Université Paris Dauphine

Sur une maladie dégénérative

Victoire de Salaberry, Mathieu Navarro & Maxime Berthier

2023 – 2024

Projet de Méthodes de régression et de classification - M2 ISF

Table des matières

1 Introduction

Le projet se concentre sur l'analyse d'un ensemble de données biomédicales de 42 individus atteints d'une maladie dégénérative précoce. Ces participants sont engagés dans un essai clinique de six mois évaluant un dispositif de télésurveillance pour suivre l'évolution de leurs symptômes.

L'objectif principal est de développer un modèle prédictif capable d'estimer le score clinique des patients en fonction du temps écoulé depuis le recrutement dans l'essai, ainsi que d'autres facteurs biologiques et médicaux.

Le modèle final sera le fruit d'une analyse approfondie des données, incluant une étude détaillée des variables et de leur relation ainsi qu'une sélection des variables les plus influentes. Le but est de parvenir à un modèle robuste et généralisable, capable de fournir des prédictions précises et fiables pour une variété de profils de patients.

2 Présentation des données et premières analyses

2.1 Description des variables

L'étude est réalisée à partir de 42 individus atteints d'une maladie dégénérative. Le nombre d'individus est assez faible mais reste correct pour une étude dans le domaine de la santé. En effet, dans ce domaine, il est souvent difficile d'avoir beaucoup d'observations. Les données disponibles mesurent l'évolution de la maladie des patients sur une durée de 6 mois, en observant le score clinique et 16 métriques en fonction du temps, qui sont détaillées ci-dessous (FIG. 1).

Variable	Description	Type	Support
sujet	Numéro du patient	Catégorielle	⟦1; 42⟧
age	Age du patient	Numérique	⟦36; 85⟧
genre	Sexe du patient	Catégorielle	{0, 1}
duree	Temps écoulé depuis le recrutement dans l'essai	Numérique	[-4; 138]
score	Score clinique	Numérique	[5; 38]
FF, FF.Abs, FF.RAP, FF.PPQ5, FF.DDP	5 mesures de la variation de la fréquence fondamentale (FF) de la voix	Numérique	[2.25e-06; 0.173]
AV, AV.dB, AV.APQ3, AVAPQ5, AV.APQ11, AV.DDA	6 mesures de la variation de l'amplitude de la voix (AV)	Numérique	[0.0019; 2.107]
BTC1, BTC2	2 mesures du rapport entre le bruit et les composantes tonales de la voix	Numérique	[0.0002; 38]
CDNL	Une mesure de complexité dynamique non linéaire	Numérique	[0.151; 0.97]
EFS	Exposant d'échelle fractale du signal	Numérique	[0.5; 0.87]
VFNL	Une mesure non linéaire de la variation de la fréquence fondamentale	Numérique	[0.02; 0.74]

FIGURE 1 – Les variables présentes dans le jeu de données

Le projet vise à développer un modèle prédictif du score clinique des patients, représentatif de l'évolution de la maladie. Pour se faire, nous avons commencé par étudier les différentes mesures relevées lors de l'étude. Une étude plus poussée a été mené dans le rapport expert.

Seulement deux variables sont catégorielles, `sujet` (correspondant à l'identifiant du patient) et `genre`. Nous avons pu constater que deux tiers de l'échantillon sont des hommes, le reste étant des femmes. La variable `age` nous indique que les patients sont âgés de 36 à 85 ans. Cependant, la plupart des individus ont entre 55 et 75 ans. Si la variable `age` a un impact sur le score, nos modèles ne seront probablement pas très performants sur des individus jeunes, car ils

n'auront pas (peu) appris sur ceux-ci. Concernant la variable `duree`, les données vont jusqu'à une valeur de 138 (nous ne connaissons pas l'unité, probablement en jours). Ensuite, nous pouvons observer 5 variables liées à la variation de la fréquence fondamentale de la voix. Les valeurs sont relativement faibles (inférieures à 0.173). Il y a 6 variables qui mesurent la variation de la voix, comprises entre 0.0019 et 2.107. Enfin, 2 variables sont associées au rapport entre le bruit et la tonalité de la voix.

