Evaluation de l'uplift sur des données biaisées dans le cas du Non-Random Assignment

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

[*]Orange Labs, 22300 Lannion, France
[**]UNICAEN, ENSICAEN, CNRS - UMR GREYC, Normandie Univ
14000 Caen, France

Résumé. L'uplift est une mesure d'impact d'une action (marketing, traitement médical) sur le comportement d'une personne. La prédiction d'uplift repose sur des groupes de personnes ayant subi des actions particulières. Ces groupes sont estimés "équivalents". Or, en pratique, on constate qu'il existe des biais entre ces groupes. Pour résoudre cet écueil nous proposons un protocole d'évaluation de l'uplift dans le cas du biais de "Non-Random Assignment". Muni de ce protocole nous évaluons les performances sur les principales méthodes d'uplift de la littérature puis nous proposons une méthode pour réduire l'effet de ce biais. Des résultats expérimentaux sur 8 jeux de données montrent que notre méthode apporte une amélioration significative des performances de l'estimation de l'uplift.

1 Introduction

La modélisation de l'uplift, également connue sous le nom de Individual Treatment Effect (ITE), est une technique de modélisation prédictive qui modélise directement l'impact incrémental d'un traitement sur le comportement d'un individu. Les applications sont multiples : gestion de la relation client pour la modélisation d'actions de marketing direct, médecine personnalisée, publicité, élections politiques, etc. Les modèles d'uplift aident à identifier les groupes de personnes susceptibles de répondre positivement à une sollicitation marketing ou à un traitement médical. Plus généralement, un modèle d'uplift est un moyen de prédire, avec un certain taux d'erreur, l'impact d'un traitement sur le comportement d'une personne. Une difficulté inhérente à la modélisation de l'uplift est que les données ne sont que partiellement étiquetées. Il est impossible de savoir pour une personne si le traitement choisi est optimal parce que ses réponses aux traitements alternatifs ne peuvent pas être observées. Plusieurs travaux portent sur des défis liés à la modélisation de l'uplift. L'uplift a d'abord été modélisé dans le cas du bi-traitement (Jaskowski et Jaroszewicz, 2012) puis en multitraitement (Zhao et al., 2017). L'évaluation des modèles d'uplift, sujet de cet article, est étudiée dans (Radcliffe et Surry, 2011).

Les méthodes d'uplift reposent sur l'hypothèse que les échantillons utilisés sont homogènes. Cela signifie que l'uplift devrait être modélisé sur des données expérimentales, c'est-à-dire des données dont la génération est maîtrisée et pour lesquelles il n'y a pas de biais entre

les données des différents traitements. Or, en pratique, les données sont de nature observationnelle et on constate l'existence de biais. Par exemple, la non réponse à un appel commercial introduit un biais entre personnes contactées et non contactées. Ce biais, en pratique, pénalise l'utilisation des méthodes de modélisation d'uplift (Olaya et al., 2020).

L'objectif de cet article est d'étudier le biais de "Non-Random Assignment" (NRA) (Berk et al., 1988), un biais très courant dans la modélisation de l'uplift. Le biais NRA arrive lorsque la probabilité qu'une personne reçoive un traitement dépend des caractéristiques de cette personne. Nous traitons deux questions de recherche liées à ce biais : quel est l'impact du biais NRA dans la modélisation de l'uplift ? Comment réduire en pratique ce biais ? Pour répondre à la première question, nous concevons un protocole expérimental qui permet de quantifier l'impact du biais NRA sur les méthodes d'uplift. Notre étude permet de dégager des classes de comportements de méthodes d'uplift. Puis, nous apportons une réponse à la deuxième question en proposant une méthode de réduction de l'effet du biais NRA. Le principe de cette méthode repose sur une pondération des personnes du groupe de traitement. Les résultats expérimentaux montrent que notre méthode obtient de meilleures performances qu'en l'absence de correction. À notre connaissance, il s'agit du premier travail qui est centré sur les biais lors de la modélisation de l'uplift.

L'article est organisé comme suit. La section 2 introduit la modélisation de l'uplift. Puis, après avoir précisé le biais NRA et la problématique traitée dans cet article, la section 3 présente notre méthodologie d'évaluation de l'impact du biais NRA. Notre méthode de réduction du biais NRA est décrite en section 4 avant de conclure à la section 5.

2 Contexte : modélisation et évaluation des modèles d'uplift

2.1 Définition

L'uplift est une notion introduite par (Radcliffe et Surry, 1999) ainsi que dans les modèles d'inférence causale de (Rubin, 1974). C'est une technique située au carrefour de 3 communautés (fouille de données, biostatistique et inférence causale) dont plusieurs études commencent à unifier les notations (Zhang et al., 2022). Dans cette section, nous explicitons l'uplift et la difficulté de sa modélisation.

