

I. Contexte

Dans le cadre du programme de répression des fraudes à la Couverture Maladie Universelle Complémentaire (CMUC) en 2010, une expérimentation des méthodes de Datamining sur le thème de la CMUC a été demandée au département statistique de la DCCRF.

Un groupe de travail national constitué de **9 CPAM** (choisies en fonction des remontées du VLR et de leur performance dans la détection des fraudes à la CMUC) a été mis en place pour ce programme

II. Données

Les 9 CPAM du groupe de travail ont établi la liste des bénéficiaires ayant fait l'objet d'un contrôle CMU C au sein de leur caisse, en distinguant les contrôles a priori (amenant à un refus à l'ouverture des droits à la CMUC), de ceux a posteriori (conduisant à un préjudice subi au titre de la CMUC accordée à tort).

Une requête élaborée au niveau national et exécutée sur ERASME régional a permis d'extraire, pour chaque caisse du groupe de travail, l'ensemble des données anonymisées relatives aux bénéficiaires contrôlés sur la profondeur de l'historique disponible (période comprise entre juillet 2008 et novembre 2010 selon la CPAM) ainsi que de disposer de cas témoins, non contrôlé, tirés au sort aléatoirement. Cependant, dans la mesure où seules 3 des CPAM du groupe de travail présentaient un nombre significatif de bénéficiaires contrôlés pour la construction des modèles, nous avons restreint les données de la base d'analyse aux seules données de 3 CPAM.

Des indicateurs portant sur les domaines « famille » et « prestations » ont été créés à partir des données extraites.

Afin de comparer les individus sur des critères équivalents, les indicateurs portant sur la consommation de soins ont été calculés, pour chaque individu, sur les 6 derniers mois couverts au titre de la CMUC.

Devant le faible nombre de cas de fraudes contenu dans la base de travail (174 fraudes a priori et 58 fraudes a posteriori), les 2 types de fraudes ont été agrégés. Les modèles testés ne discriminant pas suffisamment le phénomène de la fraude, des cas de fraude ont été générés aléatoirement afin d'équilibrer les effectifs (effectifs respectifs: 1.386 et 1.385).

III. Modélisations

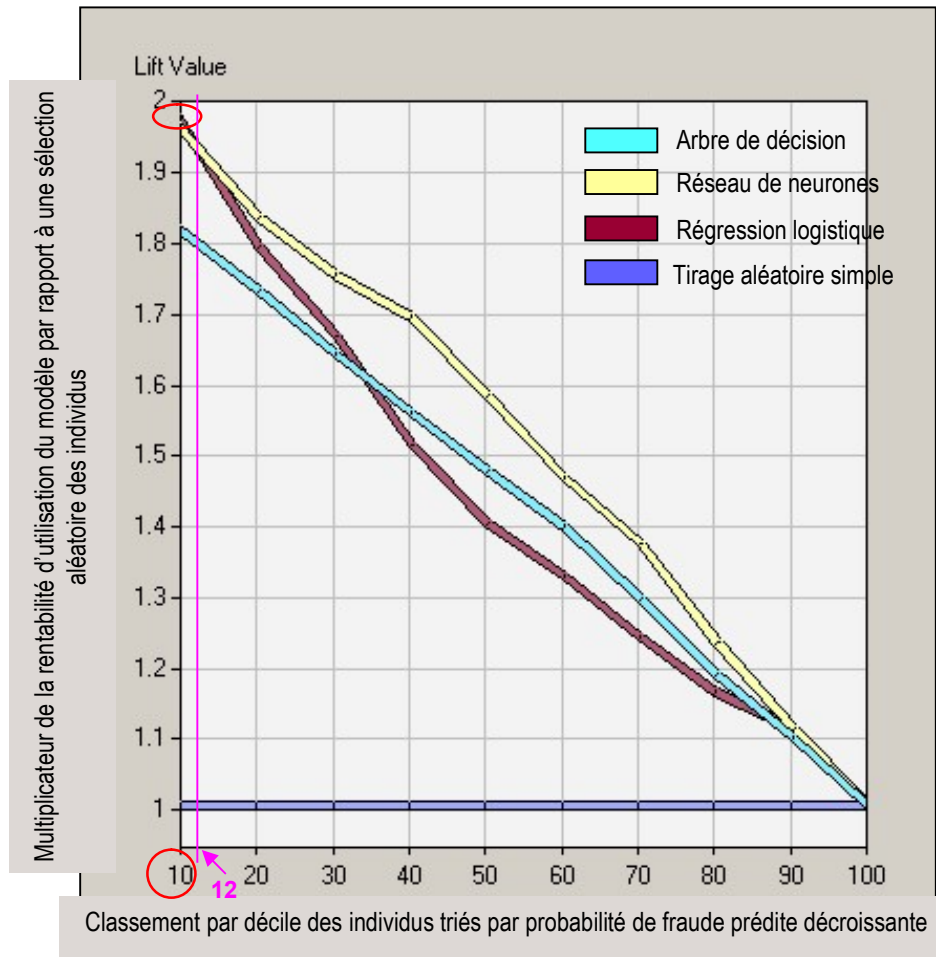
La base de travail est découpée en un échantillon d'apprentissage (**1 940** individus) utilisé pour la construction des modèles et en un échantillon de validation (**831** individus) permettant de s'assurer de leur fiabilité et de sélectionner le plus performant.

