Rapport projet SY19

Emma Falkiewitz, Lucie Béréziat

December 2024

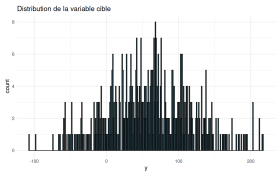
1 Introduction

Le but de ce rapport sera de présenter l'ensemble de notre démarche et les différents modèles employés pour analyser les jeux de données régression, classification ainsi que le jeu de données choisi sur Kaggle.

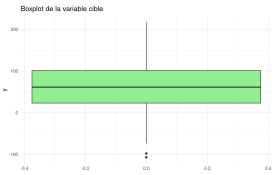
2 Jeu de données régression

2.1 Analyse exploratoire

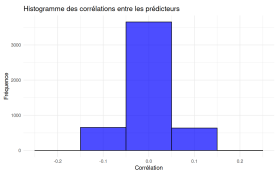
Cette analyse exploratoire avait pour objectif d'avoir une vue d'ensemble du jeu de données, de détecter de possibles outliers et de déterminer la corrélation entre les variables. Nous avons obtenu les résultats suivants:



(a) Histogramme de la variable cible



(b) Boxplot de la variable cible



(c) Histogramme des corrélations

Figure 1: Résultats de l'analyse exploratoire

Le jeu de données possède 100 prédicteurs et a seulement deux valeurs aberrantes. La variable cible a une distribution proche d'une gaussienne. On peut également noter une faible corrélation entre les variables. A partir de ces résultats (1) nous pouvons en déduire qu'une sélection de variables de type lasso ou elastic net serait intéressante pour proposer un modèle simplifié et plus

facilement interprétable. Pour cela, nous avons réalisé une validation croisée imbriquée afin d'avoir un lambda le plus précis possible. Par la suite, nous avons finalement opté pour un elastic net permettant de nous assurer du bon choix du alpha (toujours grâce à la validation croisée imbriquée) quand le modèle sera entraîné sur l'ensemble du jeu de données. On obtient comme résultats un alpha de 0.9475 et un lambda de 0.32. La valeur d'alpha confirme bien notre choix de n'avoir pas choisi lasso qui nous aurait apporté bien plus d'imprécisions. Sur les 100 prédicteurs du jeu de données, 71 ont été retenus après l'application de l'elastic net.

2.2 Modèles testés

Une série de modèles classiques a été testée. Pour nous assurer de la pertinence des modèles, nous avons appliqué une validation croisée avec 10 folds et avec une division 0.70 - 0.30. Cette validation croisée s'est faite à partir de la librairie caret permettant de simplifier le code. Nous avons obtenu les résultats suivants:

Modèles testés	MSE
Régression Linéaire	21844
Random Forest	2134.9
Smooth Splines	146.9
Elastic net	146.1

Table 1: Résultats obtenus directement à partir du jeu de données régression

Nous pouvons voir à partir des résultats (1) que le modèle Random Forest et le modèle Régression Linéaire n'étaient pas adaptés à ce jeu de données. Un travail devait donc se faire au niveau de la simplification du jeu de données. La démarche était ensuite d'aller plus loin dans la recherche d'un pertinent en se focalisant sur le type de modèle. L'elastic net nous donnant déjà une MSE intéressante, l'idée a été de l'améliorer en ajoutant un autre modèle permettant de représenter davantage des relations non-linéaires possibles entre les prédicteurs. Cependant, nous n'avons pas considéré l'utilisation de modèles de type SVM comme pertinente pour ce jeu de données. En effet, nous avons estimé qu'ils étaient trop complexes, on obtenait déjà de bons résultats avec un simple elastic net. Voici les résultats obtenus en utilisant elastic net comme pré-traitement (sélection de variables):

Modèles testés	MSE
Smooth Splines (k=10 (valeur par défaut), bs='ts')	135.5
Natural Splines	135.9
Random Forest	2102.9
Smooth Splines	144.4