Le score, notre variable cible, est compris entre 5 et 38, ce qui nous donne un ordre de grandeur pour les scores que nous souhaitons prédire. Une valeur éloignée de cette plage de données devrait nous interpeler. La tendance des scores est linéaire par morceaux et semblent s'inverser aux alentours du temps 100, avec un point d'inflexion à cette date (FIGURE 2). Sur cette-même figure, on voit que le genre n'a pas l'air d'avoir de l'influence sur le score. On étudie l'influence des variables explicatives sur la variable cible plus bas. La distribution de la variable cible peut être visualisée ci-dessous (FIGURE 3).

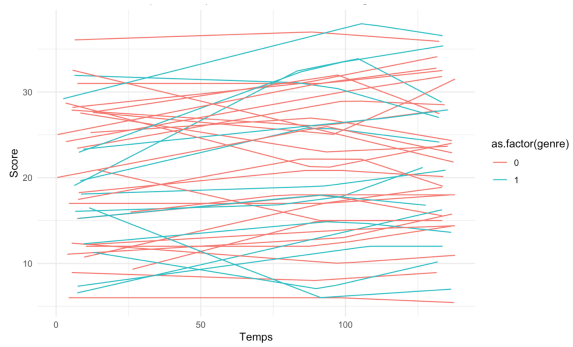

FIGURE 2 – Évolution du score en fonction de la durée et du genre

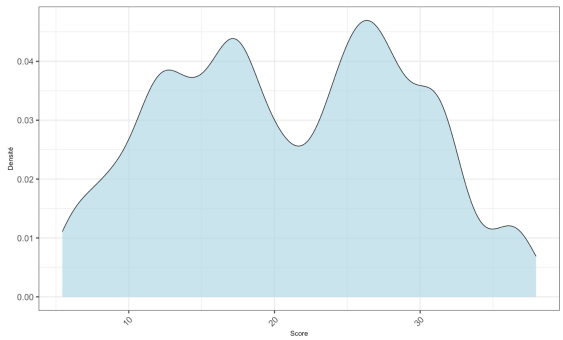

FIGURE 3 – Distribution de la variable cible `score`

2.2 Principaux retraitements

Le jeu de données ne présente pas de valeur manquante. Cependant, nous avons pu remarquer que tous les individus n'ont pas le même nombre de mesures. De plus, les intervalles de temps ainsi que les moments de la prise de mesures peuvent différer d'un individu à l'autre. Ces mesures peuvent être considérées comme des valeurs manquantes. Mais nous ne sommes pas en mesure de compléter le jeu de données. Ceci va donc peut-être réduire les performances de notre modèle. Nous avons par ailleurs constaté plusieurs observations pour certaines durées de certains patients. Nous supposons qu'il s'agit de mesures effectuées par plusieurs appareils en même temps.

Concernant les valeurs aberrantes, il y a des valeurs négatives pour la variable `duree`, ce qui n'est pas possible. N'étant pas nombreuses, nous avons décidé de supprimer ces observations.

2.3 Corrélations entre les variables

Pour avoir la meilleure analyse possible du jeu de données, il faut étudier les corrélations entre les variables. Si deux variables sont très corrélées, il faut faire une étude de leur relation et potentiellement retirer une des deux variables. En effet, garder ces deux variables pourraient créer un biais et rendre les coefficients du modèle instables. La matrice des corrélations sous forme de graphique *corrplot* (FIGURE 4) représente les coefficients de corrélation entre les variables explicatives par des couleurs.

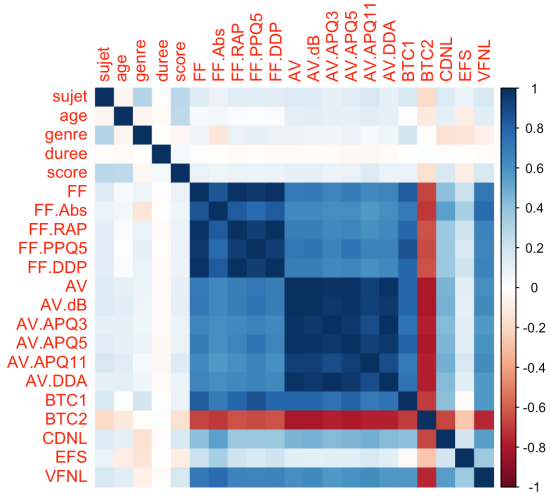

FIGURE 4 – Matrice de corrélations entre les variables explicatives

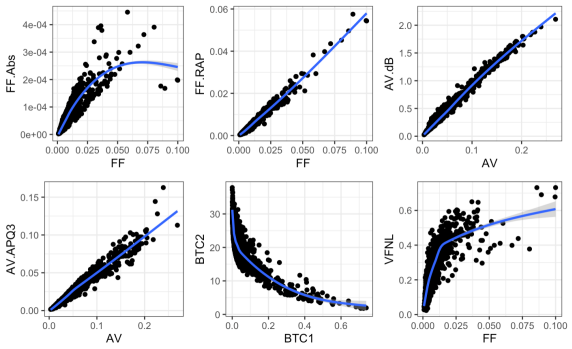

FIGURE 5 – Graphiques qui illustrent les relations entre différentes variables explicatives