Soit X un ensemble de N individus indexés par $i : 1 \dots N$ où chaque individu est décrit par un ensemble de variables noté $\mathbb{X}$. L'estimation de l'uplift repose sur deux populations : les individus ayant reçu un traitement actif (noté $T = 1$) et ceux n'ayant pas reçu de traitement appelé "groupe de contrôle" (noté $T = 0$). Soit Y une variable binaire de résultat (ou *cible*). On note $Y_i(T = 1)$ le résultat de l'individu i lorsqu'il reçoit le traitement $T = 1$ et $Y_i(T = 0)$ son résultat avec le traitement de contrôle. X_i désigne l'ensemble des valeurs de $\mathbb{X}$ pour l'individu i. L'uplift d'un individu i, noté τ_i, est défini par la différence de son résultat Y_i entre le traitement actif et celui de contrôle :

$$\tau_i = Y_i(T = 1) - Y_i(T = 0) \tag{1}$$

En pratique, il est impossible d'observer simultanément pour un même individu les valeurs $Y_i(T = 1)$ et $Y_i(T = 0)$ et l'équation 1 ne peut pas être utilisée pour calculer l'uplift d'un individu. En revanche, l'uplift d'un individu i peut-être estimé empiriquement en considérant deux groupes d'individus : un ensemble d'individus ayant reçu le traitement et un ensemble

d'individus n'ayant pas reçu de traitement. L'uplift estimé de l'individu i, noté $\hat{\tau}_i$ est alors calculé avec le CATE (Conditional Average Treatment Effect) (Rubin, 1974) ($\mathbb{E}$ note l'espérance mathématique d'une variable) :

$$CATE : \hat{\tau}_i = \mathbb{E}[Y_i(T = 1)|X_i] - \mathbb{E}[Y_i(T = 0)|X_i] \tag{2}$$

En pratique, l'impossibilité de disposer d'échantillons où τ_i serait connu fait qu'il n'est pas possible d'utiliser directement des méthodes usuelles de classification. La section suivante est consacrée à montrer comment l'uplift est modélisé dans la littérature.

2.2 Modélisation de l'uplift

La littérature de l'uplift et celle d'une branche de l'inférence causale se sont rapprochées récemment (Gutierrez et Gérardy, 2017). Les différentes méthodes de modélisation d'uplift sont classées selon une taxonomie composée de 3 principales familles (ou approches) qui sont présentées de manière synthétique ci-dessous.

Approche avec 2 modèles Cette famille de méthodes est la plus classique (Hansotia et Rukstales, 2002). Son principe repose sur la construction, de façon indépendante, de deux modèles prédictifs : l'un sur les données du groupe de traitement (modélisation de $P(Y|X, T = 1)$) et l'autre sur celles du groupe de contrôle (modélisation de $P(Y|X, T = 0)$). L'uplift estimé d'un individu est alors la différence entre la prédiction de la cible sur le groupe de traitement et celle sur le groupe de contrôle. Cette approche a comme avantage la simplicité et la possibilité d'utiliser une grande variété d'algorithmes d'apprentissage supervisé.

Approche par transformation. L'approche par transformation de classe (Jaskowski et Jaroszewicz, 2012) consiste à remplacer le problème de modélisation de l'uplift par un problème d'apprentissage supervisé traditionnel. Pour cela, la variable cible est changée en une nouvelle variable Z comme indiqué à l'équation 3 puis un algorithme d'apprentissage supervisé est utilisé. L'uplift estimé d'un individu i est $\hat{\tau}_i = 2 \times P(Z = 1|X) - 1$

$$Z = \begin{cases} 1, & \text{if } T = 1 \text{ and } Y = 1 \\ 1, & \text{if } T = 0 \text{ and } Y = 0 \\ 0, & \text{sinon.} \end{cases} \tag{3}$$

Plusieurs études (Jaskowski et Jaroszewicz, 2012), (Diemert et al., 2018) montrent la supériorité de cette famille par rapport aux méthodes de l'approche avec 2 modèles.

Approche directe. Le principe des méthodes de cette famille est de modéliser directement l'estimation de l'uplift (estimer directement $\hat{\tau}_i$) en modifiant des méthodes d'apprentissage supervisées classiques afin de les adapter au problème de la modélisation de l'uplift. Citons les méthodes fondées sur les arbres de décision (Rzepakowski et Jaroszewicz, 2011), (Zhao et al., 2017), les k plus proches voisins (Guelman, 2015), la régression logistique (Lo et Pachamanova, 2015) ou encore l'apprentissage par renforcement (Sawant et al., 2018).

2.3 Évaluation des modèles d'uplift

Puisque la valeur réelle de l'uplift pour un individu n'est pas observée, les mesures usuelles d'évaluation des algorithmes d'apprentissage supervisé sont inopérantes. C'est pourquoi, pour évaluer un modèle d'uplift, on définit une mesure de qualité issue du classement des individus selon leur uplift estimé. Les individus étant ainsi classés, un "bon" modèle d'uplift est caractérisé par le fait que les individus du groupe de traitement ayant la valeur $Y = 1$ ont une valeur d'uplift estimée plus grande que celle des individus avec $Y = 0$ et vice versa pour les individus du groupe de contrôle. Les qualités des individus sont ensuite agrégées via la mesure du qini (Radcliffe, 2007). Celle ci est une extension du coefficient de Gini pour le cas de l'uplift. Le qini prend sa valeur dans l'intervalle [-1,1], plus la valeur du qini est grande plus l'impact du traitement estimé est grand.

