Algorithmes d'apprentissage et modèles statistiques: Un exemple de régression logistique régularisée et de validation croisée pour prédire le décrochage scolaire

Eric Lacourse, Charles-Édouard Giguère & Véronique Dupéré

Référence suggérée :

Lacourse, E., Giguère, C.E., & Dupéré, V. (sous presse). Algorithmes d'apprentissage et modèles statistiques: Un exemple de régression logistique régularisée et de validation croisée pour prédire le décrochage scolaire. Dans M. Corbière & N. Larivière (Eds.), Méthodes qualitatives, quantitatives et mixtes dans la recherche en sciences humaines, sociales et de la santé, 2^e édition. Québec, QC : PUQ.

Forces

- 1) La régularisation et la validation croisée permettent de sélectionner le meilleur modèle statistique/algorithmique pour des problèmes de recherche en haute dimension (jusqu'à des millions de variables, récemment) et où le nombre de variables/dimensions p peut largement dépasser le nombre d'unité dans l'échantillon n .
- 2) Ces techniques peuvent s'appliquer à presque tous les modèles statistiques/algorithmes en permettant d'estimer les paramètres de ces derniers, alors qu'il serait difficile ou impossible de le faire avec des modèles multivariés « classiques ».

- 3) Ce sont des outils robustes de sélection de modèle qui peuvent être parfois plus performant que les indices d'ajustement proposés dans les analyses multivariées classiques, favorisant potentiellement une meilleure reproductibilité de la recherche scientifique.

Limites

- 1) Comme les techniques ont été développées pour optimiser la justesse de la prédiction, souvent au détriment de l'inférence, elles sont plutôt conçues comme des outils prédictifs et exploratoires (p. ex., forage de données).
- 2) L'interprétation de certains algorithmes est souvent peu transparente particulièrement lorsque les résultats sont issus d'algorithmes qui modélisent automatiquement les relations non-linéaires et les interactions entre les variables et qui sont souvent qualifiés de « boîtes noires ».
- 3) Ces techniques n'ont pas initialement été conçues pour accommoder des données qui contiennent de l'erreur de mesure, qui ont une structure nichée, hiérarchique ou des données qui sont manquantes.

Introduction

Au cours des dernières années, plusieurs nouveaux termes associés à l'intelligence artificielle ont fait leur apparition dans le langage populaire et scientifique comme les concepts de données massives (*big data*), de forage de données (*data mining*), d'apprentissage automatique et statistique. De plus en plus de chercheurs en sciences sociales et de la santé veulent exploiter de nouvelles données massives biologiques et comportementales provenant d'applications mobiles ou des médias sociaux. D'autres veulent explorer des bases de données déjà existantes, mais en utilisant des techniques statistiques qui permettent de dépasser les limites des analyses multivariées « classiques », principalement les règles stipulant, d'une part, qu'il faille un minimum de 5 à 10 sujets par variable et, d'autre part, qu'il faille présélectionner des variables qui sont faiblement ou modérément corrélées entre elles avant de faire des analyses. Malgré l'engouement autour de ces méthodes qui font miroiter la possibilité de découvrir de nouveaux patrons dans les données et de faire des prédictions plus justes, l'utilisation des approches algorithmiques demeurent encore limitées. Plusieurs facteurs peuvent expliquer l'utilisation restreinte de ces approches à ce jour. Premièrement, une grande proportion de chercheurs fonde traditionnellement leurs analyses quantitatives sur des hypothèses ou modèles spécifiques à tester afin de valider ou confirmer des théories préexistantes. Deuxièmement, ceux qui voudraient adopter des approches plus exploratoires se trouvent souvent confrontés à un manque de textes de référence accessibles présentant des approches rigoureuses, les textes existants portant principalement sur la régression exploratoire pas à pas (*stepwise*) qui donne souvent des résultats très peu reproductibles (McNeish, 2015). Troisièmement, les chercheurs ont souvent l'impression qu'ils doivent posséder des données massives pour utiliser les

techniques de forage de données ou que leurs données sont inadéquates pour ce genre de techniques. Ce chapitre vise à démystifier certaines de ces croyances et à démontrer que l'on peut apprendre de différentes bases de données peu importe leur taille en utilisant des outils et des idées, souvent très similaires, provenant autant de la modélisation statistique que de l'apprentissage automatique/statistique. Plusieurs revues de la littérature ont été rédigées récemment couvrant plusieurs disciplines, de la sociologie à la psychologie clinique et à la psychiatrie (Bzdok, Altman, & Krzywinski, 2018; Dwyer, Falkai, & Koutsouleris, 2018; Molina & Garip, 2019; Shatte, Hutchinson, Teague, in press).

Objectifs généraux

Un premier objectif de ce chapitre est de permettre au lecteur d'établir des liens entre des approches à l'analyse de données qui peuvent paraître très différentes dues à leur épistémologie et à leur lexique disciplinaire. Autrement dit, il s'agit de contraster la modélisation statistique « hypothético-déductive » plutôt centrée sur l'inférence, de celle utilisant des approches dites « algorithmiques », plutôt inductive, et dont l'objectif est de maximiser la capacité et la justesse prédictive du modèle (*accuracy*) tout en sélectionnant un nombre optimal et parcimonieux (*sparse*) de variables. Nous allons également aborder des concepts comme le compromis entre le biais et la variance dans la sélection des techniques de régularisation qui sont des concepts centraux afin de réfléchir à la généralisation des modèles/algorithmes. Les techniques de régularisation qui seront vues sont la régression crête (ridge) ou régularisation L₂ (Hoerl, & Kennard, 1970), la régression LASSO (Least Absolute Shrinkage and Selection Operator) ou régularisation

L₁ (Tibshirani, 1996) et la régression *elastic net* (Zhou & Hastie, 2005) qui combine la régularisation L₂ et L₁. Nous allons également aborder le concept de validation croisée qui consiste à diviser aléatoirement l'échantillon initial en données « d'entraînement » et en données « test ». La technique de validation croisée à k-plis (*k-fold*) sera aussi abordée comme outil permettant de sélectionner les différents paramètres d'ajustement/hyperparamètres pour la régularisation et d'ultimement choisir le « meilleur » modèle prédictif. Cette technique peut être parfois préférable aux indices d'ajustements que l'on retrouve plus traditionnellement en analyse multivariées, comme les modèles d'équations structurelles.

Comme deuxième objectif, ce chapitre vise à démontrer, par le biais d'un exemple, la grande similarité entre les approches basées sur des modèles « statistiques » et celles, plus exploratoires qui ne présupposent pas de modèles explicites au préalable. La régularisation et la validation croisée seront présentées comme des outils permettant la sélection de modèles, c'est-à-dire celui qui a la meilleure capacité prédictive sur de nouvelles données et celui qui est également le plus parcimonieux.

1. Éléments théoriques centraux aux algorithmes d'apprentissage automatique ou statistique

L'apprentissage automatique ou statistique est une sous-discipline de la statistique et de l'informatique qui regroupe un ensemble de techniques permettant de modéliser des données multidimensionnelles où le nombre de variables, d'observations et de paramètres peuvent théoriquement tendre vers l'infini. L'engouement des scientifiques pour ces différentes techniques est relativement récent et provient de développements parallèles dans des disciplines variées comme la statistique, l'informatique, les mathématiques et

l'ingénierie, dont les réseaux de neurones et l'apprentissage profond en sont de bons exemples (Schmidhuber, 2015). La capacité de calcul des ordinateurs a aussi joué un rôle important permettant la mise en production de ces modèles à grande échelle et de les appliquer à la résolution de problèmes très concrets dans la vie quotidienne, comme le développement de moteurs de recherche pour l'internet, la reconnaissance d'images et du son ou le traitement naturel du langage, par le biais d'agents conversationnels (p. ex., Siri et Alexa).

