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)

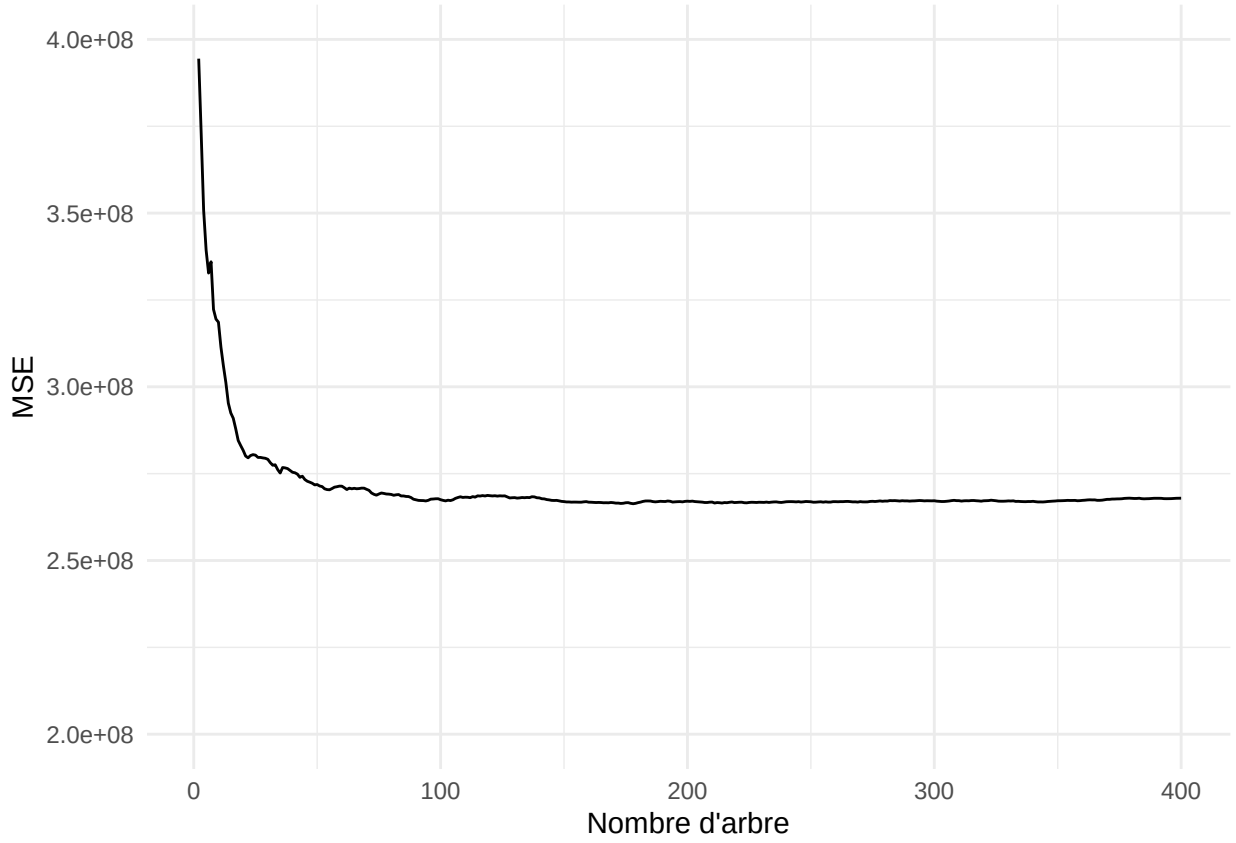
Modèle des k plus proches voisins

Arbre de décision

Bagging

Forêt aléatoire

Pour la forêt aléatoire, on commence avec quatres prédicteurs possibles pour chaque séparation, *i.e.* $m = 4$, car $\lfloor 13/3 \rfloor \approx 4$. 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.



On remarque ici que l'erreur quadratique 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. Étant en régression, la racine de l'erreur quadratique moyenne sera utilisée comme mesure de comparaison (RMSE).

```
## Random Forest
##
## 8325 samples
##   13 predictor
##
## No pre-processing
```

```
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 6661, 6659, 6660, 6660, 6660
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared    MAE
##    1    19064.43  0.03070055  9320.729
##    2    19033.52  0.08921163  9288.920
##    3    18942.42  0.11818723  9243.940
##    4    18807.79  0.13083681  9168.107
##    5    18645.09  0.13453873  9068.836
##    6    18479.90  0.14274494  8971.898
##    7    18320.93  0.14639816  8906.369
##    8    18255.60  0.14815068  8852.988
##    9    18182.85  0.15039063  8806.335
##   10    18072.50  0.15987800  8754.674
##   11    18037.38  0.15784604  8736.219
##   12    18011.15  0.15821619  8691.154
##   13    17904.56  0.17070541  8662.935
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 13.
```

La valeur optimal sera donc de 13

Pour éviter un surajustement dû à des arbres 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‘ données, les valeurs de 100 et moins seront testées par bonds de 5.

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