3 Evaluation de l'uplift en présence de biais

3.1 Problématique

Les méthodes d'uplift formulent l'hypothèse que les données des individus X des différents traitements suivent la même distribution. Or, les données étant de nature observationnelle, on constate en pratique l'existence de biais. Nous étudions ici le biais de "Non-Random Assignment" (NRA) qui est un biais très courant. Celui-ci se produit lorsque le traitement n'est pas distribué aléatoirement sur les individus, autrement dit $P(T = 1|X) \neq P(T = 0|X)$. Le plus souvent, c'est le groupe de traitement qui est biaisé suite à des difficultés éthiques, réglementaires ou économiques lors de l'application du traitement sur un individu. Les données du groupe de contrôle sont généralement plus faciles à collecter.

La problématique de biais entre échantillons est étudiée dans la littérature relative aux études cliniques. Dans ce champ de recherche, le but est d'estimer le "Average Treatment Effect" (ATE) qui est défini par $\mathbb{E}[Y_i(T = 1) - Y_i(T = 0)]$. Même si il ne s'agit pas de modéliser et d'estimer l'uplift, la définition de l'ATE montre que cette mesure partage des caractéristiques du CATE (cf. éq. 2). Pour estimer l'ATE, des méthodes de scoring connues sous le nom de "Propensity Score Methods" (PSM) sont utilisées pour extraire des échantillons équivalents (Austin (2011)). Etant donné que les méthodes d'uplift actuelles supposent que les données ne contiennent pas de biais NRA, les méthodes de PSM sont parfois utilisées avant la modélisation de l'uplift afin d'extraire des données biaisées des échantillons équivalents comme dans le benchmark construit par Olaya et al. (2020).

3.2 Conception d'un protocole d'évaluation de l'uplift

Nous présentons maintenant notre protocole expérimental pour étudier l'impact du biais NRA sur l'uplift. Le principe, pour créer un biais NRA, est de créer des déséquilibres dans les données par rapport aux distributions initiales des variables des données. Un exemple est de se ramener à une situation sans personne d'une certaine catégorie socioprofessionnelle et d'un certain âge. Ce protocole doit satisfaire plusieurs conditions pour quantifier correctement l'impact d'un biais NRA sans introduire un biais lié à l'étude. Tout d'abord, les distributions des variables choisies pour créer le biais NRA doivent être corrélées avec Y (la cible) ou la cible sachant l'uplift sinon le biais NRA sera sans effet sur la modélisation de l'uplift. En revanche, le

choix des valeurs des variables selon lesquelles le déséquilibre est créé est aléatoire afin d'éviter un biais dans la construction des populations $E1$ et $E2$ (celles-ci sont décrites ci-dessous). Les échantillons d'apprentissage auront toujours le même nombre d'individus afin d'éviter un biais lors de l'apprentissage des modèles d'uplift. Le biais NRA est créé uniquement dans le groupe de traitement puisque c'est la situation rencontrée en pratique (cf. section 3.1). Enfin, le taux de biais doit être réglable afin de quantifier l'impact du biais NRA. La figure 1 illustre l'ensemble du processus mis en place.

Plus précisément, deux populations $E1$ et $E2$ sont construites indépendamment de la cible. Pour cela, un ensemble de variables V est choisi. L'ensemble des valeurs issu de la conjonction des variables de V est aléatoirement divisé en deux groupes $C1$ et $C2$, tel que le nombre d'individus vérifiant les valeurs de $C1$ soit le même que celui vérifiant celles de $C2$. $E1$ (resp. $E2$) est défini par les individus vérifiant les valeurs de $C1$ (resp. $C2$). Nous utilisons une validation croisée à 10 volets (pour chaque apprentissage, 10% des individus de $E1$ et $E2$ sont retirés pour constituer l'échantillon test). L'ensemble complet des individus de l'échantillon d'apprentissage, qui est composé de 50% des individus de $E1$ et 50% de $E2$, est non biaisé. Cette situation, notée $b = 50$, donne une valeur de référence du qini puisque sans biais. Le biais NRA est introduit en faisant varier dans le groupe de traitement le nombre d'individus de $E1$ par rapport à celui de $E2$, cette proportion allant de 50% ($b = 50$) pour la situation sans biais, à 100% ($b = 100$), situation la plus biaisée selon le biais NRA. La valeur de b indique le niveau de biais introduit dans les données. Un modèle d'uplift est alors appris sur chaque échantillon défini par une valeur de b. Tous les modèles sont ensuite testés sur le même échantillon test et évalués selon le qini. L'évolution du qini selon b permet d'étudier le comportement d'une méthode d'uplift par rapport au biais NRA.