On peut concevoir les récentes avancées des algorithmes d'apprentissage automatique comme des extensions des modèles statistiques traditionnels pour des problèmes à haute dimensionnalité (avec infiniment beaucoup de variables) et où le nombre d'unité d'échantillonnage peut également atteindre les milliards (p. ex. le nombre d'utilisateurs Facebook). Au-delà des données massives, dans certaines situations, le nombre de variables p peut même dépasser largement la taille de l'échantillon n et c'est à ce moment que la régularisation peut devenir particulièrement utile. Par exemple, dans un domaine de recherche comme la génétique moléculaire, on se retrouve souvent avec ce genre de problématique, où le nombre de marqueurs génétiques, souvent fortement corrélés entre eux, peut largement dépasser le nombre de participants recrutés dans une étude. D'une façon plus générale, la plupart des bases de données en sciences sociales et de la santé ont souvent beaucoup plus de variables que de sujets et pourraient bénéficier des approches en apprentissage automatique présentées dans ce chapitre pour développer de nouvelles connaissances scientifiques.

Une des particularités des modèles d'apprentissage statistique/automatique est le fait qu'ils mettent une emphase particulière sur la justesse de la prédiction en lien avec une

combinaison potentiellement non-linéaire et/ou non-additive de variables (en incluant des termes d'interactions) d'une grande quantité de variables. Mettre l'accent sur les tests d'hypothèses et l'inférence devient alors secondaire. En sciences sociales et de la santé, ils sont traditionnellement utilisés pour faire des analyses exploratoires, mais certaines pistes de recherche combinant à la fois les intérêts pour la prédiction et pour l'explication des comportements en psychologie ont récemment été proposées (Yarkoni & Westfall, 2017; Jacobucci, Brandmaier, & Kievit, 2019). Comme on le voit fréquemment dans certains domaines comme l'assurance et l'attribution du crédit, on peut utiliser des algorithmes pour développer des instruments de mesure automatisés pour mesurer, d'une façon exhaustive, les risques et vulnérabilités d'une personne face à certains événements négatifs futurs. La psychométrie est un autre domaine qui utilise systématiquement des modèles statistiques/algorithmes pour construire des instruments de mesures de différents construits théoriques, comme les dimensions de la personnalité ou des mesures de la dépression et/ou l'anxiété. Chaque algorithme ou modèle statistique possède des forces et des faiblesses et la sélection des modèles/algorithmes dépend souvent du type de données à analyser (p. ex., nominal, ordinal ou intervalle/ratio) ou de la façon dont les données sont colligées (p. ex., devis multiniveaux, longitudinaux ou expérimentaux). Mais surtout, la sélection du modèle/algorithme dépend de la nature de l'objectif premier de l'analyse, selon que l'accent est placé sur l'explication des processus causaux liés à un phénomène ou bien uniquement à sa prédiction. En général, il est souhaitable de comparer différents algorithmes entre eux pour pouvoir bien évaluer leur capacité prédictive respective. Aujourd'hui, il est même commun d'utiliser plusieurs algorithmes et de faire la moyenne de leurs prédictions, ce que l'on appelle l'apprentissage par ensemble (*ensemble*

learning). Dans la panoplie d’algorithmes qui est offerte, il existe des algorithmes moins flexibles, comme la régression linéaire multiple et des algorithmes plus flexibles comme la forêt d’arbres aléatoire ou bien les « réseaux de neurones artificiels multicouches » qui peuvent automatiquement modéliser la non-linéarité et les interactions entre les variables. La capacité interprétative des paramètres estimés par un modèle/algorithme est généralement inversement proportionnelle à la flexibilité de ce dernier. On parle souvent de « boîte noire » pour décrire ce phénomène.

Plusieurs grandes familles algorithmiques ont été développées, au cours des trente dernières années, afin d’estimer plus ou moins « automatiquement » des fonctions complexes, linéaires ou non-linéaires, pouvant inclure des termes multiplicatifs (effets d’interaction) tout en permettant d’obtenir des prédictions justes et une convergence ultra-rapide en haute dimensionnalité. Ces familles d’algorithmes visent toutes à approximer une fonction liant des variables indépendantes et une ou plusieurs variables dépendantes. Pour en nommer quelques-unes, il existe les régressions régularisées (i.e., crête (*ridge*), LASSO et *elastic net*, les forêts d’arbres aléatoires, ainsi que les réseaux de neurones artificiels multicouches, nommer aussi « apprentissage profond » (*deep learning*) qui sont devenus très populaires depuis quelques années.

Deux livres proposant un survol complet des différents algorithmes d’apprentissage statistique sont fréquemment utilisés afin d’analyser des données observationnelles transversales (non-nichées et à un seul temps de mesures). Le livre intitulé *Elements of Statistical Learning* (Hastie, Tibshirani, & Friedman, 2009) s’adresse aux personnes ayant une formation avancée en statistiques, alors que le livre « *An Introduction to Statistical Learning* » (James, Witten, Hastie, & Tibshirani, 2013), du même groupe

d’auteurs, aborde des thèmes similaires, mais d’une manière moins technique et plus accessible pour un auditoire d’étudiants gradués et de chercheurs en sciences sociales et de la santé. De plus, ce dernier livre inclut des laboratoires pratiques que l’on peut réaliser avec le logiciel libre R. Les techniques présentées dans ce chapitre peuvent aussi être réalisées avec les logiciels commerciaux comme SPSS avec la procédure CATREG, avec le logiciel SAS en utilisant PROC GLMSELECT, et aussi avec le « lassopack » de STATA.

1.1 Apprentissage automatique et modélisation statistique : un jargon à démystifier

La nouvelle terminologie, développée dans le champ d’étude de l’intelligence artificielle/apprentissage automatique, peut souvent porter à confusion, même pour les utilisateurs aguerris de modèles statistiques avancés. Par exemple, lorsqu’en statistique nous estimons des paramètres d’un modèle de régression en utilisant des données d’un échantillon, en apprentissage automatique, on parlera alors de données ou d’échantillon « d’entraînement » (*training data*). Également, au lieu d’estimer un modèle statistique, on utilisera plutôt le vocable « entraîner une machine, un algorithme ou développer une architecture ». De plus, les concepts de variables indépendantes ou de covariables que l’on retrouve communément en statistique, se voient décrits comme des « caractéristiques ou entrées » (*features* ou *inputs*) d’un modèle. Lorsque la variable dépendante est catégorielle, on définira alors la tâche à exécuter par l’ordinateur comme en étant une de classification. Finalement, les problèmes statistiques classiques de prédiction sont toujours conçus en termes « d’apprentissage », comme si une machine apprenait les paramètres d’un modèle à partir des données qu’on lui présente. On parlera

« d'apprentissage supervisé » pour des tâches similaires à la régression où l'on veut prédire une variable donnée alors que « l'apprentissage non-supervisé » fera plutôt référence à des problèmes statistiques plus traditionnellement liés à l'analyse factorielle ou de cluster présumant la présence de variables latentes, continues ou catégorielles (c'est-à-dire des variables non-observées et construites à partir du modèle). Il existe également l'apprentissage par renforcement qui permet de mieux comprendre comment un acteur « machine » « apprend » par l'influence des renforcements dans son environnement. Les modèles statistiques longitudinaux qui mesurent les comportements d'une personne et les effets de covariables variant dans le temps peuvent être conçus comme des exemples « statistiques » de l'apprentissage par renforcement.