Trois méthodes ont été utilisées et mises en concurrence pour discriminer le phénomène de la fraude :

- L'arbre de décision qui a l'avantage de fournir des règles de décision lisibles et facilement interprétables,
- La régression logistique dont le point fort est de pouvoir quantifier la force de la liaison entre la fraude et les variables explicatives tout en tenant compte de l'effet des autres variables,
- Les réseaux de neurones qui affectent des pondérations à chaque variable. Le score produit est opaque et ne permet pas de comprendre les composantes du phénomène étudié.

IV. Comparaisons des modèles

Afin de comparer les modèles entre eux, une courbe représentant les zones de concentration des cas de fraude est dressée pour chacun des modèles (dans cette base, les individus sont classés par probabilité de fraude prédite décroissante).






- Explication :

Chacun des 3 modèles produits (arbre de décision, réseau de neurones et régression logistique) attribue un score, à chaque individu, équivalent à la probabilité de fraude prédite par le modèle. Ce graphique permet de mesurer, par décile d'individus triés par score décroissant, la rentabilité d'utilisation de chaque modèle en comparaison d'une sélection aléatoire d'individus (représentée par la droite en bleu foncé sur le graphe).

Par exemple, en partant de l'hypothèse d'un taux de fraudes à la CMUC de 3%, le graphique nous indique qu'en utilisant la régression logistique ou les réseaux de neurones, le modèle nous permettra d'identifier au sein du 1^{er} décile d'individus avec la plus forte probabilité de fraude, environ **1.98 fois** (Lift value=1.98) plus de cas de fraude qu'un tirage aléatoire simple (représenté par la droite en bleue). Ainsi, sur une base de 10 000 individus, le modèle détecterait 60 cas de fraude au sein des 10% d'individus (1 000) ayant la plus forte probabilité contre 30 par tirage aléatoire simple.

Jusqu'à 12% des effectifs (trait rose sur le graphe), la régression logistique offre le modèle le plus performant. Au-delà, les réseaux de neurones sont plus performants.

V. Limites

-  Les modèles ont été conçus sur une base qui ne reflète pas la réalité du phénomène (données bootstrapées),
-  Certains cas de fraude recensés relèvent d'une action spécifique sur une thématique locale. La notion de fraude n'est donc pas homogène,
-  La CMU complémentaire est octroyée à l'ensemble des individus constituant un foyer. Cette notion de foyer n'est pas identifiable dans nos systèmes d'informations ; seuls les assurés et ayants droit rattachés ont été sélectionnés. Les indicateurs ne concernent donc pas l'exhaustivité du foyer ce qui constitue un risque de biais.

VI. Conclusion

Avec le jeu de données disponibles (bootstrapées), en sélectionnant les 10 % d'individus ayant les plus forts scores de fraude et en déployant la régression logistique, on ne peut espérer en retour que doubler la détection de la fraude (Lift de 1.98) comparativement à un tirage aléatoire.

De plus, le gain apporté par ces techniques est d'autant plus à modérer que leur efficacité a été comparée à un tirage aléatoire simple alors qu'il est vraisemblable que les agents techniques des caisses auront une meilleure expertise pour la détection de fraudes CMUC compte tenu de leur connaissance métier.

Enfin, à partir des résultats des différentes simulations, on peut conclure qu'en partant d'un taux de fraudes à la CMUC relativement faible, le gain obtenu, via l'application du modèle le plus performant, n'est pas suffisant pour mettre en place un contrôle efficient au regard du nombre important de faux positifs détectés.