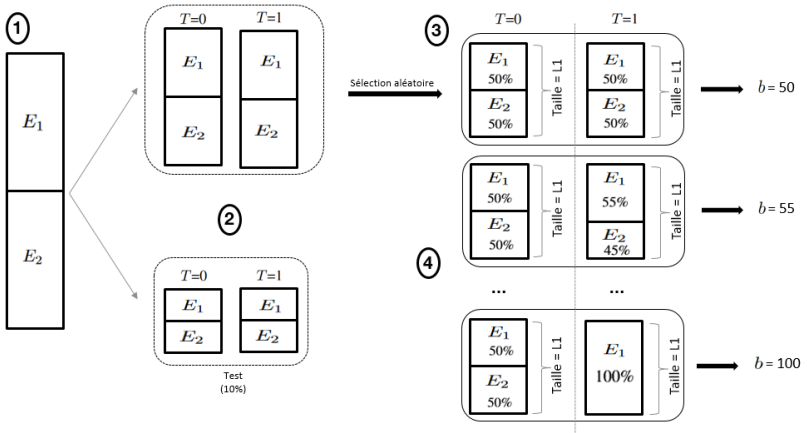


FIG. 1 – *La méthode de génération des échantillons biaisés à partir d'une base de données : (1) Variable(s) V choisies pour créer E1 et E2. (2) Création des ensembles d'apprentissage et de test avec une validation croisée à 10 volets. (3) Échantillonage aléatoire des groupes de traitement de contrôle tel que la taille de E1 et E2 soit équivalente dans chacun des groupes. (4) La taille des groupes de traitement et de contrôle est la même tout au long de la génération des échantillons biaisés.*

3.3 Expérimentation

Bases de données utilisées. Nous utilisons 4 bases de données provenant des domaines de la politique et du marketing ainsi que 4 bases de données synthétiques. Pour toutes ces bases, la variable cible est binaire [1].

1. Criteo (Diemert et al., 2018) : base de données très utilisée pour la prédiction de l'uplift, elle relève du domaine du marketing.

2. Hillstrom : base de données classique du domaine de l'uplift avec deux groupes de traitement et un groupe de contrôle. Pour le traitement, nous utilisons uniquement le groupe de personnes ayant reçu une campagne publicitaire via un mailing pour des produits relatifs aux femmes.

3. Gerber : base de données relevant du domaine de la politique et utilisée pour étudier l'effet de la pression sociale pour l'action d'inciter les électeurs à voter.

4. Retail Hero : base de données du groupe de vente X5, le traitement est l'envoi de SMS pour inciter les consommateurs à augmenter leurs achats.

5. Megafon : base de données synthétique créée pour modéliser l'uplift. Elle est générée par des entreprises de telecom de façon à reproduire les situations rencontrées par ces entreprises.

6. Zenodo : base de données synthétique contenant des motifs trigonométriques spécialement conçus pour l'évaluation de l'uplift. Nous avons utilisé un sous-ensemble de données de 20.000 lignes (données identifiées par la variable trial_id = 1 et trial_id = 2).

7. Synth1 et Synth2 : 2 bases synthétiques que nous avons construites. La génération de données permet de disposer de bases avec des caractéristiques spécifiques (Synth1 est une base avec une forte valeur d'ATE et Synth2 a un faible taux de réponse).

Le tableau 1 précise les caractéristiques de ces bases. Pour toutes ces bases, les groupes de contrôle et de traitement ont le même nombre d'individus. L'indépendance des traitements (i.e. entre $T = 0$ et $T = 1$) est mesurée avec le test C2ST (Lopez-Paz et Oquab, 2017).

Modèle prédicitifs. Nous testons 8 méthodes d'uplift qui sont classiques dans la littérature : (i) 4 méthodes fondées sur la régression logistique et XGboost : 2M_LR et 2M_Xgboost ("approche avec 2 modèles") et CT_LR et CT_Xgboost ("approche par transformation") Jaskowski et Jaroszewicz (2012) ; (ii) ainsi que 4 méthodes de forêts d'arbres relevant de "approche directe" : CTS (Zhao et al. (2017)), KL, ED et Chi (Rzepakowski et Jaroszewicz (2011)).

Détails d'implémentation. Pour chaque base de données (sauf Synth1 et Synth2) et chaque méthode d'uplift, le protocole expérimental est appliqué deux fois avec des contenus différents de V : une fois avec la variable la plus corrélée avec la cible et une fois avec la variable la plus corrélée avec la cible sachant l'uplift (i.e. Y sachant $T = 1$). Pour les bases Synth1 et Synth2, V contient les deux variables de ces bases. De plus, V étant fixé, nous avons répété deux fois l'expérience afin d'obtenir différentes divisions $C1$ et $C2$ des valeurs de V.

1. Les codes, accès aux données et résultats expérimentaux sont disponibles à https ://github.com/MinaWagdi/EvalutationUpliftBiaisNRA

Bases de données	#Lignes	#Variables	Taux de réponse (i.e $P(y=1)$)	ATE	Indépendance des traitements
Criteo	50000	13	0.16	0.08	0.1
Hillstrom	42693	8	0.129	0.04	0.33
Gerber	76419	10	0.34	0.06	0.43
RetailHero	200039	11	0.619	0.033	0.7
Megafon	600000	36	0.0.2	0.04	0.001
Synthetic Zenodo	20000	16	0.3	0.109	0.22
Synth1	40000	2	0.32	0.241	0.197
Synth2	40000	2	0.007	0.00125	0.33

