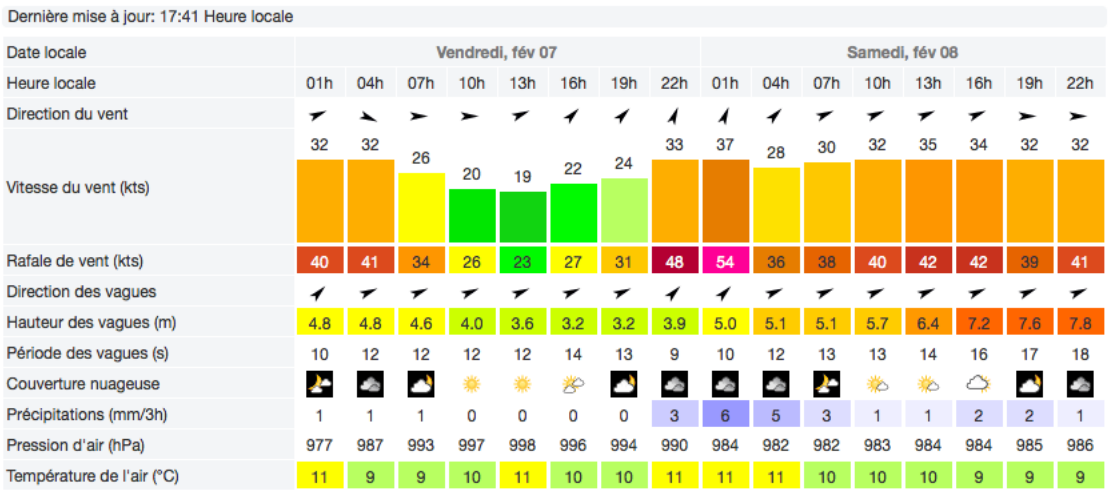
Machine Learning

Régression : Hauteur des vagues et force du vent



Dernière mise à jour: 17:41 Heure locale

Date locale	Vendredi, fév 07								Samedi, fév 08							
Heure locale	01h	04h	07h	10h	13h	16h	19h	22h	01h	04h	07h	10h	13h	16h	19h	22h
Direction du vent																
Vitesse du vent (kts)	32	32	26	20	19	22	24	33	37	28	30	32	35	34	32	32
Rafale de vent (kts)	40	41	34	26	23	27	31	48	54	36	38	40	42	42	39	41
Direction des vagues																
Hauteur des vagues (m)	4.8	4.8	4.6	4.0	3.6	3.2	3.2	3.9	5.0	5.1	5.1	5.7	6.4	7.2	7.6	7.8
Période des vagues (s)	10	12	12	12	12	14	13	9	10	12	13	13	14	16	17	18
Couverture nuageuse																
Précipitations (mm/3h)	1	1	1	0	0	0	0	3	6	5	3	1	1	2	2	1
Pression d'air (hPa)	977	987	993	997	998	996	994	990	984	982	982	983	984	984	985	986
Température de l'air (°C)	11	9	9	10	11	10	10	11	11	11	10	10	10	9	9	9

Agathe Blanvillain [1], Kevin Tran [2]

1. agathe.blanvillain.1@etudiant.univ-rennes1.fr
2. kevin.tran@etudiant.univ-rennes1.fr

Introduction

Notre étude a pour but de prédire la taille des vagues en fonction de la force du vent.

Pour cela, on va utiliser les méthodes de régression apprises en Machine Learning.

On étudiera les données qui correspondent à un échantillon de 8000 individus et 2000 variables, accompagné de leur label.

Pour travailler sur ce nombre important de variable nous allons prendre un sous ensemble d'individus aléatoire. Ainsi, nous comparerons les résultats de 3 méthodes de régression :

1. Linéaire (OLS) ;
2. Ridge ;
3. Lasso.

Pour chaque régression, nous comparerons les modèles par les méthodes de validation croisées : data-splitting et K-fold (le Leave one-out étant lui trop coûteux pour notre étude au vue du nombre d'individus).

