1. **Protocole**

**I.1 Validation croisée**

Comme il a été mentionné en introduction, une habitude à prendre lors de toute analyse de données est de séparer les données en plusieurs sous échantillons pour éviter les problèmes de surestimation des capacités du modèle.

On appelle validation croisée la technique consistant à ajuster un modèle prédictif sur un échantillon d’apprentissage et à valider ce modèle sur un échantillon test. Ces échantillons peuvent provenir du même jeu de données auquel cas il est coutume que l’échantillon d’apprentissage représente entre 60% et 80% des données et que l’échantillon test représente 20% à 40%. Il est également possible, si l’on dispose de plusieurs jeux de données différents pour le sujet d’étude, de prendre un jeu de données comme échantillon d’apprentissage et de valider le modèle sur un deuxième jeu de données. Nous avons choisi dans nos études de cas de réaliser une validation croisée à partir d’un seul jeu de données en le décomposant en un échantillon d’apprentissage représentant 80% du jeu de données et en un échantillon test représentant les 20% restant, et ce de manière aléatoire.

**I.2 Sélection des modèles**

Une fois le jeu de données séparé en deux échantillons, vient le moment de construire différents modèles prédictifs. Cependant, il n’est pas toujours évident de savoir quelles variables garder, quelles sont celles qui apportent le plus d’information, qui discriminent le mieux les groupes d’individus, etc… C’est pourquoi on s’appuie sur différents indicateurs, en plus de ceux implémentés par défaut dans les logiciels. Nous en évoquons 3 ici :

I.2.1 Les critères AIC et BIC

Les critères AIC et BIC permettent de comparer des modèles qui ne sont pas forcément emboîtés. C’est ce critère qui est utilisé dans les fonctions bestglm ou step du logiciel Rstudio.

* AIC (Akaike Information criterion) : AIC = −2ll(β\_chap) + 2p =2p+n\*log(SCR/n) (n=nb d’indiv)

Comme on cherche le modèle se rapprochant le plus de la réalité, on souhaite avoir une valeur de SCR (somme des carrées des résidus) la plus faible possible, c’est pourquoi on retiendra le modèle minimisant l’AIC.

* BIC (Bayesian information criterion) BIC = −2ll(β\_chap) + ln(n)p = ln(n)p+ n\*log(SCR/n)

C’est une version modifiée de l’AIC, motivée par l’utilisation d’un a priori sur le paramètre β.

A cause du facteur log(n) dans la pénalité, le critère BIC aura tendance à favoriser les modèles avec moins de paramètres que le critère AIC.

*Remarques : Si n tend vers l’infini, la probabilité que le BIC choisisse le « vrai » modèle tend vers 1, ce qui est faux pour l’AIC.*

*Le critère AIC choisi le modèle maximisant la vraisemblance de futures données et réalisant le meilleur compromis biais-variance.*

*L’AIC est un critère prédictif tandis que le BIC est un critère explicatif*

*Pour n fini, le critère BIC a tendance à choisir des modèles trop simples en raison de sa forte pénalisation*

* cf cours ML3-CS1 de Mr Labatte pour l’explication plus détaillée du -2ll(β\_chap). (que veut dire β\_chap)

I.2.2 La déviance

Nous verrons plus tard dans ce rapport que les paramètres des différents modèles sont le plus souvent estimés par la méthode du maximum de vraisemblance.

Pour savoir quels modèles garder, il est donc courant d’utiliser le critère de la déviance.

En effet, la déviance est égale à -2log-vraisemblance donc le modèle maximisant la vraisemblance est celui minimisant la déviance. On utilise plutôt cet indicateur car plus simple à calculer que les expressions du maximum de vraisemblance.

Il existe des fonctions dans Rstudio rendant le /les modèles minimisant la déviance, il n’est donc pas nécessaire de créer les modèles au préalable et de les comparer entre eux, le logiciel fait ce travail à notre place.

* Je ne suis pas trop sûre de ce que je dis ici, je pense qu’il manque des infos, à compléter

I.2.3 Le lambda de Wilks

Cet indicateur est propre à l’analyse discriminante et n’est pas utilisé en régression logistique.

Le Lambda de Wilks est souvent utilisé dans les logiciels comme critère pour ne garder que les variables apportant de l’information sur l’appartenance ou non d’un individu à un groupe.

Le Lambda de Wilks est une approche paramétrique permettant de tester si plusieurs variables continues distinctes  sont liées à une variable qualitative Y à K ≥ 2 groupes, lorsqu’elles sont considérées avec leurs différentes interactions multivariées.

