Rapport Projet 2 SY19

Données Communities

Données manquantes

Le jeu de données *Communities* est composé de 128 variables explicatives et d'une donnée à prédire : *ViolentCrimesPerPop*. Ce jeu de données, en plus d'être très gros et complexe, possède beaucoup de valeurs manquantes, représentées par des NA. Il faut donc tout d'abord gérer ces données avant de les fournir aux modèles à apprendre.

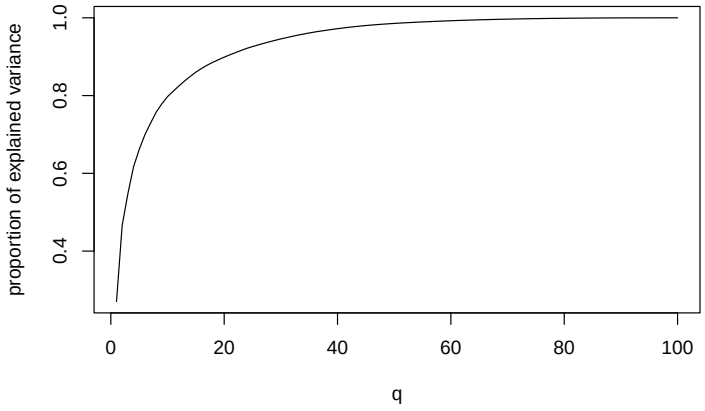
Plusieurs solutions sont envisageables pour gérer les données manquantes. Dans notre cas, nous avons vérifié le taux de NA présents pour chaque variables explicatives. Si ce taux est supérieur à 50%, nous considérons que cette variable n'est pas assez "complète" pour pouvoir vraiment expliquer la variable à prédire. De ce fait, nous la supprimons carrément du jeu de données. Cette méthode nous a permis de supprimer 26 variables.

Dans le cas où le taux de NA dans une colonne est inférieur à 50%, nous remplaçons ces valeurs manquantes par la moyenne de cette variable, disponible dans le fichier *communities.names*. Cette solution s'appelle l'imputation de valeurs manquantes, et elle crée du biais dans le jeu de données. Ainsi, si nous abusons de cette méthode, l'apprentissage pourrait être totalement erroné.
Dans notre cas, seule une variable possédait un taux de NA inférieur à 50%, et de plus, elle ne possédait en fait qu'une seule valeur manquante. Ainsi, l'imputation de cette valeur manquante n'ajoute qu'un biais très faible et négligeable, ce qui est un très bon compromis pour garder cette variable.

Le jeu de données ainsi obtenu ne contient plus que 102 variables sans aucune valeur manquante.

Analyse rapide des données

Avant de se lancer dans l'apprentissage des modèles, nous nous sommes intéressés à la corrélation des variables entre elles. Nous nous sommes aperçus que les variables sont pour la plupart correlées mais peu ont coefficient de corrélation supérieur à 50/60%. Il est donc probable que certaines variables ne soient pas toutes utiles à l'apprentissage des modèles. Ceci est visible également lorsque nous faisons une ACP : 21 composantes suffisent pour avoir 90% de variance expliquée.



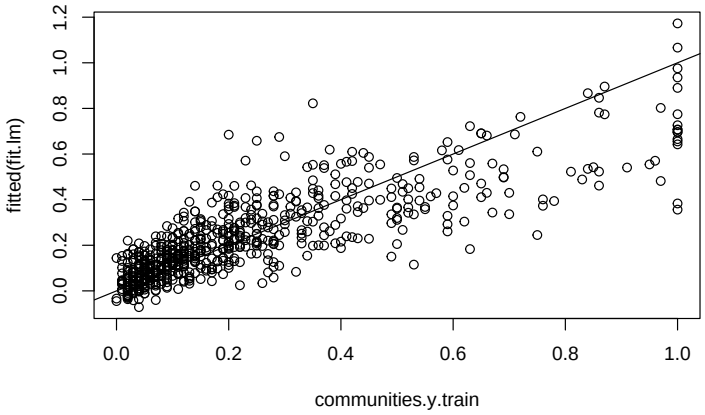
Nous avons donc décidé de privilégier des modèles simples, avec peu de variables explicatives, même si nous avons également testé des modèles plus complexes.

Modèles testés

Les modèles testés sont les suivants : Régression linéaire, régression ridge et lasso, Arbre de décision, Random Forest, et SVM.