On remarque alors que beaucoup de variables sont très corrélées (couleurs foncées), c'est-à-dire qu'elles ont un coefficient de corrélation proche de 1 en valeur absolue.

Graphiquement (FIGURE 5), on peut confirmer que les variables ayant un coefficient de corrélation élevé, ont un lien fort. C'est le cas par exemple, de toutes les mesures concernant le fréquence fondamentale (FF) et l'amplitude (AV) de la voix.

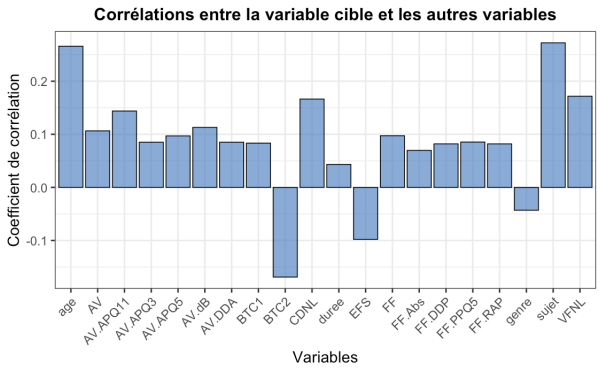

FIGURE 6

On regarde ensuite le coefficient de corrélation entre les variables explicatives et la variable cible (FIGURE 6) afin de répondre à notre objectif d'explication de la variable cible. Les variables qui semblent avoir le plus d'importance sont : `age`, `BTC2`, `CDNL`, `sujet` et `VFNL` bien que le coefficient reste relativement petit (plus proche de 0 que de 1). Les graphiques des relations entre ces variables et la variable cible sont consultables dans le rapport expert. Ils nous permettent d'observer des petites tendances, mais rien de très concluant. Nous approfondirons cette étude lors de la création des modèles.

3 Sampling

Afin d'étudier au mieux les performances de nos modèles, nous divisons le jeu de données mis à notre disposition en deux ensembles : un ensemble d'entrainement et un ensemble de validation.

L'ensemble d'entraînement est utilisé pour construire et entraîner le modèle. C'est sur cet ensemble que le modèle apprend les relations entre les variables explicatives et la variable cible. Une fois que le modèle est entraîné, l'ensemble de validation est utilisé pour évaluer sa performance. Cela permet de mesurer à quel point le modèle est capable de généraliser et de prédire sur de nouvelles données qu'il n'a pas encore vues.

Nous effectuons deux méthodes de découpages différentes sur lesquels nous entrainerons chacun

des modèles. La première méthode consiste en un découpage aléatoire, largement utilisé dans la pratique. Mais ce découpage n'est pas tout à fait juste dans notre contexte. En effet, avec un découpage aléatoire, notre modèle est exposé à des données provenant de différentes périodes, ce qui peut entraîner un apprentissage inadéquat des tendances temporelles. Or, notre objectif est d'entrainer nos modèles sur des données antérieures et les tester sur des données plus récentes. Le deuxième découpage que nous adoptons est basé sur une segmentation temporelle. Cette méthode de découpage semble mieux adaptée à nos données, car elle garantit que nos modèles sont entraînés sur des données antérieures à celles utilisées pour les tests.

Une fois que le modèle a été entraîné et optimisé grâce à l'ensemble d'entraînement et de validation, on évalue les performances finales du modèle sur un troisième ensemble : l'ensemble de test.

4 Métriques utilisées

Pour étudier nos prédictions nous choisissons la RMSE (racine de l'erreur quadratique moyenne) comme métrique de référence. Elle est facilement interprétable étant donnée que son ordre de grandeur est liée aux valeurs de la prédiction. De plus, les erreurs positives et négatives ne s'annulent pas mutuellement, et les erreurs importantes sont pénalisées de manière plus significative. Nous prendrons aussi en compte l'AIC et le BIC des modèles, qui mesurent respectivement l'information et la complexité du modèle. Ces mesures sont importantes car elles aident à choisir le modèle le plus approprié en tenant compte à la fois de sa capacité à expliquer les données et de sa complexité.