Régression

Régression linéaire (OLS)

On commence par faire une régression linéaire (OLS). L'objectif ici est de tracer le nuage de points et de déterminer l'erreur de prédiction avec la méthode du K-fold. Nous avons commencé par essayer de faire ces prédictions avec 1000 individus parmi les 8000 donnés. On remarque rapidement que la prédiction n'est pas idéale. Le nombre d'individus étant inférieur au nombre de variables, on se retrouve devant un système surdéterminé. Comme on le voit sur le graphique suivant, tous les points semblent être sur la droite. Le modèle semble donné un résultat parfait. Cependant, la méthode du 5-fold donne une erreur de prédiction de 1.35.

FIGURE 1 – Régression linéaire pour 1000 individus

Nous avons décidé de changer le nombre d'individus pour un nombre supérieur au nombre de variables (2000) pour une visualisation plus fiable, mais inférieur au nombre d'individus initial (8000) pour une réponse plus rapide. Prenons alors 4000 individus. On remarque que les points semblent moins bien disposés que la figure 1. Pour ce qui est de l'erreur elle est pourtant meilleure, la méthode du 5-fold nous donne une erreur de prédiction de 1.17.

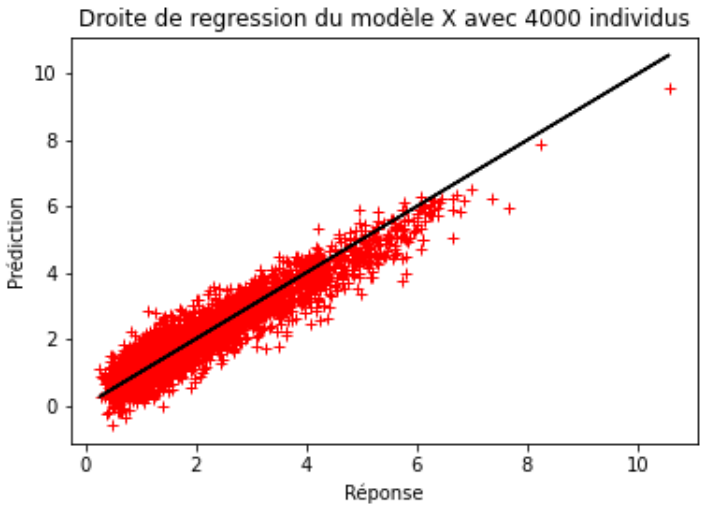


FIGURE 2 – Régression linéaire pour 4000 individus

Choisir 4000 individus parmi les 8000 semble donc être une bonne méthode.
On sait de plus que théoriquement, la hauteur des vagues est plutôt linéaire avec le carré de la vitesse du vent (donné dans l'énoncé du rendu). On va donc essayer de comparer les résultats de régressions des modèles X,X^2, X^3 et X^4 avec 4000 individus.

