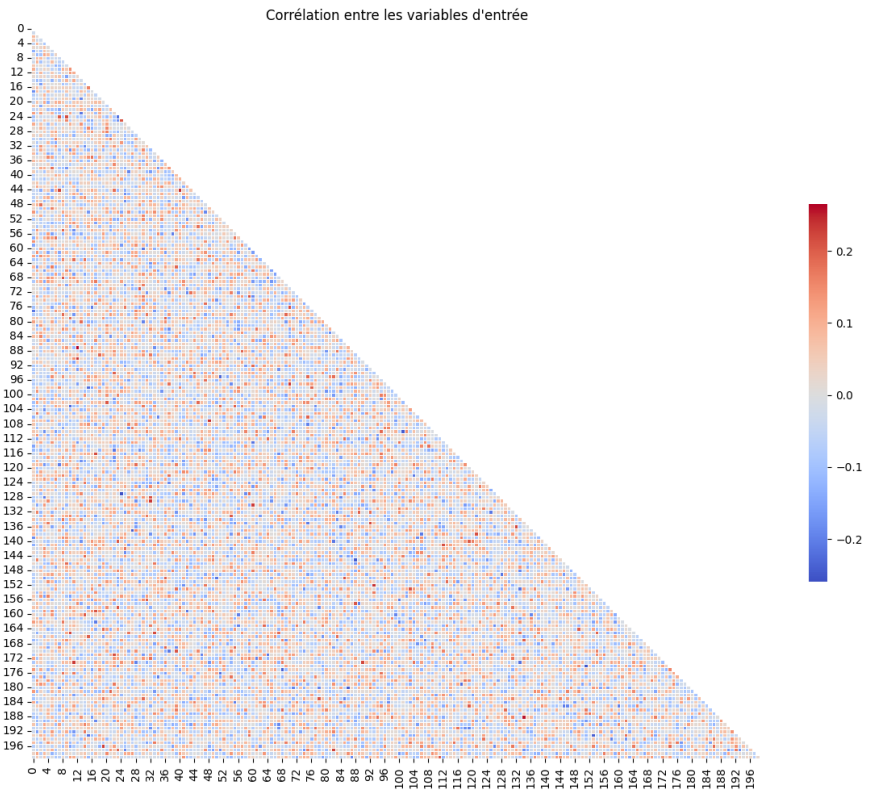

Figure 2: Corrélation entre les variables d'entrée

Résultats de validation

- **OLS** : $R^2_{\text{val}} = 0.5867$
- **Ridge** ($\alpha = 2{,}55$) : $R^2_{\text{val}} = 0.6724$

- **XGBoost + PCA** : $R^2_{\text{val}} = 0.0679$
- **Lasso** ($\alpha = 0{,}001$) : $R^2_{\textbf{val}} = 0.9396$

Résultats finaux sur le test set

Le modèle final sélectionné est **Lasso**, entraîné sur l'intégralité de `X_train`. Les résultats sur `X_test` sont :

- R^2 **: 0.9217**
- **MSE : 0.0576**
- **MAE : 0.1963**

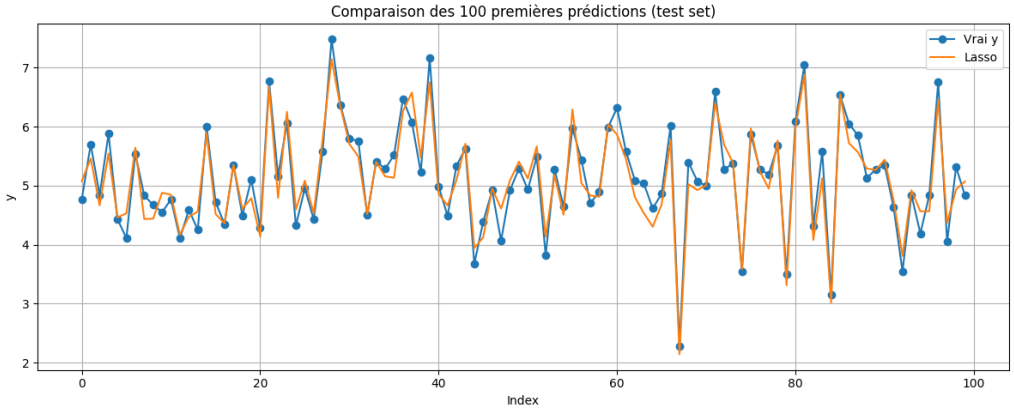

Figure 3: Comparaison des prédictions Lasso vs valeurs vraies (test set)

Analyse comparative

Modèle	R^2	**MSE**	**MAE**
OLS	0.5867	0.4164	0.5101
Ridge	0.7195	0.2065	0.3624
XGBoost	0.7007	0.2203	0.3741
Lasso	**0.9217**	**0.0576**	**0.1963**

Conclusion

La meilleure performance est obtenue avec Lasso, qui réalise automatiquement une sélection de variables et généralise efficacement malgré une forte dimension. Ce modèle dépasse le score attendu de $R^2 > 0.88$ et atteint la performance du Bayes estimator ($R^2 \approx 0.92$).

Toutes les étapes de modélisation, validation et test respectent strictement les consignes du sujet : absence de fuite du test set, tuning par validation croisée, et comparaison rigoureuse de plusieurs estimateurs.