L'apprentissage dit « supervisé » implique de développer un modèle statistique afin d'expliquer les liens existants entre des entrées (*inputs*) et une ou plusieurs sorties (*outputs*) dont on connaît les valeurs ou les étiquettes. Dans ce contexte, la régression linéaire multiple ou la régression logistique sont des exemples dits « classiques » d'algorithmes d'apprentissage supervisé. Dans ce chapitre, nous allons nous arrêter à l'utilisation d'un algorithme relativement flexible et populaire d'apprentissage supervisé, soit la régression logistique régularisée. Il peut être utile de connaître les techniques de modélisation statistique utilisant la régression linéaire et logistique, tel que décrit dans le **chapitre X du présent livre**. Il est important de comprendre que l'ensemble des concepts et des techniques qui sont couverts dans ce chapitre sont généralisables à presque tous les algorithmes d'apprentissage automatique ou aux modèles statistiques communément utilisés en sciences, comme les équations structurelles (voir chapitre X) ou les modèles multiniveaux (voir chapitre X). Il faut comprendre que seuls les objectifs de la

modélisation et la flexibilité de modèle/algorithme peuvent changer, particulièrement d'avoir un nombre de paramètres qui tend vers l'infini.

1.2 Quel est l'objectif principal de l'apprentissage automatique ou statistique?

Le but ultime de l'apprentissage automatique est de déterminer une fonction entre une variable quantitative Y (*output* ou variable dépendante) et p différentes variables prédictives $X_1, X_2, X_3, \dots, X_p$ (*input* ou variable indépendante). Il est présumé qu'il existe une association entre Y et $X = (X_1, X_2, X_3, \dots, X_p)$ déterminé par une fonction qui suit une forme très générale et qui peut être linéaire ou non-linéaire (incluant les effets d'interaction)

$$y = f(X) + \epsilon \quad (1)$$

Dans ce cas, f représente une fonction fixe, mais inconnue, liant les $X_1, \dots, X_p$ à Y et incluant un terme d'erreur e qui se veut aléatoire, indépendant de X et ayant une moyenne de 0. Tout dépendant de la distribution de Y , e suivra une distribution normale ou une autre forme. L'importance que l'on accorde à la distribution de l'erreur varie selon que l'on veut pouvoir faire de l'inférence sur les paramètres ou non. En statistique classique, ces postulats ont une grande importance car la validité des tests statistiques d'inférences dépend généralement de la normalité et de l'indépendance des termes d'erreur. En apprentissage automatique, l'interprétation des paramètres est de moindre importance et l'accent est mis sur la capacité et la justesse prédictive du modèle dans son ensemble, d'où le surnom de « *boîte noire* » qui est souvent associé à certains algorithmes permettant d'approximer des fonctions non-linéaires.

1.2.1 Pourquoi s'intéresser à une fonction f ?

L'approximation d'une fonction permet principalement de faire deux choses : soit faire des prédictions, soit faire des inférences qui peuvent être causales ou encore les deux à la fois.

Un premier objectif de la modélisation est donc de pouvoir prédire un output Y , à partir de caractéristiques X qui sont facilement mesurables et antécédentes ou concomitantes à notre output. Comme le terme d'erreur prend la valeur de 0 dans l'équation précédente, on peut prédire Y en suivant cette équation

$$\hat{y} = \hat{f}(X), \quad (2)$$

où $\hat{f}$ représente une estimation de f et où $\hat{y}$ représente les valeurs « prédites » ou « attendues » (*expected*) de notre modèle en lien avec Y . Dans plusieurs situations $\hat{f}$ est considéré comme une « *boîte noire* » et les paramètres suscitent peu d'intérêt en tant que tel, puisque c'est la justesse de la prédiction qui importe le plus. Par exemple, dans une tâche de classification d'images où X représente des milliers, voire des millions de pixels (dont les valeurs varient entre 0 et 255), il apparaît moins intéressant de comprendre les relations (poids) entre les X et Y . Par contre, on souhaitera que cette classification soit la plus précise possible pour pouvoir bien différencier les pommes des oranges ou des bananes lorsque qu'on utilise les recommandations d'un moteur de recherche. Par exemple, dans le cas de la voiture autonome, la détection d'objet par les senseurs doit atteindre 99,999% de justesse prédictive, sinon le risque d'accident devient trop élevé. L'utilisation d'algorithmes pour prendre des décisions critiques soulèvent aussi beaucoup d'enjeux éthiques. Par conséquent, le degré de précision requis dépend souvent de l'objet d'étude et du risque associé aux décisions prises par l'humain ou l'ordinateur à partir des valeurs « attendues ou prédites » du modèle développé. Dans la tradition de l'intelligence

artificielle on souhaite ultimement que le modèle ou l'algorithme développé performe aussi bien ou mieux qu'un humain en termes de justesse de la prédiction lorsqu'on lui présente une tâche similaire avec de nouvelles données ou informations.

1.2.2 Les concepts d'erreur réductible et irréductible

Dans le cas qui nous concerne, nous avons une liste de caractéristiques $X_1, \dots, X_p$, qui sont potentiellement associées au décrochage scolaire et l'on veut prédire Y (le décrochage scolaire entre 14 et 18 ans) en fonction de celles-ci. Par exemple, nous pourrions « potentiellement » vouloir intervenir sur ces facteurs X afin d'éviter l'événement Y . La justesse prédictive (*predictive accuracy*) de $\hat{y}$ comme étant une prédiction de Y dépend de deux valeurs que l'on appelle l'erreur réductible et l'erreur irréductible. En général, $\hat{f}$ est un estimateur imparfait de f , ce qui va créer de l'imprécision dans la prédiction, donc une certaine forme d'erreur. Cette erreur est réductible en utilisant l'algorithme ou modèle qui est le plus adéquat pour la tâche à réaliser (le chercheur peut en essayer autour de 5 ou 6 provenant des grandes familles décrites plus haut). La régression logistique incluant la non-linéarité et des termes d'interactions peut souvent être tout à fait adéquate tout en étant plus facilement interprétable (Cheng, Khomtchouk, Matloff, & Mohanty, 2018). La sélection exhaustive des caractéristiques X pour établir le modèle prédictif est aussi très importante et il faut éviter de les sélectionner a posteriori selon les corrélations observées avec Y (ce que l'on appelle communément du « *data snooping* »), sinon la généralisation du modèle sera alors compromise. Dans le contexte où $\hat{f}$ serait un estimé parfait de f , Y demeure quand même une fonction du terme d'erreur ϵ , ce qui peut affecter également la précision des

prédictions. Cette erreur est dite irréductible puisqu'elle est indépendante de l'estimation de f . Par contre, cette erreur peut contenir des variables non-mesurées qui peuvent permettre de prédire Y et qui n'ont pas été incluses dans le modèle initial. Elle peut aussi contenir des erreurs systématiques ou aléatoires concernant les variables que l'on mesure. L'intérêt de tester différents modèles est d'estimer f pour obtenir l'erreur réductible qui sera la plus petite afin de maximiser la précision des prédictions de Y .

Un des objectifs de l'entreprise scientifique et du développement des connaissances est de comprendre les liens, potentiellement causaux, qui peuvent exister entre Y et les différentes valeurs que peuvent prendre $X_1, \dots, X_p$. Dans ce cas, l'emphase ne porte pas tant sur la justesse de la prédiction que sur une meilleure compréhension des associations existants entre $X_1, \dots, X_p$ et Y . Donc, $\hat{f}$ peut difficilement être considéré comme une « boîte noire » puisque les valeurs des coefficients de X ainsi que leur signification statistique devient au centre du questionnement ou des hypothèses de recherche à valider. Les questions classiques lorsque l'on se préoccupe à la fois de l'inférence et de la prédiction sont les suivantes : 1) Quels sont les variables indépendantes ou « inputs » qui sont significativement liés à notre variable dépendante « output »? 2) Quelles sont la taille et la direction des effets entre X et Y ? 3) Est-ce que la relation entre Y et chacun des prédicteurs est linéaire ou faut-il utiliser des fonctions plus complexes pour bien décrire l'association? 4) Existe-t-il des interactions complexes entre les variables? Par contre, l'omission de variables indépendantes importantes dans le modèle peut grandement compromettre la validité de celui-ci, donc une emphase simultanée sur la causalité et la prédiction peut être souhaitable. De plus, la présomption que la relation entre les variables est linéaire devrait être plus souvent questionnée. Donc,

même si ces deux motivations ou objectifs liés à la modélisation peuvent apparaître distincts, dans plusieurs cas, on voudra préférablement joindre les deux.

