Rapport final

Apprentissage statistique en actuariat
ACT-4114

Équipe 09

Rapport

Inondations en Californie

Par	*Numéro d'identification*
Maryjane Bastille	111 268 504
Danny Larochelle	111 174 586
Henri Lebel	111 286 185
Isabelle Legendre	536 768 666
Félix-Antoine Paris	536 776 223

Travail présenté à
Monsieur
Olivier Côté

16 avril 2023

UNIVERSITÉ LAVAL

Faculté des sciences et de génie
École d'actuariat

Table des Matières

Introduction

Modèle de base

Ajustement des modèles

Modèle linéaire (À spécifier)

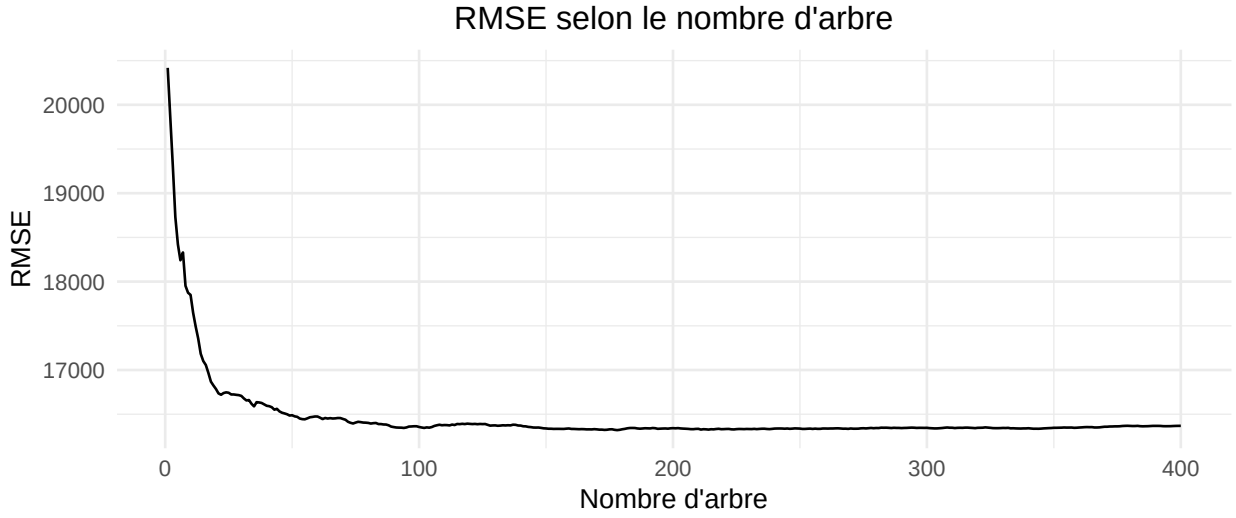
Modèle des k plus proches voisins

Arbre de décision

Bagging

Forêt aléatoire

Pour la forêt aléatoire, on commence avec quatres prédicteurs possibles pour chaque séparation, *i.e.* $m = 4$, car $\lfloor 12/3 \rfloor = 4$. Cette valeur correspond à la "règle du pouce" en régression où l'on utilise la partie entière du nombre de valeurs explicatives divisé par 3. Deplus, en utilisant une proportion de de 50% pour les échantillons bootstrap, on aide a diminuer la corrélation entre les arbres.