TAB. 1 – *Caractéristiques des bases de données.*

3.4 Résultats

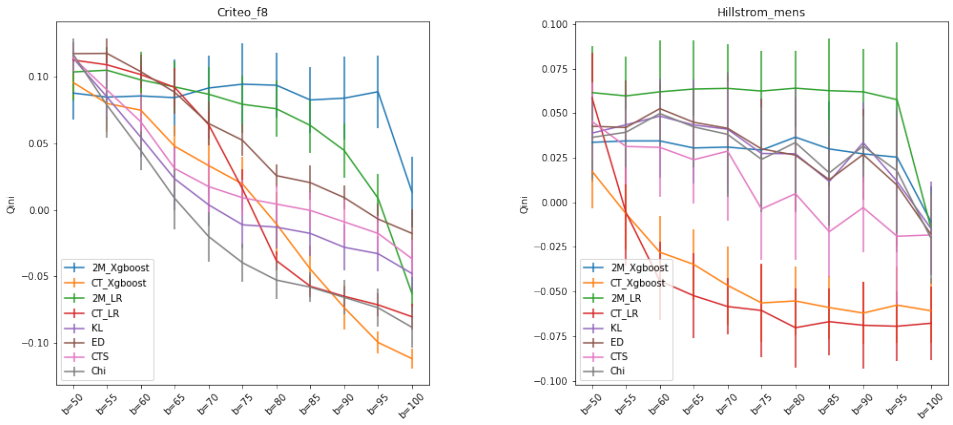
Évolution du qini en fonction de b. La figure 2 illustre les résultats obtenus (pour une raison de place, il n'est pas possible de donner l'ensemble de tous les résultats). On constate que le biais NRA affecte de façon effective les modèles d'uplift et que plus le niveau de biais est élevé, plus la baisse du qini est importante [2]. Pour disposer d'une vue plus globale des résultats, nous calculons pour chaque base et chaque méthode d'uplift le qini moyen, c'est-à-dire la moyenne des valeurs du qini suivant les niveaux de biais allant de $b = 50$ à $b = 100$ (cf. table 2).

Ces expériences permettent de faire ressortir des comportements de méthodes d'uplift face au biais NRA : (i) les modèles les plus résistants au biais NRA parmi ceux testés sont les deux modèles relevant de "approche avec 2 modèles" : le qini se dégrade fortement uniquement lorsque le niveau de biais est élevé ; (ii) les modèles où le qini se dégrade lentement au fur et à mesure que le niveau de biais augmente : il s'agit des modèles fondés sur les arbres de décision (KL, Chi , ED et CTS) et (iii) les modèles fortement affectés par le biais même avec des faibles valeurs de celui-ci : ces modèles relèvent de "approche par transformation" avec la régression logistique et Xgboost.

Moyenne des rangs. Afin de mieux classer les méthodes selon leur résistance au biais NRA, nous regardons le rang obtenu par chaque méthode selon le qini moyen et pour chaque expérience (toutes les divisions utilisées pour les valeurs de V sont prises en compte). La figure 3 indique la moyenne des rangs des 8 méthodes. Cette figure montre que les méthodes les plus résistantes au biais NRA sont la régression logistique avec "approche avec 2 modèles" et la méthode ED fondée sur les arbres de décision. A contrario, les méthodes de la famille "approche par transformation" sont les plus impactées par le bias NRA.

Validité statistique : test de Friedman avec le test post-hoc de Nemenyi. Nous étudions maintenant la validité statistique de la comparaison des méthodes. Pour le choix du test, nous nous appuyons sur l'étude de (Demšar, 2006) qui indique l'utilisation du test de Friedman avec le test post hoc pour comparer deux méthodes entre elles en présence de plusieurs méthodes. À partir des résultats du tableau 2, nous obtenons une carte de chaleur (cf. figure 4) qui indique si le résultat d'une méthode est statistiquement supérieur ou pas au résultat d'une autre méthode. L'hypothèse nulle signifie qu'il n'y a pas de différence significative en performance selon le qini moyen entre deux méthodes sur l'ensemble des bases de données. Avec une valeur de p

2. Lorsque la comparaison avec l'état de l'art est possible, les valeurs des qinis obtenus sans biais ($b = 50$) sont celles habituellement rapportées dans la littérature (Diemert et al., 2018)



(a) **(b)**

FIG. 2 – *Qini obtenu par les méthodes d'uplift selon le niveau de biais NRA dans la base Criteo avec les variables f8 (gauche) et la base Hillstrom avec les variables mens (droite).*

(p-valeur) plus petite que 0.05, l'hypothèse nulle est rejetée (en vert à la figure 4). Cette figure montre que "approche avec 2 modèles" ainsi que la méthode ED sont les plus résistantes au biais NRA.