Il existe deux grandes méthodes pour estimer f , les méthodes paramétriques et les méthodes non-paramétriques. La régression logistique entre dans la catégorie des modèles paramétriques.

1.2.3 Le compromis entre la justesse des prédictions du modèle et la capacité d'interprétation

Les modèles les plus restrictifs, comme la régression linéaire ou la régression logistique « classique » (additive), ne permettent d'estimer qu'une variété limitée de fonction f . Dans la grande famille des algorithmes, on peut définir ces modèles comme étant moins flexibles. Il est important ici de comprendre que la régression logistique peut devenir beaucoup plus flexible en introduisant des termes non-linéaires (quadratique, cubique, etc.) ainsi que des termes d'interactions entre les variables. Récemment, certains chercheurs ont démontré, par le biais du « *package* » **polyreg** dans R, que la régression polynomiale pouvait être comparable ou même mieux performée que certains modèles d'apprentissage profond, et ce, sur plus de cinq bases de données différentes (Cheng, Khomtchouk, Matloff, & Mohanty, 2018). D'autres familles d'algorithmes, comme les forêts aléatoires, sont aussi considérées comme étant plus flexibles puisqu'elles peuvent estimer un plus grand spectre de formes de f , en modélisant « automatiquement » (avec peu ou pas d'intervention humaine) la non-linéarité et les interactions entre des centaines de variables. L'automatisation du processus de sélection de modèles peut être très utile lorsque l'on est en présence de problèmes en haute dimensionnalité et dont les données

peuvent souvent être corrélées plus ou moins fortement entre elles. La sélection du modèle optimal devient alors très difficile pour l'humain puisque le nombre de possibilités et de combinaisons de variables devient trop grande et prohibitive à analyser visuellement. Pourquoi utiliser une approche en particulier? En général, si l'on s'intéresse à l'inférence, on privilégiera des modèles moins flexibles parce qu'ils sont plus faciles à interpréter. À l'autre extrême, certains modèles sont excellents pour approximer f afin d'avoir une classification plus juste particulièrement lorsque les problèmes sont complexes (non linéairement séparables), comme dans la classification de certaines données non structurées (p. ex., images, sons, langage naturel). Par contre, ces modèles rendent presque impossible la compréhension des liens qui existent entre Y et X . Donc, le chercheur est presque toujours confronté à un compromis dans la sélection du modèle ou de l'algorithme et pourra utiliser une technique comme la validation croisée, que l'on détaillera dans les prochains paragraphes, comme une approche facilitant le choix du modèle final sur des bases empiriques pour ultimement valider le modèle sur l'échantillon « test », qui lui, fournit de nouvelles données pour faire de nouvelles prédictions.

1.3 La validation croisée

La validation croisée souvent appelé « test au-delà de l'échantillon » (*out-of-sample testing*) comprend une série de techniques pour évaluer la validité d'un modèle statistique ou d'apprentissage automatique développé dans le but de voir la capacité de ce dernier à se généraliser et à permettre des prédictions au-delà des données de l'échantillon d'entraînement. On utilise traditionnellement ces techniques lorsque

l'échantillon est assez grand pour permettre de le diviser en plusieurs sous-échantillons, sinon on peut préférer les indices d'ajustement qui ont aussi été conçu pour assurer la reproductibilité. Certains indices ont même été développés pour les problèmes en haute dimension.

Lors du développement d'un modèle prédictif, on utilise généralement deux bases de données : une qui permet la construction du modèle à partir de « données d'entraînement » (*training data*) et une autre qui contient des données nouvelles ou inconnues (*first seen or unknown data*) que l'on appelle aussi « base de données test ou de validation » (*test or validation dataset*). L'objectif principal de la validation croisée est de s'assurer de la performance du modèle sur de nouvelles données, afin qu'il n'y ait pas de problèmes de sous-ajustement ou sur-ajustement (*under- or overfitting*). Il y a plusieurs types de validation croisée, les approches non-exhaustives comme celles présentées ci-dessus (méthode « holdout ») ainsi que la technique de validation croisée « k-plis », mais également des techniques exhaustives qui utilisent toutes les possibilités de diviser l'échantillon en données d'entraînement et de validation. La description détaillée de ces approches dépasse le cadre du présent chapitre.

Pour permettre la sélection optimale du modèle et des variables à inclure, on utilise souvent la procédure « *holdout* » qui est une pratique populaire en apprentissage automatique, mais que l'on retrouve plutôt rarement en sciences sociales et de la santé puisque les échantillons peuvent souvent être trop petits ($n < 100$) pour pouvoir l'appliquer. Cette technique consiste à diviser aléatoirement l'échantillon initial en deux groupes pour créer un échantillon d'entraînement et un échantillon « test ». Ce dernier permettra de valider la capacité prédictive de notre modèle sur de « nouvelles données ».

Même si dans le domaine de la statistique des indices d'ajustement des modèles ont été développés lorsqu'on utilise un seul échantillon, la validation croisée peut souvent être plus robuste, surtout lorsque les postulats des modèles ne sont pas entièrement respectés. De plus, les indices d'ajustement traditionnels pour sélectionner un modèle ne sont pas toujours adéquats en haute dimension ou lorsque la taille d'échantillon est très grande ou, à l'inverse, lorsqu'elle est beaucoup plus petite que le nombre de variables ou dimensions. Les différentes façons de sélectionner un modèle « optimal » demeurent un champ de recherche très actif et il est souvent pertinent de comparer différentes approches.

1.3.1 Le compromis entre le biais et la variance et la validation croisée

Cette section tentera d'expliquer un concept plus abstrait qui est la valeur attendue du taux d'erreur « test », c'est-à-dire le taux d'erreur du modèle en présence de nouvelles données. Cette valeur est plus généralement influencée par deux quantités, le biais et la variance qui sont associés aux paramètres du modèle. En statistique plus « classique », on tente d'estimer les paramètres d'un modèle qui a permis de générer les données observées au sein de l'échantillon. On souhaite que ces paramètres soient équivalents aux « vrais » paramètres populationnels. Les estimés qui n'ont pas cette propriété sont dits « biaisés ». En apprentissage statistique, le concept de biais est beaucoup moins négatif et peut même parfois être souhaitable. On contraste souvent le biais avec la variance qui consiste à ce que les paramètres du modèle dévient de leur tendance centrale, et cela, d'une base de données à une autre, comme on peut le constater en utilisant la validation croisée.

Yarkoni et Westfall (2017) offrent une illustration détaillée de ces deux concepts.

Il faut donc se rappeler que l'objectif ultime ou idéal est de choisir un modèle qui a une variance faible ainsi qu'un biais faible. Donc, la variance fait référence à une certaine mesure de variation lorsque l'on utilise une base de données d'entraînement pour ajuster notre modèle ou lorsque l'on utilise une technique de validation croisée avec K sous-échantillons. On peut présumer que les paramètres pourraient varier d'un sous-échantillon à un autre, mais ultimement, il serait souhaitable que ces variations soient minimales. En général, les modèles les plus flexibles ont aussi une plus grande variance, donc se généralisent moins bien en présence de nouvelles données.

Le concept de biais, quant à lui, fait référence à une erreur qui est inhérente au fait d'approximer un problème concret complexe en utilisant un modèle statistique ou d'apprentissage qui est fondamentalement trop simpliste ou parcimonieux. Donc, le modèle est alors sous-ajusté. Par exemple, on peut présumer qu'un modèle de régression linéaire serait un modèle potentiellement trop simple, dans plusieurs situations réelles, ce qui augmenterait le biais dans les prédictions de f . Par exemple, si f est une fonction non-linéaire ou multiplicative, une régression logistique « classique », qui est linéaire, en fonction du logit, donnera des prédictions imprécises indépendamment de la taille de l'échantillon d'entraînement. Dans des situations réelles, il est impossible de calculer simultanément le taux d'erreur de l'échantillon test, le biais et/ou la variance d'un modèle. Par contre, on doit toujours garder à l'esprit l'existence du compromis entre la variance et le biais. Donc, lorsqu'on utilise des modèles très flexibles, il arrive communément de se retrouver avec un modèle qui a un biais très faible, mais dont la variance est élevée. L'objectif ultime est d'avoir un modèle dont le biais et la variance sont faibles pour que celui-ci se généralise bien à de nouveaux ensembles de données.