	Modèle	erreur de prédiction
1	X	1.17
2	X^2	1.09
3	X^3	1.35
4	X^4	1.99

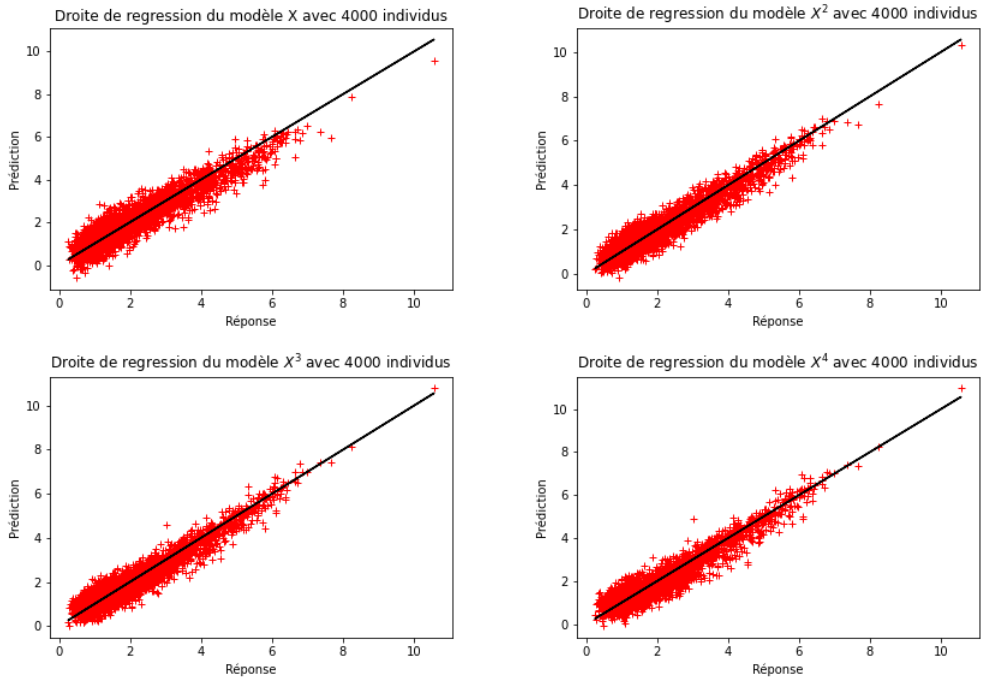


FIGURE 3 – régression linéaire avec 4000 individus pour les modèles X, X^2, X^3, et X^4

Les résultats obtenus montrent bien que le modèle X^2 est le plus représentatif.

Ridge et Lasso

Nous allons faire notre étude avec ces deux modèles de régression avec un échantillon de 1000 individus et 2000 variables (ce qui est moins problématique avec ces modèles) et allons faire une prédiction avec le carré du vent (modèle X^2).

Nous avons plusieurs résultats sur les deux modèles (Ridge et Lasso) en effectuant deux méthodes de validations croisée data-splitting et 5-Fold, la méthode de leave one-out étant trop gourmande en temps, même en réduisant l'échantillon à 1000 individus.

Pour le modèle Ridge, nous trouvons que la meilleure valeur de α pour minimiser le MSE se trouve dans l'intervalle $[10^{-1}, 10]$. Nous avons donc effectuer un data splitting avec 100 itérations et un 5-Fold avec 100 itérations dans cette intervalle avec un pas de 0.1. Nous présentons ce résultat sur la figure 4.

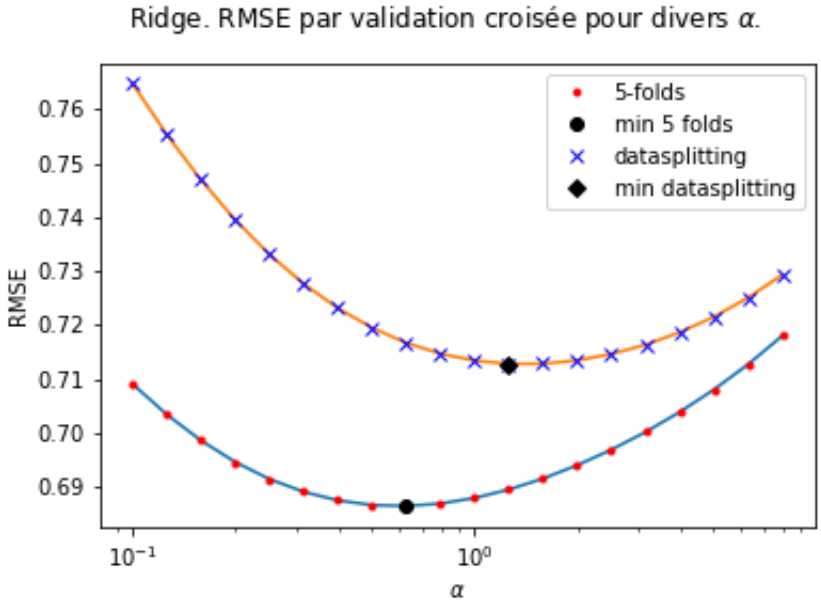


FIGURE 4 – Ridge. Data splitting et 5 Fold, RMSE en fonction des valeurs de α

Pour le modèle Lasso, nous avons un autre intervalle d'étude plus grand pour être sur de pouvoir trouver la valeur de α qui minimise le mieux le RMSE. L'intervalle d'étude est $[10^{-4}, 10^{-2}]$ avec un pas de 0.2 (pour diminuer le temps de calcul) comme nous pouvons le voir figure 5.

