Annexe 2. Fonctionnement de l'IA supervisée mise en oeuvre.

Principaux généraux

Dans la mesure où les données mobilisées rendent comptent de très nombreuses dimensions, il est possible de « reconstruire » (le verbe adéquat est « inférer ») les relations entre ces différentes données par une méthode statistique bien choisie.

À titre d'illustration, s'il est observé que les personnes non-diplômés du supérieur rapportent des risques de décrochage scolaire plus élevés pour leurs enfants, le fait de ne pas être diplômé du supérieur va constituer une variable permettant, dans une certaine mesure, de mesurer et prédire le risque de décrochage perçu par les parents.

La méthode décrite ci-dessous met en œuvre ce principe de façon automatisée dans le cas où un très grand nombre de facteurs de risque potentiels sont examinés de façon concomitante pour l'ensemble des répondants.

Description de l'algorithme d'apprentissage supervisé

Sur un plan mathématique, nous cherchons à construire un modèle appelé classifieur de la forme [y = *C*(x1, …, xi, …, xn)] permettant de déterminer si les prédicteurs x = (x1, …, xi, …, xn) associés aux personnels jouent un rôle important pour rendre compte, par exemple, du risque de décrochage scolaire perçu par les parents (y).

Ce modèle (*C*), peut être obtenu par un algorithme d'apprentissage supervisé à partir des étiquettes (y) et des prédicteurs (x).

Il est à noter que les variables médiatrices ou "médiateurs" (m) constituent une catégorie particulière de prédicteurs dont la classification est proposée par les auteurs sur le fondement d'une connaissance *a priori* pour faciliter l'interprétation des résultats.

En raison de sa relative facilité d'utilisation[1], de son fonctionnement intuitif qui rappelle les arbres de décision humains, de sa bonne performance pour des problèmes de complexité intermédiaire[2] et de sa robustesse[3], le choix a été fait de recourir à l'algorithme d'apprentissage appelé « forêt d'arbres décisionnels »[4]. En pratique, le package « ranger » a été employé[5].

Une telle forêt combine de nombreux arbres de décision afin d'améliorer les performances et la robustesse des prédictions du modèle (méthode dite d'ensemble). Un arbre de décision constitue un modèle représentant des décisions successives situées sur les branches d'un arbre logique.

---

[1] La manière dont la forêt est construite n'est pas impactée par les transformations qui conservent l'ordre des variables.

[2] Au sens du nombre de prédicteurs employés.

[3] Le risque de sur-apprentissage est limité avec cette méthode.

[4] Cf. [Léo Breiman (2001) Random Forests. *Machine Learning*], article cité plus 72 000 fois en avril 2021 (source : Google Scholar).

[5] Proposé sur le langage libre R, « ranger » constitue une implémentation en C++ de l'algorithme originel de Breiman, ce qui conduit à des gains de performance importants.

Sur un plan plus formel, est entraîné, un modèle du type « y = C(x) », par le biais d'une forêt d'arbres décisionnels de classification [C = { c1, …, cn}] où les étiquettes de supervision (= « y ») correspondent à la variable à prédire et où les prédicteurs (= « x ») sont les variables décrivant les profils des répondants et les variables de médiation.

Les valeurs manquantes sont imputées par la méthode usuelle des k plus proches voisins, où k est pris comme la partie entière de la racine carrée du nombre d'observations.

L'ensemble d'apprentissage est obtenu en tirant au hasard trois quart des observations relatives aux décisions de classement. Concernant le paramétrage de la forêt, les valeurs usuelles sont employées[6].

La précision du modèle C entraîné est calculée à partir d'un ensemble test comportant le quart des observations qui n'a pas été utilisé lors de la phase d'entraînement (données indépendantes). Les méthodes d'analyse de cette précision sont rapportées dans la section "Troisième message".

Déduction des variables prédictives à partir de l'algorithme entraîné

Une fois la forêt entraînée, les prédicteurs clés sont obtenues en mesurant l'importance des variables. Cette importance est quantifiée par la baisse de précision obtenue par permutation de chaque prédicteur[7] dans les arbres composants *C*. Cette procédure est proposée dans le cadre du package « ranger ».

Pour apporter un niveau de preuve élevé sur l'importance des prédicteurs, il a été calculé une p-valeur[8] pour chaque variable suivant la méthode de Altmann[9], qui repose sur des permutations des étiquettes associées à un grand nombre de réentrainements du modèle. Le recours à un test de permutation permet de s'assurer que le poids estimé pour un critère ne peut s'expliquer par le hasard.

Afin de faciliter la présentation des résultats dans les tableaux et graphiques, l'importance des variables a été normalisée[10]. Par ailleurs, la présentation graphique des critères employés intègre le sens dominant des variables[11].

---

[6] 1000 arbres par forêt et un nombre de variables testées à chaque division fixée à la racine carré du nombre de prédicteurs et arrondi par le bas. Une optimisation de ces deux hyper-paramètres principaux (ainsi que d'autres secondaires) a été mise en oeuvre avec le package « Caret ». Mener une telle optimisation ne conduit pas à des augmentations de performances majeures. Cette méthode n'a donc pas été employée pour gagner en vitesse de calcul.

[7] Une permutation correspond ici à un changement au hasard de l'ordre de succession des valeurs d'un prédicteur donné.

[8] En statistique, désigne la probabilité que l'effet observé soit imputable au hasard.

[9] Cf. [Altmann et al (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*]. La méthode d'Altmann est implémentée dans le package « ranger ».

[10] Chaque importance de variable a été passée à la valeur absolue et divisée par la somme des valeurs absolus des importances calculées pour l'ensemble des variables. La somme des importances normalisées par cette méthode fait donc 100 %.

[11] Le signe du coefficient de corrélation de rang entre le prédicteur et la variable à prédire a été utilisé pour représenter le sens majoritaire de l'effet du prédicteur sur la variable à prédire.