Alors qu'en statistique, on cherchera absolument à minimiser le biais, en apprentissage automatique on tentera de minimiser l'erreur de prédiction et donc considérer simultanément les concepts de biais et de variance. Il est aussi important de considérer que la taille de l'échantillon, le nombre de variables dans le modèle et surtout leur ratio p/n , peuvent grandement influencer le biais, la variance et l'erreur test. C'est pour cette raison que les modèles d'apprentissage profond (modèles très flexibles) nécessitent beaucoup de données et des problèmes complexes pour bien performer.

1.4 Validation des modèles et sélection des variables

Une première façon de valider un modèle est d'observer à quel point celui-ci est capable de classifier correctement les sujets dans l'échantillon d'entraînement suivant nos deux catégories (décrocheurs et non-décrocheurs). L'idée est de mesurer un taux d'erreur « entraînement » qui reflète la proportion de mauvaises classifications basées sur les probabilités « prédites » estimées par le modèle. On peut alors élaborer une « matrice de confusion », terme utilisé en apprentissage automatique pour représenter une table de classification, afin de rendre de compte, entre autres, des faux positifs et des faux négatifs. L'enjeu, ici, est d'avoir la meilleure classification possible afin que celle-ci se généralise à l'échantillon « test », ou en d'autres mots à de nouvelles données.

1.4.1 Comment sélectionner les variables importantes par le biais des indices d'ajustement

Il y a évidemment plusieurs techniques pour sélectionner les variables qui seront incluses dans un modèle et pour déterminer la fonction entre les variables indépendantes

et la variable dépendante. Il y a des approches qui sont fondées sur des indices d'ajustement du modèle et qui peuvent être calculés à partir de la fonction de vraisemblance. Ceux-ci considèrent à la fois le nombre de paramètres dans le modèle et la taille de l'échantillon afin de choisir le modèle le plus parcimonieux. Le *Mallow's Cp*, *Akaike information criterion* (AIC), *Bayesian Information criterion* (BIC) sont des exemples de ce genre d'indices que l'on retrouve régulièrement en modélisation statistique, dont les équations structurelles et les modèles multiniveaux/hierarchiques (voir Chapitre X et Y) et qui ont été développés pour garantir la généralisation du modèle à de nouveaux échantillons et de garantir la reproductibilité des résultats. Ces outils peuvent être adéquats lorsque le ratio p/n suit les limites des analyses multivariées, mais peuvent se comporter inadéquatement dans d'autres situations, comme en haute dimension.

Même en basse dimension, ces indices ne sont pas toujours concordants, donc le chercheur doit demeurer à l'affût tout au long du processus décisionnel afin de sélectionner le modèle optimal. La validation croisée sera préférable aux indices d'ajustement lorsque le modèle est complexe, relativement à la quantité de données disponibles. Il a été noté que les modèles d'équations structurelles se généralisent parfois peu à de nouvelles données même si les indices d'ajustement étaient satisfaisants. Certains auteurs suggèrent que la validation croisée soit plus fréquemment utilisée avec ce genre de modèle (Browne et al., 2002).

1.4.2 La régularisation ou pénalisation

Une autre méthode de sélection de modèles ou de variables qui est très populaire en apprentissage automatique s'appelle la régularisation ou la pénalisation. Ces approches sont en émergence en psychologie et plus particulièrement en équations structurelles et modèles multiniveaux (McNeish, 2015; Jacobucci et al., 2019). La régularisation peut servir à plusieurs fonctions en vue de solutionner des problèmes dans l'estimation des modèles. L'idée générale est d'ajouter une certaine quantité d'information au modèle qui sera une fonction des coefficients de régression. Par exemple, des coefficients qui seraient gonflés par la multicollinéarité ou par une petite taille d'échantillon se verront pénalisés ou réduits (*shrinkage*). Ce genre de paramètres problématiques peut parfois être diagnostiqué par le facteur d'inflation de la variance en régression linéaire multiple (voir chapitre X). Selon le concept de compromis biais/variance, on introduit un certain biais dans nos paramètres pour pouvoir réduire la variance de ceux-ci et faire que l'erreur de prédiction sera optimale avec des nouvelles données. Dans le cas de certains types de régularisation, certains paramètres seront même réduits à zéro, ce qui simplifiera le modèle et diminuera encore une fois la variance de celui-ci. Il existe trois techniques de régularisation qui sont les plus populaires dans le contexte de la régression (même si aujourd'hui on peut en retrouver plus d'une dizaine) : 1) crête, 2) LASSO et 3) *elastic net*. Ce sont des méthodes de contraction des coefficients de régression qui ont des similarités, mais aussi de grandes différences. Par conséquent, on pourra utiliser la validation croisée pour estimer un paramètre optimal qui maximisera la justesse de la prédiction tout en réduisant l'importance (le poids) des variables qui augmentent la variance du modèle. L'objectif est de réduire les coefficients de régression de ces variables pour la régression crête, mais aussi d'en fixer à zéro pour les deux autres

types de pénalisation. Ceci permettra d'obtenir un modèle plus parcimonieux et éviter le surajustement ou surapprentissage (*overfitting*) lorsque p est grand (plus de 50 variables) ou éviter des problèmes de calculs computationnels lorsque p et les poids estimés atteignent les millions. Même si la régression crête et LASSO apparaissent avoir le même objectif, elles fonctionnent de manières différentes. La régression crête applique une régularisation que l'on appelle L_2 , c'est-à-dire que l'on ajoute une pénalité équivalente au carré de la taille des coefficients. Dans le cas de la régression LASSO, une régularisation L_1 est appliquée, donc on ajoute une pénalité qui équivaut à la valeur absolue de la taille des coefficients. Le paramètre d'ajustement λ dans le cas de la régression crête varie entre 0 et l'infini. Lorsque ce paramètre vaut 0, on obtient une régression logistique classique, alors que lorsque λ tend vers l'infini, les coefficients de régression vont tous être estimés près de zéro et la justesse de la prédiction sera alors grandement détériorée. La régression *elastic net* combine les deux types de pénalité L_1 et L_2 , donc, elle est souvent préférable aux deux autres approches. Par contre, chaque approche pourra fonctionner de manière optimale, tout dépendamment des caractéristiques des variables qui sont introduites dans le modèle, dont leurs variances et leurs corrélations ainsi que le ratio nombre de dimensions p par rapport à la taille d'échantillon n .

En résumé, la régularisation permet de prévenir le surajustement du modèle. La régression crête fonctionne bien lorsque p n'est pas trop grand et lorsque les caractéristiques sont modérément à fortement corrélées entre elles. La régression LASSO est plus susceptible à la multicollinéarité et aura tendance à choisir arbitrairement une variable parmi plusieurs, surtout si celles-ci sont fortement corrélées entre elles. Par contre, la régression LASSO est efficace lorsque p est très grand puisqu'elle permet de

sélectionner les variables en éliminant celles dont les coefficients de régression sont estimés à 0. Ceci permet d'avoir un modèle plus parcimonieux., l'*elastic net* (Zhou & Hastie, 2005) performe, en général, mieux que la régression LASSO en présence de multicollinéarité, donc elle sera souhaitable dans cette situation particulière. Elle estime simultanément les pénalités L_1 et L_2 . Il est fortement recommandé d'aller lire les articles originaux qui détaillent bien, par le biais de simulations, les problèmes que différentes formes de régularisation tentent de résoudre.