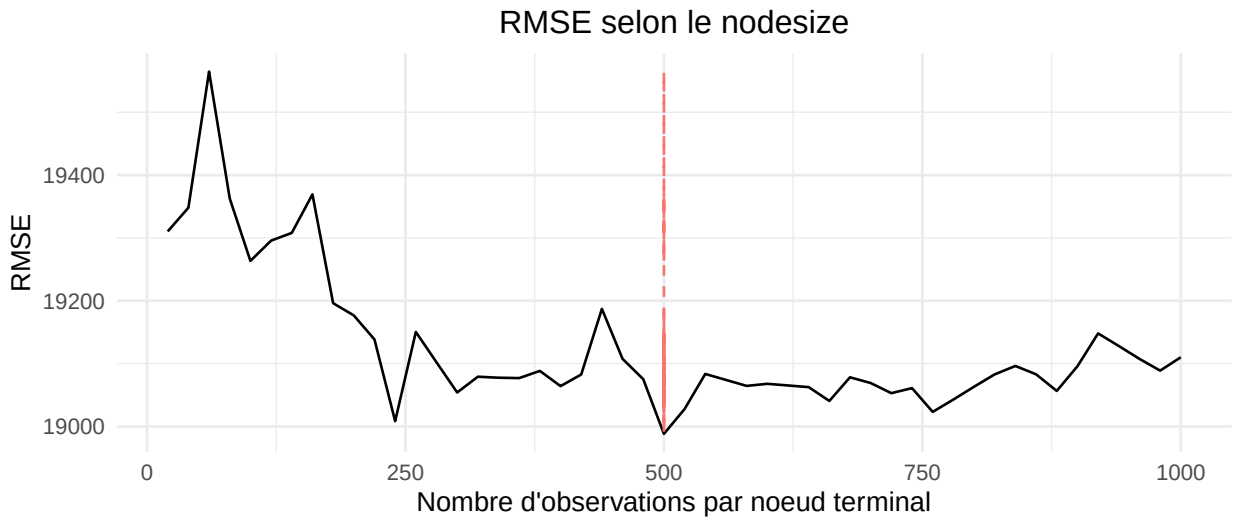
Étant en régression, la racine de l'erreur quadratique moyenne, ou RMSE, sera utilisée comme mesure de comparaison. On remarque ici (Graphique no. #) que la RMSE se stabilise aux alentours de 100-150 arbres, on utilisera alors 200 arbres pour l'optimisation des autres hyperparamètres, puisqu'on ne peut pas surajuster en ayant trop d'arbre avec les forêts aléatoires. Maintenant, on regarde plus en profondeur le nombre de prédicteurs possible à chaque séparation d'un arbre, la variable `mtry`.

Table 1: RMSE par rapport au mtry

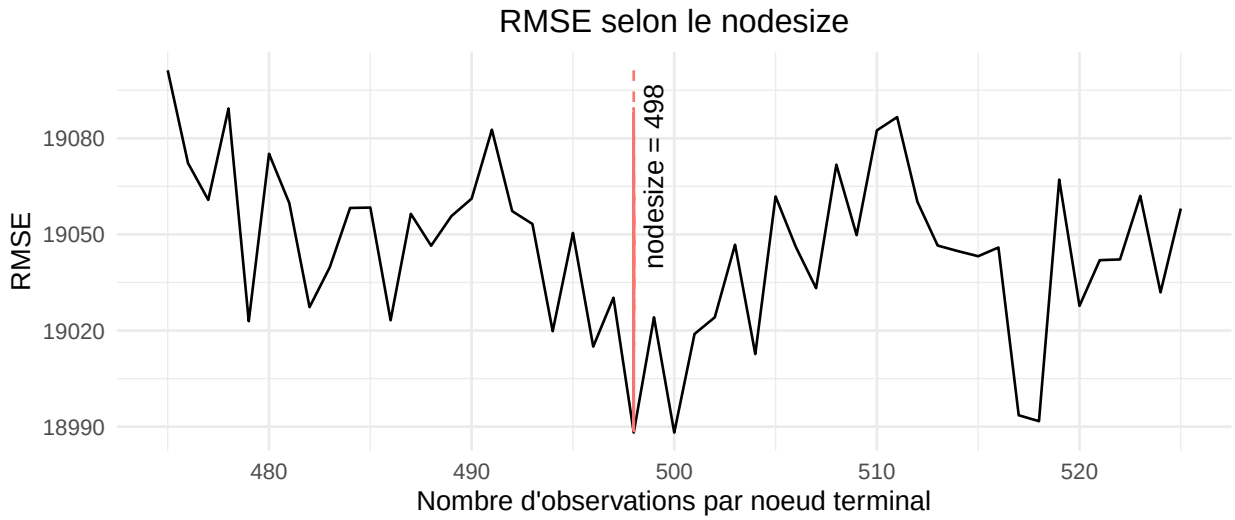
mtry	1	2	3	4	5	6	7	8	9	10	11	12
RMSE	19176	19141	19074	18920	18714	18568	18433	18331	18236	18126	18148	18042

Les résultats de la table no. # ont été obtenus par validation croisée à 5 plis, pour ainsi réduire le biais d'échantillonage. L'utilisation des 12 choix de variables explicatives à chaque noeud minimise la RMSE.

