Apprentissage machine pour la prédiction de l'attrition: une étude comparative

Louis Geiler [*] [,**], Séverine Affeldt[*] Mohamed Nadif [*],

[*] Université de Paris, CNRS, Centre Borelli UMR9010, 75006, France.
[**] Brigad, 34 Rue du Sentier, 75002, Paris

Résumé. La prédiction du taux d'attrition est une préoccupation économique majeure pour de nombreuses entreprises. Différentes approches d'apprentissage ont été proposées, toutefois le choix à priori du modèle le plus adapté reste une tâche non triviale car extrêmement dépendante des caractéristiques intrinsèques des données d'attrition. Notre étude compare huit méthodes d'apprentissage supervisé combinées à sept approches d'échantillonnage sur treize jeux de données publiques relatifs au désabonnement. Nos évaluations, rapportées en termes d'aire sous la courbe (AUC), explorent l'influence du rééquilibrage et des propriétés des données sur les performances des méthodes d'apprentissage. Nous nous appuyons sur le test de Nemenyi et l'Analyse des Correspondances comme moyens de visualisation de l'association entre modèles, rééquilibrages et données. Notre étude comparative identifie les meilleures méthodes dans un contexte d'attrition et propose une chaîne de traitements générique performante basée sur une approche ensemble.

1 Introduction

La mise en place d'une bonne gestion de la relation client est devenue un sujet crucial pour de nombreuses entreprises qui concentrent notamment leur attention sur la *rétention* des clients. En effet, il apparaît aujourd'hui clairement que les coûts d'acquisition d'un nouveau client peuvent être beaucoup plus élevés que les coûts de rétention d'un client existant (Yang et Peterson, 2004). Jusqu'à récemment, les services du marketing et de l'industrie financière préféraient l'exploitation des méthodes de modélisation statistique pour l'analyse et la prédiction du taux de *désabonnement*, telles que l'analyse de survie, la modélisation par équations structurelles ou l'analyse de la variance. L'étude que nous proposons n'explore pas ces approches traditionnelles et se concentre sur les techniques d'apprentissage machine qui sont de plus en plus utilisées dans le contexte du départ ou du désabonnement des clients.

Notre objectif principal est de comparer plusieurs variantes d'une chaîne de traitements pour l'analyse des désabonnements. Cette chaîne comporte *(i)* une étape de *rééquilibrage* des classes, *(ii)* une phase d'apprentissage supervisé et *(iii)* une procédure d'évaluation robuste (Fig. 1). Une analyse exhaustive de toutes les variantes des algorithmes de cette chaîne n'étant pas envisageable dans le cadre de cet article, nous nous concentrons sur les algorithmes d'apprentissage dans leur version originale. Par ailleurs, le départ d'un client étant un événement