2. Illustration de l'utilisation de la régression logistique avec régularisation

2.1 Description de l'échantillon, du devis et des variables

2.1.1 Échantillon

Les données ont été colligées initialement auprès de $n=6773$ d'adolescents provenant de 12 écoles où le taux de décrochage est particulièrement élevé, autour de 36%, afin de mesurer un ensemble de facteurs de risque du décrochage scolaire. Au total, dix des 12 écoles étaient situées dans des quartiers défavorisés. Un sous-échantillon a par la suite été invité à une entrevue afin d'identifier les stressseurs auxquels les adolescents étaient exposés. L'objectif était d'interviewer 45 adolescents par école (pour un total de $n=545$), suivant un devis avec cas témoins appariés. D'abord, 15 élèves qui venaient de décrocher de l'école ont été interviewés. Ensuite, 15 élèves appariés ayant un profil initial de risques similaires mais qui persévéraient ont également été interviewés. Finalement, 15 autres élèves « normatifs », également persévérants, qui avaient un niveau moyen de risque ont été interviewés.

2.1.2 Variables

La description générale des variables est présentée dans le tableau 1. La variable dépendante est une variable dichotomique (codée 0 = non / 1=oui) représentant le fait qu'un étudiant ait décroché de l'école ou non. Un étudiant est considéré comme décrocheur s'il remplit au moins une des trois conditions suivantes : 1) avoir avisé officiellement de la cessation des études avant d'obtenir leur diplôme de secondaire 5, 2) avoir été transféré au système d'éducation aux adultes, 3) être absent pendant plus d'un mois de l'école sans avoir avisé des motivations sous-jacentes. Pour plus de détails sur les variables et le devis, voir l'article original (Dupéré et al., 2018)

Une particularité de la structure des données est que la grande proportion des variables sont ordinales et recodées en variables factices (*dummy*) dichotomiques. En général, les effets de la régulation sont plus marqués : 1) avec des variables intervalles/ratio puisqu'elles ont une plus grande variance et 2) en présence de multicollinéarité. Une autre caractéristique des données est que nous ne sommes pas en contexte de haute dimensionnalité puisque nous avons 25 variables pour 1000 participants, donc $p \ll n$. En dernier lieu, l'étude originale est de nature confirmatoire (hypothético-déductive) et non exploratoire (inductive), ce qui favorisera les algorithmes les moins flexibles, comme la régression logistique « classique ».

Tableau 1. Description des variables indépendantes introduites dans le modèle de régression logistique régularisée provenant de l'étude de Dupéré et al. (2018)¹

Nom des variables	Types de variables	Nom des variables dans le fichier de données
1. Sexe	Dichotomique	MALE
2. Âge	Intervalle	AGE
3. Parent immigré	Dichotomique	PAR_IMM
4. Ethnicité	Dichotomique	MINORITY

5. Niveau de scolarité parents	Intervalle	SCOLMAX
6. Mère en emploi	Dichotomique	TRAVAILM
7. Père en emploi	Dichotomique	TRAVAILP
8. Parents séparés	Dichotomique	PAR_SEP
9. Adaptation scolaire	Intervalle	ADAPT
10. Risque décrochage scolaire	Intervalle	SRDQ
11. Difficultés chroniques sévères	Intervalle	CHRONSEVACT
12. Stresseurs sévères 0-3 mois	Dichotomique	SEVER03DICO
13. Stresseurs sévères 3-6 mois	Dichotomique	SEVER36DICO
14. Stresseurs sévères 6-9 mois	Dichotomique	SEVER69DICO
15. Stresseurs sévères 9-12 mois	Dichotomique	SEVER912DICO
16. Stresseurs modérés 0-3 mois	Dichotomique	MODER203DICO
17. Stresseurs modérés 3-6 mois	Dichotomique	MODER236DICO
18. Stresseurs modérés 6-9 mois	Dichotomique	MODER269DICO
19. Stresseurs modérés 9-12 mois	Dichotomique	MODER2912DICO
20. Stresseurs faibles 0-3 mois	Dichotomique	LOW203DICO
21. Stresseurs faibles 3-6 mois	Dichotomique	LOW236DICO
22. Stresseurs faibles 6-9 mois	Dichotomique	LOW269DICO
23. Stresseurs faibles 9-12 mois	Dichotomique	LOW2912DICO
24. Stresseurs distaux sévères	Intervalle/ratio	EVDISTSEV
25. Stresseurs distaux modérés	Intervalle/ratio	EVDISTMOD

1. Les stresseurs de faible intensité n'avaient pas été inclus dans l'article original car ils n'étaient pas associés de manière bivariée au décrochage scolaire. Ils ont été ajoutés ici pour servir la démonstration et augmenter les contrastes, en incluant des éléments de

multicollinéarité et des effets de suppression entre les variables indépendantes/prédictrices ou entrées/*inputs*.

2.2 Objectifs de l'analyse

Le décrochage scolaire à l'école secondaire est un phénomène complexe et multidimensionnel. Le phénomène est rarement expliqué par le fait d'un seul facteur, mais plutôt par une combinaison de facteurs. De plus, la nature de la fonction multidimensionnelle qui lie les multiples facteurs et la probabilité du décrochage scolaire d'un adolescent est encore mal connue puisque la plupart des études scientifiques n'ont pas nécessairement mis l'accent sur le développement de modèles prédictifs, mais plutôt sur des modèles hypothético-déductifs avec quelques variables de contrôle. La plupart de ces facteurs peuvent aussi souvent être modérément à fortement corrélés entre eux ou potentiellement non linéaires (par souci de simplicité, nous ne modéliserons pas les relations non linéaires et/ou les interactions entre les variables), ce qui rend l'estimation de leurs contributions dans la prédiction complexes.

L'objectif principal de cette section du chapitre sera de sélectionner un modèle de régression logistique, de manière exploratoire, en utilisant des techniques qui sont particulières à l'apprentissage automatique afin de potentiellement prédire le décrochage scolaire avec la plus grande justesse prédictive à partir des 25 variables indépendantes (voir tableau 1). Nous utiliserons des données simulées à partir de l'échantillon initial, ce qui permettra d'utiliser les données comme outil pédagogique.

En résumé, la tâche de classification vise à identifier à la fois le nombre optimal de prédicteurs du décrochage scolaire, mais également l'algorithme de régularisation qui

permettra le meilleur ajustement du modèle aux données, considérant la spécificité des variables introduites dans le modèle.

2.3 Régression logistique régularisée : procédures et modèles

Comme décrit précédemment, trois algorithmes seront comparés par le biais de la validation croisée à k-plis (*K-fold*): La régression crête (*ridge*), la régression LASSO, ainsi que la régression *elastic net*. L'algorithme sélectionné pourrait par la suite servir à faire des prédictions à partir de nouvelles données qui ne proviennent pas de l'échantillon initial afin d'ultimement pouvoir développer un instrument de mesure pour identifier les adolescents à risque de décrochage scolaire.

La sélection du modèle et des variables à inclure dans ce modèle se fait principalement par la qualité de la classification fondée sur une « matrice de classification » aussi appelée « matrice de confusion » en apprentissage automatique, en utilisant les données d'entraînement. On doit idéalement éviter le sur- ou sous-apprentissage, c'est-à-dire d'avoir une excellente classification au sein de l'échantillon d'entraînement, mais une mauvaise classification avec de nouvelles données.¹ Une solution pour régler ce problème est de diviser aléatoirement l'échantillon en deux, 70% de données d'entraînement et 30% de données test (c.à.d., nouvelles données). Cette proportion est arbitraire, mais l'idée est de garder un échantillon d'entraînement le plus grand possible pour pouvoir développer notre modèle et aussi un échantillon test assez grand pour valider notre modèle. Par la suite, l'échantillon d'entraînement sera aussi divisé aléatoirement en 10 sous-échantillons de 70 unités pour permettre la validation

¹ Par exemple, si la relation est réellement quadratique entre nos X et notre Y et que l'on modélise une relation linéaire, on va être dans une situation de sous-apprentissage (biais élevé). Par contre, si l'on modélise une relation cubique, on se retrouvera avec un problème de sur-apprentissage (variance élevée).

croisée à k plis ainsi que la sélection du paramètre alpha ou lambda afin de permettre une régularisation optimale. Pour évaluer la performance sur de nouvelles données, le modèle final sera évalué en utilisant les données *test* puisque la recherche du meilleur modèle par un analyste/chercheur peut faire en sorte que celui-ci soit plus ou moins consciemment sur-ajusté (*overfitted*). On utilisera également une autre technique de rééchantillonnage, le bootstrap pour calculer un intervalle de confiance autour d'un taux de classification moyen pour chacun des algorithmes.