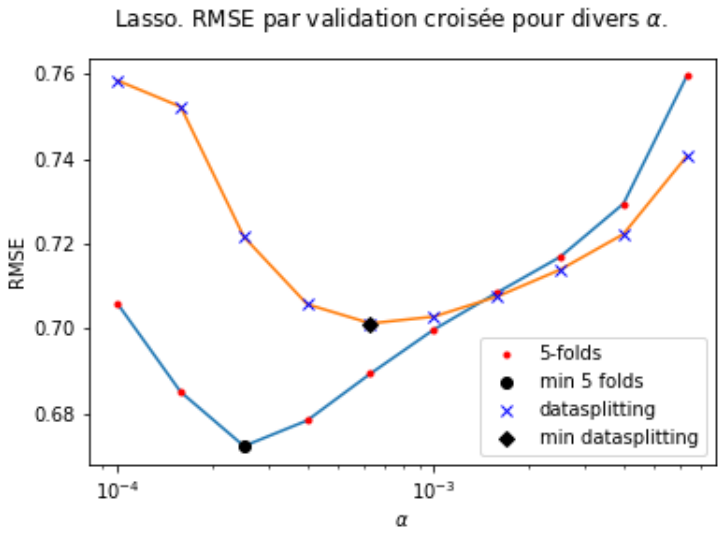


FIGURE 5 – Lasso. Data splitting et 5 Fold, RMSE en fonction des valeurs de α

La méthode de Lasso nous donne aussi un moyen de réduire considérablement le nombre de variable nécessaire pour faire de prédiction. Comme le montre les courbes figure 6. Nous arrivons à 137 variables (au lieu de 2000) pour un RMSE de 0.67 en 5-Fold.

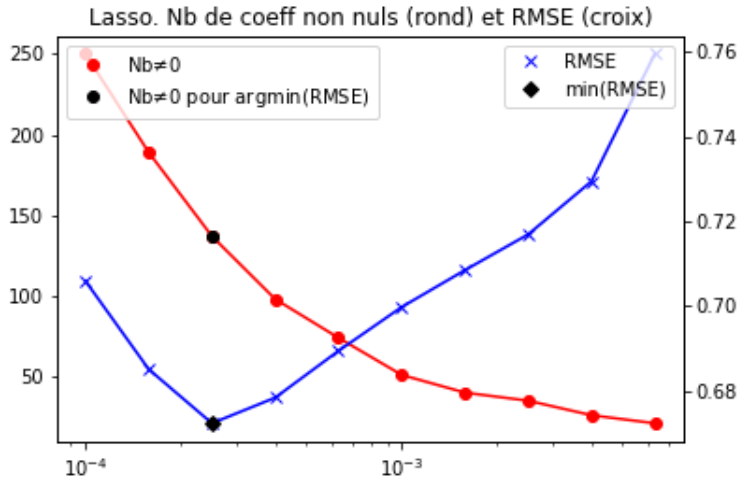


FIGURE 6 – Nombre de variable et RMSE associé

Conclusion

En regroupant les résultats des différents modèles de régression / modèles linéaire et méthodes de validation croisée. Nous trouvons un meilleur résultat de prédiction pour le Lasso avec un modèle qui prend le carré du vent avec $\alpha = 2 \times 10^{-3}$ qui minimise le RMSE pour le 5-Fold.

	valeur de α	
Modèle/Méthode	Data-splitting	5-Fold
Ridge	1.26	0.63
Lasso	6×10^{-3}	2×10^{-3}

FIGURE 7 – Valeur de α qui minimise le RMSE

	Erreur de prédiction	
Modèle/Méthode	Data-splitting	5-Fold
Linéaire (OLS)		1.09
Ridge	0.71	0.69
Lasso	0.70	0.67

FIGURE 8 – Comparaison des différents modèle selon leur erreur de prédiction