Table 2: Résultats obtenus après elastic net

2.3 Analyse du meilleur modèle obtenu

Par rapport aux résultats obtenus (2), nous avons donc choisi le modèle Smooth Splines avec une base 'ts' qui est une pénalité de lissage. Cette pénalité s'applique sur les dérivées d'ordre élevé mais également de manière plus légère à l'espace nul. Cela permet de réduire la complexité du modèle en pénalisant des termes n'apportant pas d'informations significatives. Les Smooth Splines ont été appliquées sur un GAM grâce à la librairie mgcv. Pour faciliter les calculs et donc obtenir des résultats plus rapidement, la fonction bam, correspondant à gam() pour de grands jeux de données a été utilisée. Avec la pénalisation, le modèle a bien été simplifié. Parmi les 71 variables sélectionnées avec l'elastic net, 57 ont un degré de liberté supérieur à 0.

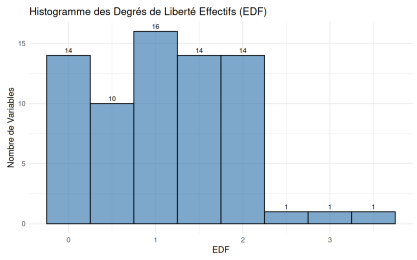
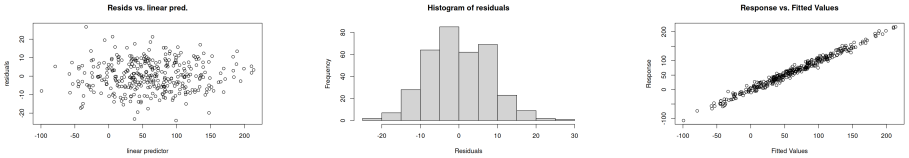


Figure 2: Histogramme de la distribution des degrés de liberté

On observe d'après le graphique 2 que la majorité des degrés de liberté sont faiblement élevés (soit compris entre 0 et 2), cela veut dire qu'une grande partie des prédicteurs ont une relation linéaire avec la variable cible. On observe également quelques relations non-linéaires entre les prédicteurs et la variable cible avec un degré de liberté compris entre 2 et 4. Ces degrés de liberté permettent de saisir la complexité des relations entre les données et la variable cible sans être trop sensible au bruit.



(a) Quantiles vs déviance (b) Histogramme des résidus (c) Réponse vs valeurs prédites

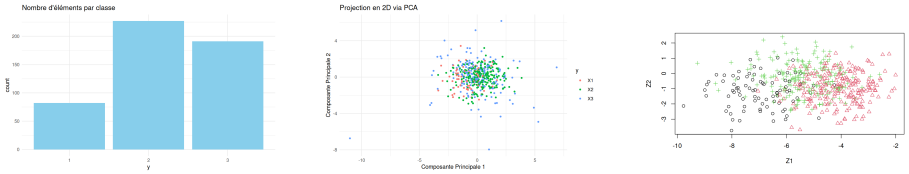
Figure 3: Analyse des résidus du modèle Smooth Splines bs='ts'

L'analyse des résidus du modèle (3) nous permet de constater que l'histogramme des résidus (3a) ressemble approximativement à une gaussienne et le diagramme des résidus en fonction des prédicteurs linéaires (3b) nous indique que les résidus semblent être centrés autour de 0. Ces informations nous permettent d'en déduire que le modèle est adapté au jeu de données. Ce modèle nous a permis d'obtenir une MSE de 135,4013 sur Maggle. La question de normaliser et centrer les données avant l'application des modèles s'est posée. Nous avons cependant observé que cela n'avait pas d'impact sur le résultat final.

3 Jeu de données classification

3.1 Analyse exploratoire

Le jeu de données possède 50 prédicteurs. Il s'agira dans cette partie d'observer la fréquence de chaque classe afin de voir s'il y a un déséquilibre entre les classes. Un PCA et une FDA nous permettront d'avoir une vue en 2 dimensions de ce jeu de données et d'orienter notre choix de modèles.