Nous utiliserons principalement la matrice de confusion pour évaluer nos modèles, mais les indices de spécificité et sensibilité peuvent aussi être des mesures utiles pour mieux détailler les erreurs de classification. Donc, chacun des modèles va nous permettre de calculer des probabilités prédites pour chacun des participants, ce qui permettra ultimement de les classer en fonction de celles-ci.²

2.4 Régression logistique régularisée : interprétations et tableaux des résultats

Une manière adéquate de débiter les analyses est d'estimer un modèle de régression logistique « classique » en utilisant les 25 prédicteurs. Il est toujours souhaitable de commencer les analyses avec des modèles linéaires simples, pour ensuite, les comparer séquentiellement avec des modèles plus flexibles (non-linéaires et non-additifs).

En premier lieu, nous modéliserons les données de notre échantillon à partir d'une régression logistique « classique ». Les 25 variables standardisées potentiellement associées au décrochage sont introduites dans ce modèle. Le tableau de classification ou matrice de confusion démontre la qualité de la classification basée sur la validation

² Il est également possible d'utiliser la courbe ROC. Avec cette technique, on utilisera l'aire en-dessous de la courbe (*area under the curve*) comme indicateur de performance du modèle.

croisée. On peut constater que 66,6% des participants sont bien classés en utilisant l'échantillon d'entraînement et la validation croisée. Le tableau 2 présente l'ensemble des paramètres et les différents modèles/algorithmes, de la régression logistique « classique » à la régression elastic net en utilisant les données de l'échantillon test.

Tableau 2. Coefficient de régression et justesse de la classification sur l'échantillon « test » pour la régression logistique « classique », crête, LASSO et elastic net.

Variables	Coefficient de régression			
	Logistique	Crête	LASSO	Elastic net
Intercept	0,01	0,01	0,01	0,01
Male	0,12	0,04	0,09	0,11
Age en années	0,38	0,10	0,33	0,35
Parent immigrant	-0,02	0,02	-	-0,01
Minorité visible	0,01	-0,01	-	-
Niveau de Scolarité max. des parents	0,28	0,12	0,27	0,28
La mère a un travail	0,12	0,06	0,10	0,11
Le père a un travail	-0,11	-0,03	-0,07	-0,09
Séparation des parents	0,55	0,23	0,51	0,53
Adaptation scolaire	-0,12	-0,05	-0,11	-0,12
SRDQ (score de risque de décrochage)	-0,04	0,00	-0,01	-0,03
Événements distaux sévères	0,27	0,05	0,16	0,19
Événements distaux modérés	-0,50	-0,05	-0,31	-0,40
Événements sévères (0-3 mois)	0,44	0,10	0,33	0,39
Événements sévères (3-6 mois)	-0,06	0,04	-	-
Événements sévères (6-9 mois)	-0,46	-0,03	-0,30	-0,37
Événements sévères (9-12 mois)	-0,02	0,00	-	0,03
Événements modérés (0-3 mois)	-0,37	0,02	-0,24	-0,31
Événements modérés (3-6 mois)	0,37	0,03	0,20	0,26
Événements modérés (6-9 mois)	0,45	0,01	0,24	0,33
Événements modérés (9-12 mois)	0,10	-0,05	-	0,01
Événements mineurs (0-3 mois)	0,26	0,12	0,24	0,25
Événements mineurs (3-6 mois)	0,12	0,05	0,10	0,12
Événements mineurs (6-9 mois)	-0,23	-0,12	-0,22	-0,23
Événements mineurs (9-12 mois)	-0,17	-0,06	-0,14	-0,16
Difficultés chroniques sévères	0,61	0,23	0,56	0,58
% accord	66,7%	66,7%	67,0%	67,0%
sd % accord (bootstrap)	2,75%	2,83%	2,98%	2,76%

Il est important de noter qu'aucune mesure de signification statistique n'est présentée puisqu'il n'est pas simple de les calculer directement lorsqu'on utilise la régularisation. Par contre, on pourrait utiliser une technique de rééchantillonnage comme le bootstrap pour estimer des intervalles de confiance à 95% autour de nos paramètres.

La régression crête (*ridge*) est par la suite testée. Pour sélectionner la valeur du lambda nous utiliserons la validation croisée avec 10 sous-échantillons. Un paramètre lambda correspondant au log (-3,48) a été choisi puisque c'est le paramètre d'ajustement ou « l'hyperparamètre » permettant la meilleure classification ou la déviance la plus faible. On peut voir à la Figure 1, l'évolution de l'indice de déviance en fonction du paramètre lambda. On voit qu'il y a une légère baisse de la déviance jusqu'à une valeur de -3.48 et par la suite, l'ajustement du modèle se détériore avec une augmentation de la déviance. À la Figure 2, on peut voir qu'il y a une réduction des coefficients de régression jusqu'au lambda optimal. Par contre, aucun coefficient n'est fixé à zéro, ce qui ne permet pas de réduire la taille du modèle. Par contre, comme on peut le voir au Tableau 2 les coefficients de régression sont beaucoup plus faibles que pour les autres modèles. La multicollinéarité, à laquelle la régression crête est particulièrement sensible, semble une explication plausible pour expliquer cet effet sur le modèle.

Figure 1. Indice de déviance pour les prédictions de la régression crête en fonction de l'hyperparamètre lambda et de la régularisation L_2 .

Figure 2. Les coefficients standardisés de la régression crête en fonction du paramètre lambda et de la régularisation L_2 . La ligne rouge représente la valeur de l'hyperparamètre lambda optimal.

On constate peu de différences avec le modèle de régression « classique », mais il y a une légère amélioration de la classification. La matrice de confusion démontre que 70,6% des participants sont bien classifiés.

Maintenant, nous allons comparer les performances de la régression LASSO aux deux modèles précédents. On se rappelle que la régression LASSO peut à la fois réduire la taille des paramètres de régression et aussi en fixer à zéro, ce qui permettra d'obtenir un modèle plus parcimonieux, donc avec une variance moins grande. Nous utilisons la même technique que précédemment pour sélectionner le λ . Le λ optimal équivaut à approximativement $\log(-5,83)$ (voir Figure 3). Comme on peut le voir à la Figure 4, lorsque certains coefficients croisent la ligne de zéro, ils sont fixés à cette valeur. Les degrés de liberté de l'ensemble du modèle varient alors en fonction du nombre de variables qui sont fixés à zéro. La matrice de confusion démontre que 70,6% des participants sont bien classifiés. Donc, le modèle LASSO semble légèrement mieux s'ajuster à nos données que la régression logistique « classique » tout en simplifiant notre modèle initial, en éliminant 5 variables (voir Tableau 2). La justesse de la prédiction est similaire à la régression crête.

Figure 3. Indice de déviance pour les prédictions de la régression LASSO en fonction de l'hyperparamètre λ et de la régularisation L_1 .

Figure 4. Les coefficients standardisés de la régression LASSO en fonction du paramètre lambda et de la régularisation L_1 . La ligne rouge représente la valeur de l'hyperparamètre lambda optimal.