	Approche à 2 modèles		Approche par transformation		Arbres de décisions			
	Xgboost	LR	Xgboost	LR	KL	ED	CTS	Chi
Criteo_f2	6.6(1.7)	**7.2(1.6)**	0.2(1.9)	1.9(1.2)	0.6(1.4)	4.9(1.3)	2.1(1.5)	-2.1(1.6)
Criteo_f8	**8.1(2.6)**	6.3(2.0)	0.1(1.7)	1.7(1.0)	1.2(1.6)	5.2(1.2)	2.4(1.6)	-1.4(1.6)
Hillstrom_mens	2.7(2.1)	**5.5(2.6)**	-4.1(2.0)	-4.6(2.2)	2.8(2.6)	2.9(2.5)	1.0(2.8)	2.8(2.6)
Hillstrom_newbie	2.8(2.2)	**6.2(2.7)**	0.1(2.1)	2.4(1.9)	4.2(2.2)	4.3(2.5)	4.3(2.5)	4.1(2.4)
Gerber_p2002	-2.4(2.0)	**1.1(1.1)**	-2.1(1.5)	-0.4(1.2)	-1.5(1.8)	-0.9(1.5)	-0.1(1.7)	-1.5(1.8)
Gerber_p2004	-2.1(2.0)	**0.8(1.1)**	-1.8(1.7)	-1.2(1.3)	-1.7(1.8)	-1.5(1.9)	-0.6(1.9)	-1.9(1.8)
retailHero_age	0.7(0.4)	**1.2(0.3)**	0.3(0.4)	0.8(0.4)	0.8(0.3)	0.9(0.3)	0.9(0.4)	0.9(0.3)
retailHero_trNum	0.8(0.4)	**1.2(0.3)**	0.4(0.3)	1.1(0.4)	0.7(0.4)	0.7(0.4)	0.6(0.4)	0.7(0.4)
Megafone_X16	**17.8(0.5)**	3.5(0.4)	8.6(0.6)	3.2(0.4)	13.2(0.5)	13.7(0.5)	11.6(0.7)	13.2(0.4)
Megafone_X21	**18.2(0.4)**	3.5(0.4)	12.0(0.4)	2.4(0.5)	13.9(0.5)	14.0(0.6)	10.7(0.8)	13.7(0.5)
zenodoSynth_X10	9.7(1.8)	12.6(1.9)	7.0(2.2)	12.1(1.5)	12.8(1.9)	**13.0(1.9)**	10.6(2.6)	12.8(1.8)
zenodoSynth_X31	9.8(2.4)	12.2(2.0)	6.6(2.0)	12.0(1.9)	12.7(1.9)	**13.2(2.0)**	10.2(2.2)	12.8(1.8)
Synth1	7.0(0.9)	0.9(1.6)	1.7(0.9)	-2.9(1.3)	**9.7(1.2)**	8.8(1.6)	8.7(1.2)	9.6(1.8)
Synth2	9.8(0.1)	1.9(0.1)	8.1(0.5)	1.1(0.2)	9.7(0.1)	9.6(0.2)	8.7(0.1)	**9.9(0.2)**

TAB. 2 – *Qini moyen (multiplé par 100), sa variance (indiquée entre parenthèses) suivant les bases et les méthodes d'uplift. Le nom de la base est suivi par les noms des variables de V utilisées pour générer le biais NRA (pour une raison de place, les résultats sont donnés pour une seule division des valeurs de V).*

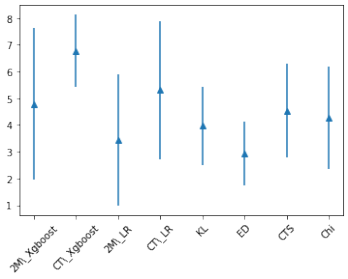


FIG. 3 – *Moyenne des rangs des méthodes d'uplift.*

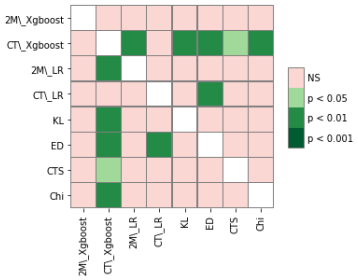


FIG. 4 – *Carte de chaleur pour visualiser la comparaison entre les méthodes d'uplift. Une valeur de p plus petite que 0.05 signifie que l'hypothèse nulle est rejetée.*

4 Méthode pour réduire le biais

Pondération du groupe de traitement avec le rapport des gaussiennes. Afin de réduire le biais NRA, nous nous inspirons de la littérature du *Domain Adaptation* où des méthodes de pondération sont proposées afin de rééquilibrer la population d'un échantillon par rapport à celle de l'autre. Notre idée est de faire ressembler la population biaisée à celle qui serait non biaisée. Notre méthode consiste à appliquer la technique de pondération de (Shimodaira, 2000) pour pondérer les individus du groupe de traitement suivant leur importance dans le groupe de contrôle. Soit $w(X_i)$ le poids de l'individu X_i. Ce poids est le rapport de la probabilité que l'individu appartienne au groupe de contrôle par rapport à la probabilité qu'il appartienne au groupe de traitement.

Pour estimer chacune de ces probabilités, nous formulons l'hypothèse classique (Shimodaira, 2000) que chaque groupe de traitement est issu d'une distribution gaussienne multivariée, c'est-à-dire, $P(X|T = t) = \mathcal{N}\left(X_i \mid \mu_{T=t}, \sigma^2_{T=t}\right)$. Il s'ensuit que le poids de X_i est équivalent à :

$$w(X_i) = \mathcal{N}\left(X_i \mid \mu_{T=0}, \sigma^2_{T=0}\right) / \mathcal{N}(X_i \mid \mu_{T=1}, \sigma^2_{T=1})$$

Afin d'éviter des poids de très grandes valeurs ou, au contraire, proches de zéro, nous ajoutons 2 paramètres a et b pour les borner tels que :

$$w_{borne}(X_i) = \begin{cases} a, & \text{if } w(X_i) < a \\ b, & \text{if } w(X_i) > b \\ w(X_i), & \text{sinon.} \end{cases} \tag{4}$$