Les hypothèses d’utilisation de ce test sont:  suivent une loi normale et leur matrice de covariance respective sont égales (homoscédasticité).

La statistique du test du Lambda de Wilks se définie de la manière suivante :

Où SCR est la matrice de variance-covariance intragroupe et SCT la matrice de variance-covariance globale.

Cette statistique de test suit une loi de Wilks à  degrés de liberté et l’hypothèse  est : « Indépendance entre  et  ».

Une variable a un bon pouvoir discriminant si la dispersion intra-groupe est faible et si la dispersion intergroupe est forte. Donc plus le Lambda de Wilks sera faible, plus la variable considérée est discriminante.

C’est ce critère qu’utilise la commande « greedy.wilks » de Rstudio que nous avons utilisée pour trouver les variables les plus discriminantes dans notre jeu de données et ainsi se focaliser sur un nombre de modèles plus réduit.

**I.3 Qualité du modèle**

I.3.1 Matrices de confusion

L’analyse factorielle discriminante probabiliste et la régression logistique sont utilisées pour affecter des scores à des individus. Mais donner un score à un individu sans contexte d’étude est assez absurde. C’est pourquoi, il faut évaluer le modèle choisi pour savoir si celui-ci permet une bonne prédiction des groupes d’appartenance.

Il est coutume de tester la qualité du modèle grâce aux matrices de confusions qui donnent le taux de bon et mauvais classement des individus dans chaque groupe. Nous construisons donc souvent plusieurs modèles et évaluons leurs capacités prédictives grâce à ces matrices.

Dans le cas de deux groupes les matrices de confusions se représentent souvent de la façon suivante :

TN = vrai négatif FN = faux négatif

FP = faux positif TP = vrai positif

0 = négatif

1 = positif

|  |  |  |
| --- | --- | --- |
|  | Y=0 | Y=1 |
| Y\_pred = 0 | TN | FN |
| Y\_pred = 1 | FP | TP |

En fonction de ce que l’on cherche à faire grâce au scoring, on préfèrera retenir le modèle minimisant les faux négatifs ou les faux positifs, ou encore le modèle maximisant les vrais positifs ou vrai négatif.

I.3.2 Courbes Roc et histogrammes des scores

Une fois notre modèle choisi grâce aux comparaisons des différentes matrices de confusion, il est possible de visualiser graphiquement la qualité globale de ce modèle grâce à une courbe ROC.

Une courbe ROC (Receiver Operating Characteristic) est un graphique représentant les performances d’un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs :

* Le taux de vrais positifs (TVP) (sensibilité) est défini comme : T V P = V P / (V P + F N)
* Le taux de faux positifs (TFP) (spécificité) est défini comme : T F P = V N / (V N + F P)

La courbe ROC résume les performances de toutes les règles de classement que l’on peut obtenir en faisant varier le seuil de décision. Les termes "positifs" et "négatifs" dépendent des conventions que l’on aura choisies au préalable.

La sensibilité se définit comme le pourcentage de vrais positifs : 1 − β. D’un point de vue médical cela veut dire être testé positif à un test détectant la présence de maladie quand on est bien malade.

La spécificité se définit quant à elle comme le pourcentage de vrais négatifs : 1−α. D’un point de vue médical cela signifie être testé négatif à un test détectant la présence de maladie, quand on est bien sain.

Aperçu d’une courbe ROC : mettre un graphe d’exemple.

Plus la courbe se rapproche du coin supérieur gauche, plus le modèle est considéré comme bon.

Cependant, la lecture du seuil optimal sur cette représentation est assez difficile, c’est pourquoi on utilise d’autres graphiques pour déterminer avec plus de précision ce seuil.

I.3.3 Histogrammes de score

Le seuil utilisé dans l’algorithme de recherche du meilleur modèle (pour la problématique considérée) est par défaut fixé à 0.5 (dans le cas de 2 groupes) pour décider du groupe d’appartenance. Or toute l’essence du scoring est justement de trouver le seuil qui donnera un risque d’erreur le plus faible possible, tout en prenant les contraintes de coût financier en compte.

C’est pourquoi il est courant de représenter les histogrammes de score pour déterminer le seuil optimal.

On trace les histogrammes des scores des individus en fonction de leur vrai groupe d’appartenance.

Dans le cas où les 2 histogrammes sont disjoints, alors il est possible de trouver un seuil qui annulera l’erreur prise, mais ce cas est plutôt rare. Les histogrammes sont toujours plus ont moins superposés et c’est donc le travail de l’analyste de choisir le seuil qui minimisera le plus possible le taux d’erreur pris, tout en prenant en compte encore une fois toutes les contraintes financières ou matérielles.

* Graphique d’exemple