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

pour la classification, il y a très peu de propositions pour la modélisation de l'uplift (Zhao et al., 2020). Cette observation peut s'expliquer par le fait que l'uplift crée de nouveaux défis tels que l'impossibilité d'observer deux résultats de traitement pour un même individu. La conception de méthodes pour l'uplift nécessite de surmonter cette difficulté. Cet article vise à répondre au besoin de méthodes de sélection de variables pour l'uplift.

Nous présentons une méthode de sélection de variables sans paramètres pour la modélisation de l'uplift, fondée sur une approche bayésienne. En s'inspirant des idées de la littérature sur la sélection des variables, nous décrivons tout d'abord une méthode de discrétisation automatique des variables pour la modélisation de l'uplift que nous appelons UMODL (pour Uplift MODL). UMODL s'appuie sur le critère bayésien MODL (Minimum Optimized Description Length) (Boullé, 2006) que nous avons étendu au problème de l'uplift. Ensuite, sur la base d'UMODL, nous présentons UMODL feature selection (UMODL-FS en abrégé) une méthode de sélection de variables pour l'uplift. Cette approche est présentée de façon plus approfondie dans Rafla et al. (2022).

La section 2 présente le contexte et l'état de l'art. Nous décrivons UMODL en section 3 et UMODL-FS et les expérimentations en section 4. Nous concluons dans la section 5.

2 Contexte et état de l'art

2.1 Modélisation de l'uplift

Définition de l'uplift. L'uplift est une notion introduite par Radcliffe et Surry (1999) et définie dans les modèles d'inférence causale de Rubin (1974) comme le *Individual Treatment Effect*. La littérature sur la modélisation de l'uplift et une branche de la littérature sur l'inférence causale se sont récemment rapprochées (Gutierrez et Gérardy, 2016). Nous présentons maintenant la notion d'uplift.

Soit D un groupe de N individus indexés par $n : 1 \dots N$ où chaque individu est décrit par un ensemble de variables $\mathbb{X}$. X_n désigne l'ensemble des valeurs de $\mathbb{X}$ pour l'individu n. Soit T une variable indiquant si un individu a reçu ou non un traitement.

La modélisation de l'uplift repose sur deux groupes : les individus ayant reçu un traitement (noté $T = 1$) et ceux sans traitement (noté $T = 0$). Soit Y la variable cible (par exemple, l'achat ou non d'un produit). On note $Y_n(T = 1)$ le résultat d'un individu n lorsqu'il a reçu un traitement et $Y_n(T = 0)$ son résultat sans traitement. L'uplift d'un individu n, notée τ_n, est définie comme : $\tau_n = Y_n(T = 1) - Y_n(T = 0)$. La principale difficulté réside dans le fait que la valeur d'uplift n'est pas directement mesurable, c'est-à-dire que pour chaque individu, nous pouvons soit observer $Y_n(T = 1)$, soit $Y_n(T = 0)$ mais nous ne pouvons pas observer simultanément les deux résultats. Cependant, le gain τ_n peut être estimé empiriquement en considérant deux groupes : un groupe de traitement (individus ayant reçu un traitement) et un groupe de contrôle (individus n'en ayant pas reçu). L'uplift estimé d'un individu n, désignée par $\hat{\tau}_n$, est alors la différence entre les taux de réponse des deux groupes et est calculée en utilisant la méthode CATE [1] (Conditional Average Treatment Effect) (Rubin, 1974) : CATE : $\hat{\tau}_n = \mathbb{E}[Y_n(T = 1)|X_n] - \mathbb{E}[Y_n(T = 0)|X_n]$.

1. Les termes *effet du traitement* et *uplift* traitent la même notion. CATE est une estimation de l'uplift et nous utilisons "CATE" pour parler des valeurs estimées d'uplift.

Comme la valeur réelle de τ_n ne peut être observée, il est impossible d'utiliser directement des algorithmes d'apprentissage automatique tels que la régression pour déduire un modèle permettant de prédire τ_n.

2.2 Sélection de variables pour les modèles d'uplift

L'accessibilité des ensembles de données de haute dimension avec des centaines de variables rend l'utilisation de techniques de sélection de variables cruciale pour les tâches d'apprentissage automatique et l'uplift. L'objectif des techniques de sélection de variables est de sélectionner un sous-ensemble de variables qui pourraient décrire efficacement les données tout en éliminant les variables non pertinentes (Guyon et Elisseeff, 2003). Cela peut améliorer de manière significative les performances des modèles et le temps de calcul. En ce qui concerne la modélisation de l'uplift, les études portant sur la sélection des variables sont très limitées. À notre connaissance, seuls deux articles de recherche traitent de ce défi.