5 Modèle final

Comme mentionné précédemment, nous avons observé un point d'inflexion approximativement autour du temps 100 pour tous les individus. Ce point d'inflexion fait sens car l'évolution de certaines maladies peut se traduire par un changement brutal de l'état de santé. En réponse à cette observation, nous avons introduit une nouvelle variable, `inflx`, pour tenir compte de ce phénomène.

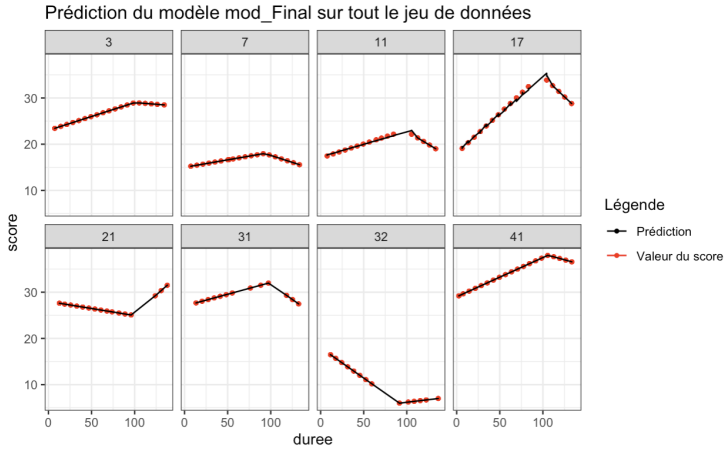

FIGURE 7

Le modèle mixte est particulièrement adapté à nos données car il combine à la fois des effets fixes qui sont communs à tous les individus, et des effets aléatoires, qui capturent la variabilité entre les différents individus du jeu de données, pour modéliser les relations entre les variables. Graphiquement, il semble être le plus pertinent. Les prédictions sur l'ensemble des données collent parfaitement aux observations (FIGURE 7). Le modèle mixte, entraîné à partir des variables `duree, FF, AV, BTC1, CDNL, EFS, VFNL` et variant selon l'interaction entre `inflx` et `sujet`, s'est avéré être le meilleur modèle.

Pour choisir le modèle final, nous nous sommes d'abord fiés au deuxième découpage, ce découpage étant le plus pertinent. Sur ce découpage, ce modèle présente le meilleur équilibre entre RMSE, AIC et BIC. Sa RMSE est l'une des plus petites parmi celles de tous les autres modèles testés (voir section 6.5). En termes de AIC et BIC, ce modèle est le meilleur. C'est également l'un des meilleurs modèles sur le premier découpage.

Les sections suivantes détaillent rapidement les autres modèles que nous avons testés. Ces modèles ont contribué à orienter notre réflexion vers la création du modèle final.

6 Autres modèles testés

Les modèles présentés dans cette section ont été entrainé sur chacun des deux découpages (section 3). De plus, pour chacune des méthodes, nous avons créé un modèle avec toutes les variables explicatives puis un modèle en sélectionnant certaines variables. En effet, nous avions remarqué que beaucoup de variables étaient très corrélées (section 2.3), et donc en supprimer peut améliorer les performances du modèle. C'est le cas notament de notre modèle final qui ne garde qu'une seule mesure de la fréquence fondamentale (`FF`) et une seule de l'amplitude de la voix (`AV`) par exemple.

6.1 Une première approche : les modèles linéaires

Les modèles linéaires sont une classe de modèles statistiques qui supposent une relation linéaire entre la variable réponse et les variables explicatives. Nous avons pu constater que le modèle linéaire sous-estime ou sur-estime les données des individus. En effet, bien que l'allure générale du score est la même pour tous les indivius (en forme de « chapeau pointu »), l'ordonnée à l'origine et la pente de la fonction varient d'un sujet à l'autre. Le modèle linéaire simple n'est donc pas le plus adapté pour des données complexes telles que les nôtres. Il est nécessaire d'élargir notre modèle afin de tenir compte de cette variabilité interindividuelle.