(a) Répartition des éléments par classe (b) Analyse en Composantes Principales (c) Factor Discriminant Analysis

Figure 4: Résultats de l'analyse exploratoire

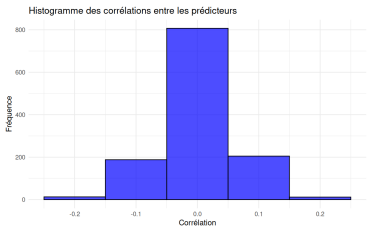


Figure 5: Corrélation entre les prédicteurs

Ces graphiques (4) nous permettent de voir qu'il y a un déséquilibre concernant les classes. En effet, les éléments de la classe 1 sont sous-représentés par rapport aux deux autres. La PCA ne nous permet pas d'avoir une représentation en 2 dimensions adaptée pour appliquer un modèle dessus. En revanche, la FDA donne un meilleur résultat et on peut visualiser plutôt "correctement" 3 groupes distincts. On peut également voir d'après le graphique 5 que les prédicteurs sont faiblement corrélés entre eux.

3.2 Modèles testés

Pour l'ensemble des méthodes employées, nous avons normalisé et centré les données.

3.2.1 Méthode 1

A partir de cette analyse, plusieurs options se sont présentées à nous. L'idée a tout d'abord été de rééquilibrer les classes et d'appliquer une PCA en gardant l'ensemble des variables tel que le pourcentage de variance expliquée soit égale à 90%. A partir de ce jeu de données nous avons testé des modèles. Les résultats obtenus sont les suivants :

Modèles testés	Accuracy
Multinom	0.6534
SVM Radial	0.8284
Random Forest	0.7273
SVM Poly	0.8186
SVM Linear	0.6373
rda	0.7784
knn	0.5147

Table 3: Résultats obtenus après la méthode 1

Modèles testés	Sensitivité	Spécificité	Balanced Accuracy
Multinom	1: 0.7794 2: 0.6324 3: 0.5882	1: 0.9191 2: 0.8235 3: 0.7574	1: 0.8493 2: 0.7279 3: 0.6728
SVM Radial	1: 0.9265 2: 0.7353 3: 0.8235	1: 0.9706 2: 0.9191 3: 0.8529	1: 0.9485 2: 0.8272 3: 0.8382
Random Forest	1: 0.8088 2: 0.7794 3: 0.6912	1: 0.9559 2: 0.8382 3: 0.8456	1: 0.8824 2: 0.8088 3: 0.7684
SVM Poly	1: 1.0000 2: 0.8529 3: 0.6029	1: 0.9412 2: 0.8382 3: 0.9485	1: 0.9706 2: 0.8456 3: 0.7757
SVM Linear	1: 0.8529 2: 0.6324 3: 0.4265	1: 0.8603 2: 0.8088 3: 0.7868	1: 0.8566 2: 0.7206 3: 0.6066
rda	1: 0.9265 2: 0.7941 3: 0.6912	1: 0.9485 2: 0.8750 3: 0.8824	1: 0.9375 2: 0.8346 3: 0.7868
knn	1: 0.9412 2: 0.26471 3: 0.3382	1: 0.4926 2: 0.93382 3: 0.8456	1: 0.7169 2: 0.59926 3: 0.5919

Table 4: Autres mesures

Cette méthode présente cependant quelques soucis, placer le smote avant la PCA entraîne des erreurs au niveau du test et au niveau de l'estimation des paramètres. Ces erreurs ont également une certaine influence sur la PCA. On observe que pour l'ensemble des modèles testés, la sensitivité est assez mauvaise pour la classe 3. L'autre problème de cette méthode est le sur-apprentissage. Nous avons testé le modèle rda sur Maggle et nous avons seulement eu un score de 0.6782 pour l'Accuracy. La taille du jeu de données a également des répercutions sur la différence entre les deux scores.