Zhao et al. (2020) proposent des méthodes de sélection de variables pour l'uplift de type filtres et intégrées. Le principe est de supprimer les variables qui ne sont pas corrélées à la variable cible ou à l'uplift. Les méthodes de type filtres sont utilisées dans une étape de prétraitement indépendamment d'un modèle d'uplift, tandis que les méthodes intégrées effectuent la sélection des variables pendant l'apprentissage d'un modèle et sont spécifiques à un algorithme d'uplift. Dans Zhao et al. (2020), les méthodes de filtrage présentées sont les *méthodes à bins* (inspirées de (Rzepakowski et Jaroszewicz, 2012)), *F-filtre* et *LR-filtre*. Les expériences menées dans Zhao et al. (2020) montrent que les méthodes de filtrage basées sur les bins ont les meilleures performances, tandis que les méthodes *F-filter*, *LR-filter* et intégrées ont des performances médiocres. Un autre article très récent (Hu, 2022) utilise certaines des méthodes de filtrage données dans Zhao et al. (2020) ainsi qu'un coefficient de corrélation pour éliminer les variables redondantes.

2.3 L'approche MODL

L'approche MODL (Minimum Optimized Description Length) est une approche bayésienne non paramétrique pour la discrétisation et l'estimation des probabilités conditionnelles (Boullé, 2006). Elle est basée sur le principe de la longueur de description minimale (LDM) (Grünwald, 2007). L'approche MODL consiste à définir un critère pour un modèle de discrétisation et, à l'aide d'un algorithme de recherche, l'approche MODL peut noter tous les modèles de discrétisation possibles et sélectionner celui qui a le meilleur score.

3 UMODL

Cette section présente UMODL, un nouveau critère pour la modélisation de la discrétisation de l'uplift.

3.1 Critère UMODL

Bien que MODL exploite correctement la discrétisation pour l'estimation de la densité, il n'est pas adapté à la modélisation de l'uplift. En effet, l'uplift porte sur deux groupes de

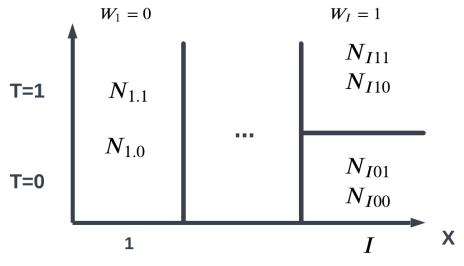


FIG. 1 – Le modèle de discrétisation est décrit par un ensemble de paramètres : le nombre d'intervalles I, La présence d'un effet de traitement ($W_i = 1$) ou l'absence d'un effet de traitement ($W_i = 0$) les fréquences des intervalles N_i et les fréquences des classes dans les intervalles ($N_{i.j}$ ou N_{itj}).

traitement et l'estimation des probabilités conditionnelles de la variable cible Y étant donné un attribut X dépend également de la variable de traitement T.

Nous présentons maintenant le nouveau critère que nous proposons pour définir le meilleur modèle de discrétisation pour l'uplift. Soit M un modèle de discrétisation de l'uplift et D les données. D'un point de vue bayésien, le meilleur modèle de discrétisation de l'uplift est trouvé en maximisant la probabilité postérieure du modèle étant donné les données $P(M|D)$. Considérons la règle de Bayes :

$$P(M \mid D) = \frac{P(M)P(D \mid M)}{P(D)} \tag{1}$$

$P(D)$ étant constant, maximiser $P(M|D)$ est équivalent à maximiser $P(M)P(D|M)$.