Nous évaluons expérimentalement cette méthode de réduction du biais NRA. Nous testons les méthodes d'uplift relevant des familles "approche avec 2 modèles" et "approche par transformation" car ces méthodes reposent sur des algorithmes d'apprentissage traditionnels où il est possible d'intégrer en entrée des poids pour chaque ligne (une ligne étant ici un individu). En revanche, les méthodes de la famille "approche directe" ne peuvent pas prendre en

compte des poids et c'est pourquoi nous ne les utilisons pas. Après une étude préliminaire sur les valeurs de a et b, la pondération de chaque individu du groupe de traitement est faite avec $a = 0.8$ et $b = 15$. Les résultats montrent une hausse du qini moyen avec les méthodes de "approche par transformation" mais pas avec celles de "approche avec 2 modèles". Le tableau 3 détaille ces résultats avec les méthodes de "approche par transformation" : CT_LR_pond (resp. CT_Xgboost_pond) correspond à la méthode CT_LR (resp. CT_Xgboost) mais avec notre technique de pondération ("pond" signifiant "pondération").

	Ref.qini	CT_LR	CT_LR_pond	Ref.qini	CT_Xgboost	CT_Xgboost_pond
Criteo_f2	11.1(0.9)	1.9(1.2)	**10.2(0.9)**	9.1(2.6)	0.2(1.9)	**8.9(0.9)**
Criteo_f8	11.2(1.0)	1.7(1.0)	**10.0(0.8)**	9.6(1.2)	0.1(1.7)	**8.8(1.0)**
Gerber_p2002	0.8(1.6)	-0.4(1.2)	**1.2(0.9)**	-1.9(2.0)	-2.1(1.5)	**-1.2(2.0)**
Gerber_p2004	1.1(1.4)	-1.2(1.3)	**1.2(1.1)**	-1.6(2.1)	-1.8(1.7)	**-1.5(2.0)**
Hillstrom_mens	5.9(2.5)	-4.6(2.2)	**2.1(2.4)**	1.7(2.1)	-4.1(2.0)	**0.3(2.3)**
Hillstrom_newbie	6.3(1.7)	2.4(1.9)	**5.1(2.1)**	1.7(1.9)	0.1(2.1)	**1.5(2.7)**
Megafone_X16	3.2(0.5)	**3.2(0.4)**	2.1(0.5)	17.3(0.6)	8.6(0.6)	**11.8(0.5)**
Megafone_X21	3.2(0.4)	**2.4(0.5)**	**2.4(0.5)**	17.2(0.5)	12.0(0.4)	**13.1(0.4)**
Synth1	-0.2(3.4)	-2.9(1.3)	**1.2(2.1)**	2.5(2.4)	1.7(0.9)	**3.8(0.6)**
Synth2	1.8(0.0)	1.1(0.2)	**1.8(0.0)**	10.7(0.0)	8.1(0.5)	**8.3(0.7)**
retailHero_age	1.2(0.4)	**0.8(0.4)**	0.7(0.5)	0.6(0.4)	**0.3(0.4)**	0.2(0.5)
retailHero_trNum	1.2(0.3)	**1.1(0.4)**	0.5(0.8)	0.7(0.4)	**0.4(0.3)**	0.3(0.4)
zenodoSynth_X10	12.3(1.3)	**12.1(1.5)**	**12.1(1.9)**	8.0(3.1)	7.0(2.2)	**7.5(2.1)**
zenodoSynth_X31	11.7(2.3)	12.0(1.9)	**12.2(1.9)**	6.9(1.9)	6.6(2.0)	**7.4(2.0)**

TAB. 3 – *Qini moyen (multiplié par 100) et sa variance (indiquée entre parenthèses) avec CT_LR et CT_Xgboost sans et avec pondération. Le nom de la base de données est suivi par les noms des variables de V utilisées pour générer le biais NRA. Ref.qini indique le qini de référence c'est-à-dire le qini d'une méthode sans biais et sans pondération (i.e., $b = 50$).*

Discussion. Il est possible d'expliquer le faible impact de la pondération sur les méthodes de la famille "approche avec 2 modèles". En effet, le biais NRA n'affecte pas la distribution de la cible connaissant les populations E_1 et E_2 (définies à la section 3.2) dans le groupe de traitement. Les estimations des probabilités $P(Y|T = 1, X)$ et $P(Y|T = 0, X)$ sont alors peu perturbées et les performances avec ou sans pondération sont similaires. Il en va différemment avec les méthodes de "approche par transformation" qui estiment directement Z en se fondant sur l'hypothèse que les groupes de traitement et de contrôle sont équivalents. Or, cette hypothèse n'est plus valable avec le biais NRA et l'estimation de Z devient biaisée. Dans cette situation, la pondération du groupe de traitement améliore l'estimation de Z et donc l'uplift.