6.2 Un modèle plus adpaté : le modèle mixte

Nous avons alors ajusté des modèles mixtes, plus adaptés à notre jeu de données. Les modèles semblent un peu mieux *fitter* avec les données mais ils ne sont toujours pas convaincants. Ils ne parviennent toujours pas à capturer le changement brutal du coefficient directeur (c'est-à-dire le point d'inflexion).

6.3 Encore mieux : notre modèle final

Pour tenir compte de cette observation, nous avons créé des modèles mixtes en intégrant une nouvelle variable : `inflx`. Nous avons exploré diverses combinaisons de variables et créé des modèles pour chaque configuration, pour finalement retenir le meilleur, décrit dans la section 5.

6.4 Un dernier modèle : le XGBoost

Pour compléter notre étude, nous avons investigué un modèle XGboost. En effet, XGboost est reconnu pour sa capacité à gérer des ensembles de données complexes et à produire des prédictions précises, ce qui en fait un choix intéressant dans notre contexte. La création de modèles XGBoost a enrichi notre analyse en fournissant des performances très prometteuses, avec notamment, de petites RSME.

6.5 Comparaison des modèles

Modèle	1er découpage		2ème découpage	
	Avec toutes les var.	Sélection de var.	Avec toutes les var.	Sélection de var.
Modèle linéaire	1.94	1.95	//	2.47
Modèle mixte	1.76	1.84	//	2.46
Modèle mixte avec `inflx`	0.24	**0.07**	0.45	**0.46**
XGboost	0.30	//	//	1.49

Le tableau ci-dessous compare les RMSE des différents modèles. Nous en retiendrons plusieurs choses :

- Le modèle linéaire classique, même après sélection de variables, n'est pas adapté puisqu'il obtient une RMSE élevée, proche de 2 sur le premier découpage classique, et de 2.5 pour le découpage temporel.
- L'utilisation d'un modèle à effet mixte améliore les performances puisqu'il s'adapte aux spécificités de chaque sujet, mais les RMSE sont tout de même élevées. Il n'arrive pas à appréhender l'inflexion de la trajectoire des scores.
- Notre modèle final, un modèle linéaire à effet mixte qui prend en compte le moment de l'inflexion, a une très bonne RMSE (en rouge). Sur le 2^{eme} découpage, le modèle mixte avec toutes les variables à certes une plus petite RMSE mais des AIC et BIC plus élevés. Nous préférons conserver un modèle avec des AIC et BIC plus faibles, tout en gardant une RMSE peu élevée (bien que plus élevée que le modèle avec toutes les variables mais l'écart est faible). C'est le modèle qui a le meilleur équilibre entre RMSE, AIC et BIC.
- Notre modèle final est largement meilleur que le modèle construit avec XGboost, une méthode pourtant particulièrement performante !

7 Conclusion

À travers l'utilisation de modèles prédictifs, nous avons cherché à mieux comprendre l'évolution de la maladie pour améliorer le suivi des patients. Les performances des modèles construits, notamment celles du modèle mixte avec la variable `inflx`, sont très prometteuses.

Cependant, l'une des principales limites réside dans la taille restreinte de l'échantillon de données, ce qui pourrait limiter la généralisation des résultats à une population plus large. De plus, un ensemble de données avec des mesures prises simultanément et à des intervalles réguliers aurait probablement permis une modélisation plus précise et fiable. Enfin, la nature des données biomédicales peut introduire des biais et des sources d'incertitude, notamment en raison de la variabilité interindividuelle et de la complexité du suivi d'une maladie. Il est également important de noter que l'identification d'une inflexion dans l'évolution de la maladie, et l'utilisation de cette observation dans la construction du modèle, reposent sur une hypothèse simplificatrice. En effet, supposer que l'évolution de la maladie ne contiendra qu'un changement brutal peut être erroné car cette évolution peut être plus complexe et comporter plusieurs phases distinctes. Il est nécessaire de rester prudent quant à cette hypothèse et d'envisager d'autres modèles ou analyses pour confirmer ou infirmer cette observation.

Finalement, une extension intéressante consisterait à inclure un suivi plus regulier et plus long des patients afin de mieux comprendre l'évolution de la maladie. De plus, l'intégration de données (mesures) supplémentaires ainsi que d'individus, pourraient enrichir l'analyse et améliorer la précision des prédictions. Enfin, il serait intéressant de construire des modèles à partir d'autres méthodes, par exemple des modèles à seuil.