Nous définissons un modèle de discrétisation de l'uplift M pour une variable X par le nombre d'intervalles I, les bornes des intervalles, la présence ou l'absence d'un effet du traitement, les fréquences des classes par intervalle ou pour chaque traitement par intervalle. En d'autres termes, un modèle M est défini par la hiérarchie des paramètres (cf. Fig. 1) :

$$\{I, \{N_i\}, \{W_i\}, \{N_{i.j}\}_{W_i=0}, \{N_{itj}\}_{W_i=1}\}$$

On définit $C(M)$ le coût d'un modèle M de discrétisation de l'uplift par : $C(M) = -\log\big(P(M) \times P(D|M)\big)$. En prenant le log négatif, on transforme le problème de maximisation en un problème de minimisation. M est optimal si $C(M)$ est minimal.

$$\begin{aligned} C(M) = \log N + \log \binom{N+I-1}{I-1} + I \times \log 2 \\ + \sum_{i=1}^{I} (1 - W_i) \log \binom{N_i + J - 1}{J-1} + \underbrace{\sum_{i=1}^{I} (1 - W_i) \log \frac{N_i!}{N_{i.1}!..N_{i.J}!}}_{Likelihood} \\ + \sum_{i=1}^{I} W_i \sum_t \log \binom{N_{it.} + J - 1}{J-1} + \underbrace{\sum_{i=1}^{I} W_i \sum_t \log \frac{N_{it.}!}{N_{it1}!..N_{itJ}!}}_{Likelihood} \end{aligned} \tag{2}$$

Estimation de l'uplift L'estimation du CATE pour chaque intervalle est simple. Comme le montre la Fig. 1, en supposant une variable cible binaire Y et étant donné $W_i = 1$, nous avons

$P_i(Y = 1|T = 1) = N_{i11}/(N_{i11} + N_{i01})$ et $P_i(Y = 1|T = 0) = N_{i10}/(N_{i10} + N_{i00})$, donc $CATE_i = P_i(Y = 1|T = 1) - P_i(Y = 1|T = 0)$. Pour les intervalles avec $W_i = 0$, $CATE_i$ est considéré comme non significatif.

L'algorithme de recherche de l'optimum est de type glouton [2] il est décrit dans l'article Rafla et al. (2022). Dans cet article, nous avons aussi montré de façon expérimentale que UMODL est un estimateur efficace et précis de l'uplift et qui ne sur-apprend pas.

4 Sélection des variables avec UMODL

Description de la sélection des variables UMODL. Nous définissons la mesure de divergence de l'effet du traitement sur les intervalles trouvés par UMODL $imp.s(X)$ comme suit. En supposant que $p_i = P_i(Y = 1|T = 1)$ et $q_i = P_i(Y = 1|T = 0)$, nous définissons :

$$imp.s(X) = \begin{cases} \sum_{i=1}^{I} \frac{N_i}{N} D(p_i : q_i), & \text{if } I > 1 \\ 0, & \text{sinon.} \end{cases} \tag{3}$$

où la mesure de divergence de distribution D est la distance euclidienne.

La méthode UMODL-FS consiste à :

1. Étant donné une variable X, nous appliquons la méthode de discrétisation UMODL présentée dans l'article Rafla et al. (2022)
2. Calcul pour X d'un score d'importance (cf. équation 3) désigné par $imp.s(X)$: qui est la mesure de divergence de l'effet du traitement sur les intervalles trouvés.
3. Nous répétons ces étapes pour chaque variable de l'ensemble de données. La découverte d'un seul intervalle par UMODL fait référence à des variables non pertinentes.
4. Toutes les variables avec $imp.s(X) > 0$ sont considérées comme pertinentes pour l'estimation de l'uplift, tandis que toute variable avec $imp.s(X) = 0$ est éliminée.

Lorsque UMODL ne trouve qu'un seul intervalle pour une variable, cela signifie qu'il n'y a qu'une seule distribution pour toutes les instances et donc une variable non informative (i.e . $imp.s(X) = 0$). Contrairement aux méthodes de sélection de variables de la littérature, notre approche ne nécessite pas de paramètres à définir, et il n'est pas nécessaire de donner le nombre de variables à conserver ou à supprimer.

Protocole expérimental. Pour comparer UMODL-FS aux méthodes de sélection de variables de l'uplift de l'état de l'art (cf. section 2.2), nous avons conçu le protocole expérimental suivant :

1. Pour chaque ensemble de données, nous générons onze variantes de celui-ci, chacune avec un nombre total allant de 0 à 100 de variables de bruit. Les variables de bruit sont échantillonnées dans $\mathcal{N}(0, 1)$ pour chacun des groupes de traitement et de contrôle.
2. Pour chaque variante, nous appliquons les méthodes de sélection des variables suivantes : (a) KL-filter (b) Chi-filter (c) ED-filter (d) LR-filter (e) F-filter (f) UMODL-FS.

   Pour les méthodes KL-filter, Chi-filter et ED-filter, nous fixons le nombre d'intervalles à 10.