Test de Wilcoxon. Nous étudions la validité statistique de notre méthode de pondération pour réduire le biais NRA. Pour cela, nous utilisons le test de Wilcoxon (Wilcoxon, 1945). Ce test est utilisé pour comparer deux méthodes (dans notre cas, une méthode d'uplift sans et avec pondération) sur plusieurs bases de données Demšar (2006). Les résultats sont à la table 4. Avec un seuil de confiance de 95%, les p-valeurs obtenues indiquent que l'hypothèse nulle est rejetée pour "approche par transformation" et n'est pas rejetée pour "approche avec 2 modèles". Autrement dit, notre méthode de pondération amène une amélioration significative pour les méthodes de "approche par transformation" mais pas pour celles de "approche avec 2 modèles" comme on pouvait s'y attendre suite à notre discussion ci-dessus.

Méthodes	p-valeur
CT_LR vs CT_LR_pond	0.013
CT_Xgboost vs CT_Xgboost_pond	0.0011

Méthodes	p-valeur
2M_LR vs 2M_LR_pond	0.48
2M_Xgboost vs 2M_Xgboost_pond	0.09

TAB. 4 – *p-valeur obtenues avec le test de Wilcoxon.*

5 Conclusion et perspectives

Dans cet article nous avons étudié l'impact du biais NRA lors de la modélisation des méthodes d'uplift. À notre connaissance, ce travail est le premier qui se concentre sur l'étude des effets des biais sur les modèles actuels d'uplift. Nous avons conçu un protocole expérimental qui permet, en faisant varier le niveau de biais sur le groupe de traitement, d'étudier l'impact du biais NRA sur les méthodes d'uplift et de dégager des classes de comportements pour ces méthodes. Inspiré de la littérature du *Domain Adaptation*, nous avons proposé une méthode de réduction de l'effet du biais NRA reposant sur une pondération des individus du groupe de traitement. Les résultats expérimentaux sur 8 jeux de données montrent que notre méthode apporte une amélioration significative des performances de l'estimation de l'uplift pour les méthodes "approche par transformation".

Ce travail ouvre plusieurs perspectives. La possibilité de correction de l'effet du biais NRA pour les méthodes "approche par transformation" suggère de concevoir de nouvelles méthodes relevant de cette famille. D'autre part, il sera utile d'étudier d'autres biais comme (i) les biais de déploiement qui surviennent lorsque les modèles d'uplift sont appliqués sur des populations différentes et (ii) les biais de non-réponse qui forment un véritable défi pour la modélisation de l'uplift avec des données observationnelles.

Références

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research 46*, 399 – 424.

Berk, R. A., G. K. Smyth, et L. W. Sherman (1988). When random assignment fails : Some lessons from the minneapolis spouse abuse experiment. *Journal of Quantitative Criminology 4*(3), 209–223.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res. 7*, 1–30.

Diemert, E., A. Betlei, C. Renaudin, et M.-R. Amini (2018). A large scale benchmark for uplift modeling. In *KDD*.

Guelman, L. A. (2015). Optimal personalized treatment learning models with insurance applications (doctoral thesis). Universitat de Barcelona.

Gutierrez, P. et J.-Y. Gérardy (2017). Causal inference and uplift modelling : A review of the literature. In *International Conference on Predictive Applications and APIs*. PMLR.

Hansotia, B. et B. Rukstales (2002). Incremental value modeling. *Journal of Interactive Marketing 16*(3), 35–46.

Jaskowski, M. et S. Jaroszewicz (2012). Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, Volume 46.

Lo, V. et D. Pachamanova (2015). From predictive uplift modeling to prescriptive uplift analytics : A practical approach to treatment optimization while accounting for estimation risk. *Journal of Marketing Analytics 3*(2), 79–95.

Lopez-Paz, D. et M. Oquab (2017). Revisiting classifier two-sample tests. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Olaya, D., K. Coussement, et W. Verbeke (2020). A survey and benchmarking study of multi-treatment uplift modeling. *Data Mining and Knowledge Discovery 34*, 273–308.

Radcliffe, N. (2007). Using control groups to target on predicted lift : Building and assessing uplift model. *Direct Marketing Analytics Journal*, 14–21.

Radcliffe, N. et P. Surry (1999). Differential response analysis : Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV*.

Radcliffe, N. J. et P. D. Surry (2011). Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 1–33.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Rzepakowski, P. et S. Jaroszewicz (2011). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems 32*, 303–327.

Sawant, N., C. B. Namballa, N. Sadagopan, et H. Nassif (2018). Contextual multi-armed bandits for causal marketing. *CoRR abs/1810.01859*.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference 90*(2), 227–244.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*(6), 80–83.

Zhang, W., J. Li, et L. Liu (2022). A unified survey of treatment effect heterogeneity modelling and uplift modelling. *ACM Comput. Surv. 54*(8), 162 :1–162 :36.

Zhao, Y., X. Fang, et D. Simchi-Levi (2017). Uplift modeling with multiple treatments and general response types. *CoRR abs/1705.08492*.

Summary

Uplift modeling measures the impact of an action (marketing, medical treatment) on a person's behavior. Uplift prediction is based on groups of people who have received different treatments. These groups are assumed to be equivalent. However, in practice, we observe that there are biases between these groups. In this paper, we propose a protocol to evaluate and study the impact of non-random assignment bias (NRA) on the performance of the main uplift methods. Then we present a weighting technique to reduce the effect of NRA bias. Experimental results show an improvement in the performance of the uplift methods.