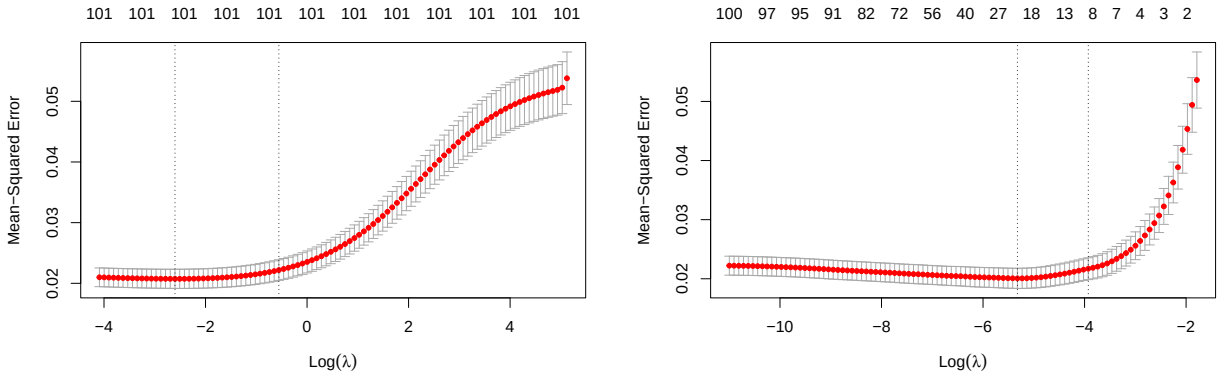
Régression linéaire avec sélection de variables Nous avons premièrement testé un modèle linéaire avec toutes les variables du jeu de données. Les résultats obtenus étaient très élevés comparés aux autres (10 à 30 fois plus élevés). Nous avons donc testé un modèle linéaire en choisissant dix variables explicatives, sélectionnées grâce à la Backward subset selection. Nous avons ainsi obtenus les 10 variables avec les coefficients les plus significatifs, et nous les avons injecté dans notre modèle linéaire : *racepctblack, pctUrban, MalePctDivorce, PctKids2Par, PctPersDenseHous, PctHousOccup, RentLowQ, MedRent, MedOwnCostPctIncNoMtg,* et *NumStreet.*

Les résultats ainsi obtenus sont bien plus satisfaisants malgré la simplicité du modèle. Le modèle a donc l'air de bien s'y adapter et d'être robuste. Le MSE tourne autour de 0,019 et 0,022.



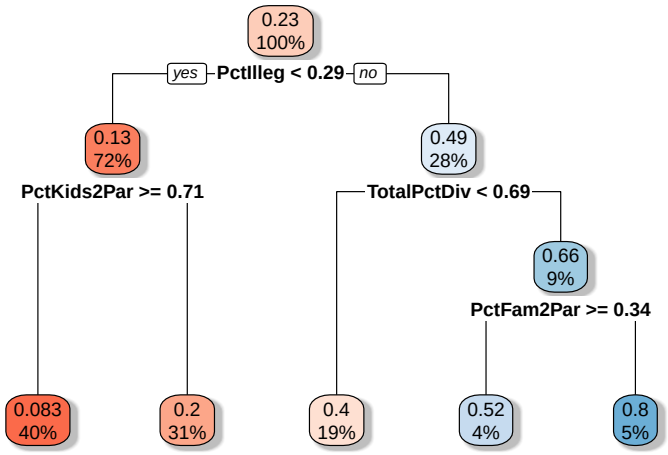
A noter que nous avons également testé une GAM avec ces mêmes variables. Le résultat obtenu est plus ou moins le même lorsque nous n'utilisons pas de splines. L'influence de ces variables est donc potentiellement additive.

Régression ridge et lasso Toujours dans l'optique de réduire le nombre de variables, nous avons testé ces deux modèles.