Pour éviter un surajustement dû à des arbres inutillemnet trop profonds, on devra ajuster la valeur de `nodesize`, mais il est impossible de le faire directement avec le package `caret`. Puisque le modèle est entraîné sur 8325 observations, les valeurs de 1000 et moins seront testées et comparées. Pour limiter le temps de calcul, un premier entraînement sera fait par bond de 20.

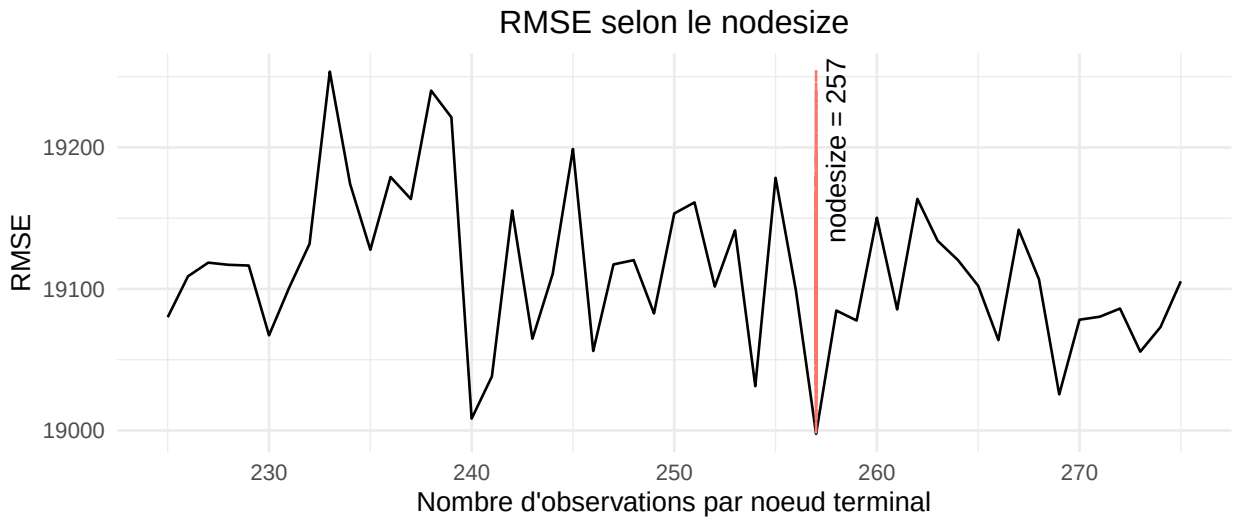


Dans le graphique no. #, la valeur minimale de nodesize est de 500. Puisque l'analyse précédente à été effectuée par bonds de 20, on la fera à nouveau de manière plus précise entre 475 et 525. On remarque aussi qu'il y ait un minimum local près de 250, par prudence on vérifiera aussi les valeurs entre 225 et 275.



Dans le graphique no. #, la valeur minimale de nodesize est de 500. Puisque l'analyse précédente à été effectuée par bonds de 20, on la fera à nouveau de manière plus précise entre 475 et 525.

On remarque aussi qu'il y ait un minimum local près de 250, par prudence on vérifiera aussi les valeurs entre 225 et 275.



Tel qu'on peut le voir dans les graphiques no. # et no. #, les valeurs des minimums locaux sont de 498 et de 257 pour l'hyperparamètre nodesize. Le nodesize qui minimise la RMSE est de 498, tel que l'on peut dans la table no. #.

Table 2: RMSE aux minimums locaux

Nodesize	RMSE
498	18988.15
257	18997.59

Par conséquent, les hyperparamètres finaux pour le modèle "Forêt aléatoire" sont ceux décrits dans la table suivante.

Table 3: Valeurs des hyperparamètres du modèle final

Hyperparamètre	Valeur
Nombre d'arbres	200
Nombre de choix de variables à chaque noeud	12
Nombre d'observation dans les noeuds terminaux	498

Boosting

Gradient Boosting

Extreme gradient boosting

Comparaison des modèles

Interprétation des meilleurs modèles

Conclusion

Bibliographie

The Federal Emergency Management Agency (2023). FIMA NFIP Redacted Claims - v1.

Récupéré de https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v1