3.2.2 Méthode 2

Afin d'améliorer nos résultats et pour ne pas obtenir des erreurs à cause du smote, l'idée a été d'appliquer cette fois-ci une PCA en gardant l'ensemble des variables tel que le pourcentage de variance expliquée soit égale à 90% puis de rééquilibrer le train lors de l'application du modèle. On obtient cette fois-ci les résultats suivants :

Modèles testés	Accuracy
Multinom	0.5705
SVM Radial	0.5973
Random Forest	0.6577
SVM Poly	0.5973
SVM Linear	0.5705
rda	0.6443
knn	0.3289

Table 5: Résultats obtenus après la méthode 2

Modèles testés	Sensitivité	Spécificité	Balanced Accuracy
Multinom	1: 0.54167 2: 0.6176 3: 0.5263	1: 0.88800 2: 0.7654 3: 0.6630	1: 0.71483 2: 0.6915 3: 0.5947
SVM Radial	1:0.41667 2:0.6618 3: 0.5965	1: 0.96800 2: 0.7407 3: 0.6196	1: 0.69233 2: 0.7013 3: 0.6080
Random Forest	1: 0.37500 2: 0.7794 3: 0.6491	1: 0.97600 2: 0.7778 3: 0.6739	1: 0.69633 2: 0.7639 3: 0.6615
SVM Poly	1: 0.33333 2: 0.6912 3: 0.5789	1: 0.94400 2: 0.7284 3: 0.6522	1: 0.63867 2: 0.7098 3: 0.6156
SVM Linear	1: 0.6667 2: 0.6618 3: 0.3158	1: 0.8080 2: 0.6914 3: 0.7717	1: 0.7373 2: 0.6766 3: 0.5438
rda	1: 0.41667 2: 0.7941 3: 0.6491	1: 0.92000 2: 0.8148 3: 0.7500	1: 0.66833 2: 0.8045 3: 0.6996
knn	1: 1.0000 2: 0.2353 3: 0.4035	1: 0.5120 2: 0.9506 3: 0.7717	1: 0.7560 2: 0.5930 3: 0.5876

Table 6: Autres mesures

Les résultats (5) ne sont pas vraiment concluants. Les modèles simples ou complexes n'arrivent pas à représenter correctement les données. En utilisant la méthode 1 avec rda, nous trouvions 0.6782% sur Maggle. Nous n'avons donc pas testé la méthode 2. On observe que pour l'ensemble des modèles testés avec la méthode 2 (6), on a un taux de sensitivité plus faible pour la classe 3 et la classe 1.

3.2.3 Méthode 3

La troisième méthode consiste à utiliser la FDA pour y appliquer des modèles. Le but était d'obtenir de meilleurs résultats en maximisant la séparation inter-classes. Un rééquilibrage se fait dans le train, lors de l'apprentissage de chaque modèle. Les résultats obtenus sont les suivants :

Modèles testés	Accuracy
Multinom	0.7047
Random forest	0.7248
SVM Radial	0.6980
SVM Linear	0.6913
SVM Poly	0.7047
rda	0.7047
knn	0.7181