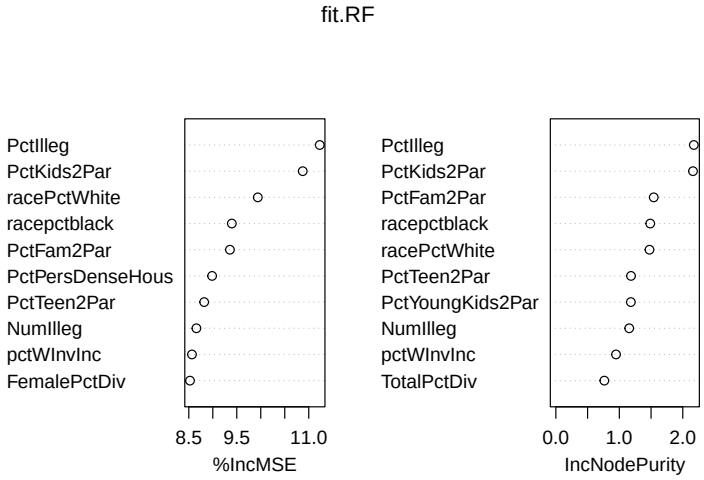
Les résultats ensuite obtenus grâce aux deux modèles sont relativement similaires, avec un MSE tournant au voisinage de 0.0187 pour le ridge, et 0.0182 pour le lasso. En observant les coefficients beta de chaque modèle, nous observons que le nombre de coefficients mis à zéro dans le lasso est très élevé (environ 80 coefficients pour une centaine de variables). Ceci est vérifiable également dans le régression ridge, car ces mêmes coefficients ont des valeurs très faibles (souvent d'un ordre de grandeur des centimèmes ou des millièmes). Ainsi, cela prouve bien que toutes les variables n'ont pas une grande influence sur la variable de réponse, et que certaines peuvent être rejetées, comme nous l'avons fait dans notre modèle linéaire précédemment.
Ces deux modèles de régression ridge et lasso sont robustes et donnent quasiment les mêmes résultats quelques soient les splits train/test donnés.

Arbre de décision Nous avons ensuite testé une méthode plus couramment utilisée pour la classification: les arbres de décision. Cet arbre a été généré à partir de toutes les variables explicatives puis élagué selon la méthode vu dans le TP7.



Nous observons que seules 5 variables explicatives sont utilisées dans cet arbre. Malgré la simplicité de cet arbre, le MSE de ce dernier est d'environ 0.028. Cette valeur n'est pas si mauvaise contrairement au modèle linéaire contenant toutes les variables, mais ce modèle reste moins performant que ceux testés plus haut.

Random Forest Une option plus robuste que l'arbre de décision sont les Random Forests. L'apprentissage de ce modèle nous a, sans surprise, donné de meilleurs résultats que ceux de l'arbre de décision, avec un MSE tournant autour de 0.019/0.020. Cette performance est similaire voire meilleure que les modèles que nous avons testé précédemment. Il est intéressant de noter que le graphe de l'importance des variables nous montre que les variables les plus importantes obtenues avec cette méthode sont parfois les mêmes que celles choisies par Backward Subset selection, vu plus haut. Enfin, grâce à la robustesse du modèle dû aux mélanges de plusieurs arbres de décision, cette méthode a l'air d'être un choix pertinent pour ce jeu de données. En effet, les Random Forests calculant différents arbres en retirant certaines variables prédictives puis en en faisant une moyenne globale, ils permettent de discriminer les variables qui n'ont que très peu d'influence sur la variable de réponse.



SVM Enfin, nous avons testé un SVM en utilisant la méthode vue en TD pour optimiser le métaparamètre C. Le MSE obtenu après avoir déterminé ce paramètre varie beaucoup en fonction du split train/test des données, allant de 0.020 jusqu'à 0.024 et plus. Cette méthode n'a donc pas l'air de bien s'adapter au jeu de données.

Conclusion

Les différents modèles donnent des résultats plus ou moins bons. Malgré cette divergence de résultats, certains modèles permettent de confirmer l'importance (ou non) des différentes variables pour expliquer la variable *ViolentCrimesPerPop*. De plus, nous observons que la régression ridge et le lasso, le modèle linéaire à 10 variables explicatives, ainsi que le Random Forest sont plutôt performants et robustes de part leur simplicité. Nous avons donc choisi d'utiliser le Random Forest comme modèle final.

Pour conclure, ces modèles nous ont permis de mettre en avant l'importance de l'influence de certaines variables sur la variable de réponse, par exemple le taux de population afro-américaine, le pourcentage de foyer avec 2 parents possédant des enfants, ou le pourcentage d'habitations denses (avec plus d'une personnes par chambre). Il est intéressant de noter que d'un modèle à un autre, les variables les plus "importantes" sont différentes, mais certaines se retrouvent dans plusieurs modèles. Nous pouvons donc croire que ces dernières ont une forte influence sur la détermination de la variable à expliquer.