Une approche bayésienne non paramétrique de sélection de variables pour la modélisation de l'uplift

Mina Rafla[*,**], Nicolas Voisine[*], Bruno Cremilleux[**], Marc Boullé[*]

[*] Orange Innovation, 22300 Lannion, France
{mina.rafla, nicolas.voisine, marc.boulle}@orange.com
[**] UNICAEN, ENSICAEN, CNRS - UMR GREYC, Normandie Univ
14000 Caen, France
bruno.cremilleux@unicaen.fr

Résumé. Le présent article est un résumé de l'article Rafla et al. (2022) publié à la conférence ECML/PKDD 2022. La modélisation de l'uplift vise à estimer l'impact d'un traitement sur un individu, tel qu'une campagne de marketing ou d'un médicament. Les données d'uplift des banques ou des télécoms comportent souvent des centaines voire des milliers de variables. Dans de telles situations, la détection des variables non pertinentes est une étape essentielle pour réduire le temps de calcul et augmenter la performance du modèle. Nous présentons une méthode bayésienne de sélection de variable sans paramètres pour la modélisation de l'uplift. Cette méthode repose sur une méthode de discrétisation automatique des variables selon une approche bayésienne. Les expériences montrent que la nouvelle méthode permet à la fois d'éliminer les variables non pertinentes et d'obtenir de meilleures performances que les méthodes de l'état de l'art.

1 Introduction

La modélisation de l'uplift vise à estimer l'impact d'un traitement sur un individu, tel qu'une campagne de marketing ou un médicament. Les modèles d'uplift permettent d'identifier les groupes de personnes susceptibles de répondre positivement à un traitement *uniquement parce* qu'ils en ont reçu un. Ce domaine de recherche a de multiples applications comme la gestion de la relation client, la médecine personnalisée, la publicité. L'estimation de l'uplift est fondée sur des groupes de personnes qui ont reçu différents traitements. Une difficulté majeure est que les données ne sont que partiellement connues : il est impossible de savoir pour un individu si le traitement choisi est optimal car ses réponses aux traitements alternatifs ne peuvent pas être observées. Plusieurs travaux abordent les défis liés à la modélisation de l'uplift (Jaskowski et Jaroszewicz, 2012; Zhao et al., 2017).

De nombreuses bases de données sont volumineuses et contiennent des centaines de variables (Hu, 2022). Conserver toutes les variables est coûteux et inefficace pour construire des modèles d'uplift. Un processus de sélection des variables est alors une étape essentielle pour éliminer les variables non pertinentes, améliorer la précision de l'estimation et accélérer la construction du modèle. Alors qu'il existe de nombreuses méthodes de sélection de variables