Table 7: Résultats obtenus après la méthode 3

Modèles testés	Sensitivité	Spécificité	Balanced Accuracy
Multinom	1: 0.7917 2: 0.7206 3: 0.6491	1: 0.9440 2: 0.8272 3: 0.7500	1: 0.8678 2: 0.7739 3: 0.6996
Random Forest	1: 0.7083 2: 0.7941 3: 0.6491	1: 0.9520 2: 0.8025 3: 0.7935	1: 0.8302 2: 0.7983 3: 0.7213
SVM Radial	1: 0.7917 2: 0.7941 3: 0.4912	1: 0.8880 2: 0.7654 3: 0.8370	1: 0.8398 2: 0.7798 3: 0.6641
SVM Linear	1: 0.7083 2: 0.7794 3: 0.5439	1: 0.9360 2: 0.7160 3: 0.8152	1: 0.8222 2: 0.7477 3: 0.6795
SVM Poly	1: 0.7917 2: 0.7500 3: 0.5439	1: 0.8800 2: 0.8025 3: 0.8152	1: 0.8358 2: 0.7762 3: 0.6795
rda	1: 0.8333 2: 0.7353 3: 0.6140	1: 0.9360 2: 0.8148 3: 0.7717	1: 0.8847 2: 0.7751 3: 0.6929
knn	1: 0.8333 2: 0.7941 3: 0.5088	1: 0.8960 2: 0.7407 3: 0.8696	1: 0.8647 2: 0.7674 3: 0.6892

Table 8: Autres mesures

A partir des résultats obtenus (7 et 8) nous pouvons faire les mêmes constats que pour la méthode 2. La sensitivité est bien moins bonne pour la classe 3, ce qui nous permet d'en déduire que le rééquilibrage des classes est à l'origine de ce problème. On peut voir également que plus le modèle est complexe et représente des relations non-linéaires, plus le taux de sensitivité pour la classe 3 diminue. On remarque aussi au niveau du Balanced Accuracy que l'Accuracy la plus faible est de la classe 3. Inversement, la classe 1, sous-représentée au départ, possède de bons résultats sur l'ensemble des modèles testés. Nous n'avons pas pu observer si les modèles sur-apprenaient en faisant un test sur Maggle.

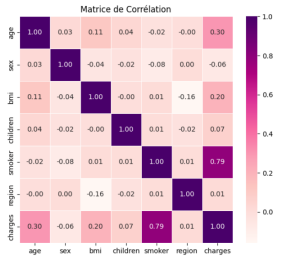
4 Jeu de données Kaggle

Le jeu de données que nous avons choisi peut se trouver via le lien suivant : lien.

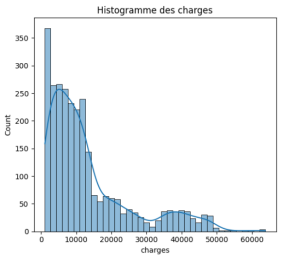
4.1 Analyse exploratoire

Ce dataset contient 2 700 lignes pour 7 variables prédictives dont une variable cible quantitative appelée "charges". Nous avons dû pour 3 variables catégorielles les convertir en numérique : 'sex', 'smoker' et 'region'. De plus, il n'y a aucune valeur manquante. Ce jeu de données sera utilisé pour faire de la régression. Afin d'appréhender ce jeu de données, nous avons tout d'abord voulu connaître si des variables étaient corrélées entre elles. Ainsi que l'apparence de la variable cible "charges".

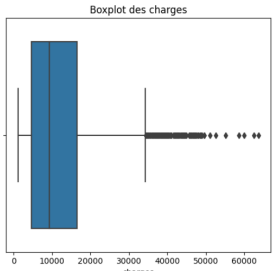
D'après la figure 6b, nous pouvons observer que les charges se trouvent sur une plage très grande ainsi que comme la figure 6c le montre, il y a beaucoup de valeurs extrêmes. De plus, le fait que la variable "charges" ne soit pas normalement distribuée va nous conduire en plus des métriques R^2 et RMSE à utiliser le MAE, Mean Absolute Error comme métrique non influençable par des valeurs extrêmes. Enfin, la figure 6a montre bien que 3 variables sont corrélées avec "charges" : "age", "bmi: body mass index" et "smoker".