En dernier lieu, nous allons comparer la régression *elastic net* qui est un modèle qui comprend à la fois une combinaison de paramètres L_1 et L_2 , donc un modèle qui est un peu plus complexe que le LASSO. Dans ce modèle, on fera varier l'hyperparamètre

alpha qui est fixé à 0 pour la régression crête et à 1 pour la régression LASSO. On préférera une approche incrémentale, en faisant varier le paramètre alpha de 0,1 à 0,9 par incrémentation de 0,1. Il y a très peu de différences entre les différents modèles de régularisation lorsque l'on tient compte de la déviance, celle-ci est toujours approximativement 1,24 peu importe le niveau du paramètre alpha. La Figure 5 représente l'évolution de l'indice de déviance en fonction des hyperparamètres alpha et lambda.

Figure 5. Indice de déviance pour les prédictions de la régression *elastic net* en fonction des hyperparamètres alpha et lambda et de la régularisation L2 et L1.

La matrice de confusion démontre que 67,7% des participants sont bien classifiés. Cette performance est légèrement moins bonne que la performance de la régression LASSO. Donc, il n'y a pas de « valeur ajoutée » à utiliser un algorithme/modèle plus complexe. Nous aurions pu également inclure dans nos modèles des termes non-linéaires ou des termes d'interactions, mais pour les fins de l'exercice, nous allons demeurer avec des modèles linéaires.

En dernier lieu, l'objectif est de valider les différents modèles avec les données « test ». On peut se rendre rapidement compte que la qualité de la classification avec les données de validation est relativement semblable à ce qui avait été obtenu avec l'échantillon d'entraînement. Ceci est rassurant et ce qui est d'ailleurs attendu. De plus, il y a peu de variation entre les quatre différents modèles appliquer aux données « test ». Comme il est présenté dans le bas du Tableau 2, la régression logistique « classique » obtient un taux de classification avec l'échantillon test de 66,7%, la régression crête de 66,7% la régression LASSO de 67,0% et la régression *elastic net* de 67,0%. Lorsque l'on utilise une technique de rééchantillonnage de type bootstrap pour calculer l'écart-type de cette classification (variant de 2,8% à 3,0%), on voit à quel point les modèles ont une performance similaire. Dans cette situation, un chercheur pourrait vouloir choisir la régression LASSO parce que ce modèle est plus parcimonieux, en fixant les coefficients de 5 variables à exactement zéro.

Il peut être intéressant de visualiser l'importance relative des variables qui ont été sélectionnées par les différents algorithmes. Les coefficients de régression des différentes variables pour les quatre modèles sont présentés dans le Tableau 2. Les paramètres sont relativement similaires d'un modèle à l'autre. Pour le modèle LASSO, les variables qui semblent les plus contributives à la prédiction du décrochage scolaire sont en ordre décroissant, l'exposition à des stressors chroniques ($\beta = 0,56$), la séparation des parents ($\beta = 0,51$) et l'âge ($\beta = 0,33$). Encore une fois, il faut faire attention dans l'interprétation des paramètres puisque ceux-ci sont biaisés. De plus, c'est l'ensemble du modèle et la justesse prédictive sur laquelle nous mettons l'accent et comme on peut le constater, il n'y a pas de modèles qui se distingue franchement des autres. Dans le même ordre d'idée, le modèle actuel est additif et il pourrait être pertinent d'évaluer la performance d'algorithmes qui modélisent automatiquement les interactions comme les forêts aléatoires et les réseaux de neurones artificiels multicouches. Juste avec 25 variables, il est déjà prohibitif de tester tous les modèles possibles, donc des algorithmes plus automatisés pourraient être une option intéressante.

Conclusion

Dans un article important en statistique, Breiman (2001) argumentait qu'il y a deux cultures en modélisation statistique. La culture majoritaire était formée de statisticiens qui modélisent les relations entre les données afin d'estimer des paramètres non-biaisés en présumant d'avance du processus qui a généré les données observées. À l'autre extrême se trouve d'autres statisticiens (et aussi les chercheurs en apprentissage automatique) qui se définissent comme provenant d'une culture plutôt « algorithmique » et qui se préoccupent peu du processus de génération des données, mais qui tente de

laisser la « machine » définir l'algorithme/modèles qui lient des *inputs* à des *outputs*.

Vingt ans plus tard, la culture dominante a bien changé. Cette conception est très similaire à la tension qui peut exister entre les chercheurs qui sont centrés sur l'explication des processus et ceux qui mettent l'accent sur la justesse de la prédiction et sur la prise de décision basée sur ces prédictions.

Outre ces différences de culture statistique, les approches algorithmiques et la modélisation statistique « classique » utilisent des outils qui sont très similaires et les différences perçues viennent souvent du jargon développé dans ces deux types de culture. Une autre spécificité provient du type de données, qualifiées de non structurées, et qui sont plus fréquemment utilisées par la culture algorithmique (sons, images et traitement du langage naturel). Il faut aussi comprendre que l'utilisation des modèles statistiques en apprentissage automatique n'a pas les mêmes objectifs que l'utilisation scientifique « classique » de la statistique qui est plutôt hypothético-déductive ou qui s'intéresse au « pourquoi » et au « comment » (processus conditionnels) différentes variables peuvent être associées entre elles. L'objectif principal en apprentissage automatique est de définir une fonction qui lie un ensemble de variables entre elles dans le but de maximiser la justesse de la classification ou de la prédiction et de par la suite les mettre en production pour faire des machines/robots qui prennent des décisions et des actions basées sur ces prédictions.

Aujourd'hui, au lieu d'utiliser un code informatique pour développer un logiciel, on utilise les « poids de modèles de régression » qui permet de calculer des valeurs prédites et ainsi prendre des décisions automatiquement à partir de nouvelles données. Plus particulièrement, la régression régularisée a été développée dans le but d'estimer des

modèles en présence de multicollinéarité ou d'identifier des modèles plus simples face à un problème statistique donné, souvent en haute dimensionnalité. Entre autres, la régression LASSO ou *elastic net* vont permettre de sélectionner « automatiquement » des variables sans utiliser des indices d'ajustement ou des techniques moins stables comme les approches pas à pas (*stepwise*) qui sont généralement moins fiables. Le chercheur doit donc choisir un modèle statistique/algorithmique en fonction de la complexité d'un problème donné. Il fait alors face à un compromis entre le biais et la variance du modèle choisi et la capacité interprétative des paramètres. La validation croisée devient alors un outil important pour sélectionner le modèle afin de le soumettre ultimement à de nouvelles données. Ultimement, les approches algorithmiques peuvent être conçus comme des outils complémentaires aux analyses multivariées « classiques » qui peuvent potentiellement permettre une meilleure reproductibilité des recherches en science sociales et de la santé et aussi le développement de nouvelles théories en explorant les données afin de générer de nouvelles hypothèses. Par contre, ces outils deviennent indispensables pour la sélection et la convergence des modèles en haute dimension.

Références

- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199-231.
- Browne, M. W., MacCallum, R. C., Kim, C. T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological methods*, 7(4), 403.
- Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature methods*, 15(4), 233.
- Cheng, X., Khomtchouk, B., Matloff, N., & Mohanty, P. (2018). Polynomial regression as an alternative to neural nets. *arXiv preprint arXiv:1806.06850*.
- Dupéré, V., Dion, E., Leventhal, T., Archambault, I., Crosnoe, R., & Janosz, M. (2018). High school dropout in proximal context: The triggering role of stressful life events. *Child development*, 89(2), e107-e122.
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, 14, 91-118.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: prediction, inference and data mining. *Springer-Verlag, New York*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple-indicators, multiple-causes models. *Advances in methods and practices in psychological science*, 2(1), 55-76.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, 50(5), 471-484.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*.

Shatte, A., Hutchinson, D., & Teague, S. (in press). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 1-23.

doi:10.1017/S0033291719000151

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122.