(a) Matrice de corrélation entre les différentes variables



(b) Histogramme des charges (y cible)



(c) Boxplot des charges (y cible)

Cependant, la réduction de dimensions avec la RFE, en limitant les variables à 3, a entraîné une perte importante d'informations, empêchant le modèle de capturer les interactions complexes nécessaires pour de bonnes prédictions. De plus, l'utilisation de la PCA n'a pas amélioré les performances, probablement parce que cette méthode, adaptée aux relations linéaires, n'a pas saisi efficacement la complexité des données. Ces résultats montrent que conserver l'ensemble des variables, avec un simple scaling des données tout en supprimant les valeurs extrêmes, est suffisant pour de bons résultats. Ainsi en maintenant l'ensemble des interactions potentiellement utiles, nous pouvons améliorer les performances du modèle.

4.2 Modèles testés

Nous avons pu tester plusieurs modèles à partir d'un simple scaling, en supprimant ou non les valeurs extrêmes. La PCA dans la plupart des cas n'améliorait pas la performance. Le choix de supprimer les valeurs extrêmes au vu des schémas précédents se justifie et a été testé. Les meilleurs modèles avec leur résultats obtenus sont les suivants 9 et 10:

Modèles testés avec simple scaling	MAE	R^2	RMSE	MSE
KNN	556	0.958738	2516	6.333000e+06
Decision Tree	704	0.928175	3320	1.102382e+07
Random Forest	1314	0.951507	2728	7.442726e+06
Regression lineaire	4167	0.74	6318	3.9922479e+07
Elastic Net	4484	0.722312	6528	4.261984e+07

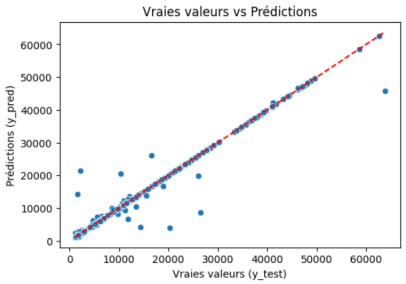
Table 9: Résultats obtenus avec un simple scaling

Modèles testés avec un scaling et suppression des valeurs extrêmes	MAE	R^2	RMSE	MSE
KNN	450	0.911747	2102	4.420435e+06
Decision Tree	640	0.818251	3017	9.103515e+06
Random Forest	1253	0.881897	2432	5.915573e+06
Regression lineaire	2455	0.61	4393	1.9302360e+07
Elastic Net	2730	0.593166	4514	2.037766e+07

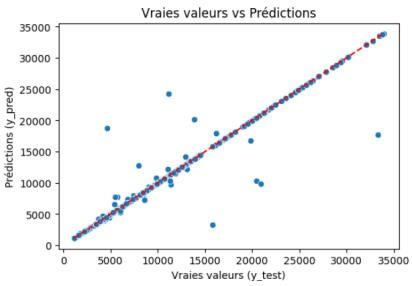
Table 10: Résultats obtenus avec un scaling et suppression des valeurs extrêmes

4.3 Analyse des résultats

Nous pouvons dire tout d'abord que la métrique MSE n'apporte pas grand chose comme information au vu des nombreuses valeurs extrêmes que la variable cible contient. De plus, la métrique MAE plus robuste que le R^2 donne un meilleur aspect de la performance. On peut noter qu'avec la suppression des valeurs extrêmes, le modèle s'ajuste mieux sur des données sans outliers. Du côté du RMSE, étant plus sensible aux grandes erreurs que le MAE, sa réduction indique que les valeurs extrêmes contribuaient de manière disproportionnée aux erreurs totales. En somme, KNN être le meilleur modèle que ce soit avec ou sans suppression des valeurs extrêmes voir les figures suivantes 7a et 7b.



(a) KNN avec scaling simple



(b) KNN avec scaling et suppression des valeurs extrêmes

5 Conclusion

Ces trois jeux de données nous ont permis d'appliquer de manière concrète des méthodes de Machine Learning vues en cours. Nous avons obtenu dans l'ensemble des résultats correctes. Les pistes d'améliorations seraient d'aborder d'une autre manière le jeu de données classification et de trouver des modèles qui ne sont pas influencés par le déséquilibre entre les classes.