2. Notre implémentation est fournie sur `https://github.com/MinaWagdi/UMODL`

3. Pour avoir le même nombre de variables pour chaque méthode de sélection des variables et effectuer une comparaison équitable, nous choisissons les M variables les plus importantes, où M est le nombre de toutes les variables jugées informatives par UMODL-FS.
4. Avec ces ensembles de variables, nous construisons des modèles d'uplift : une approche à deux modèles avec régression logistique (Hitsch et Misra, 2018) et X-Learner avec régression linéaire (Jacob, 2021).
5. Le processus d'apprentissage se fait par validation croisée stratifiée à 10 volets. Les échantillons de test sont utilisés pour évaluer les performances des modèles d'uplift construits à partir des variables sélectionnées.
6. La métrique du coefficient de qini (Devriendt et al., 2020) est utilisée pour évaluer la performance du modèle d'uplift. 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.

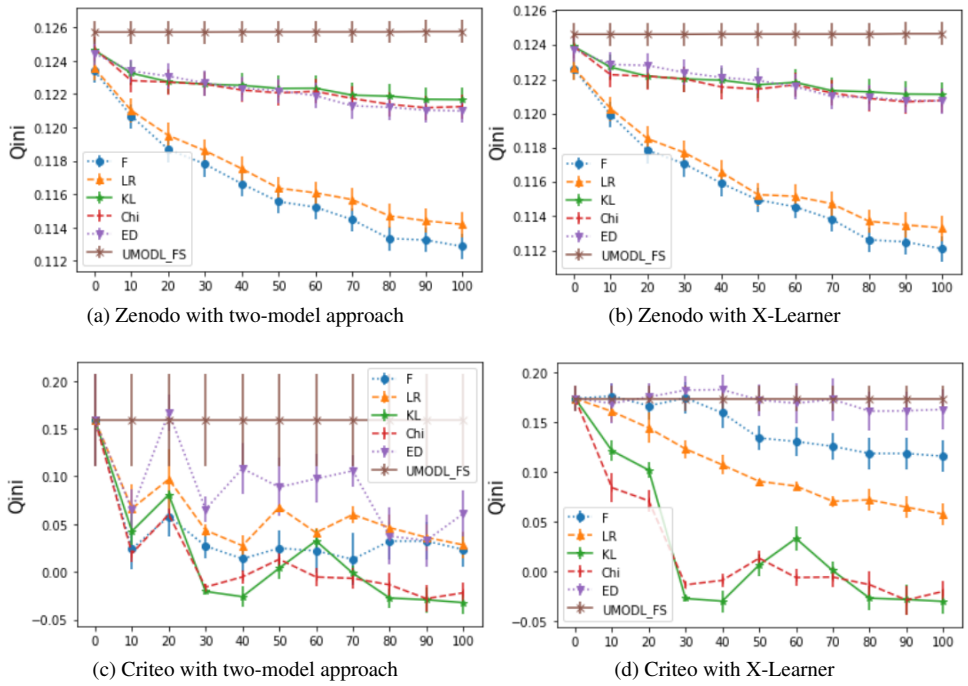
Jeux de données. Les expériences sont menées sur deux ensembles de données continues disponibles publiquement et habituelles dans la communauté uplift : Criteo dataset (Diemert et al., 2018) un véritable jeu de données à grande échelle construit en assemblant les données résultant de plusieurs tests dans la publicité et Zenodo synthetic dataset [3] un jeu de données créé pour évaluer les méthodes de sélection de variables pour la modélisation de l'uplift.
Résultats. La Fig. 2 présente les résultats de l'utilisation d'UMODL-FS pour la modélisation de l'uplift. Dans toutes les expériences, UMODL-FS sélectionne l'ensemble des variables conduisant au modèle d'uplift avec le meilleur qini (donc le meilleur modèle d'uplift) quelle que soit l'approche d'uplift utilisée. Il est remarquable de constater que plus on ajoute de variables de bruit, plus la différence de qini entre UMODL-FS et les autres méthodes de sélection de variables augmente.

UMDOL-FS ne sélectionne jamais une variable de bruit. Cela illustre la capacité évidente d'UMODL-FS à supprimer les variables de bruit. A contrario, toutes les autres méthodes sélectionnent des variables de bruit et le pourcentage de variables de bruit sélectionnées augmente avec le nombre de variables de bruit ajoutées.

5 Conclusion et travaux futurs

Dans cet article, nous avons proposé une nouvelle approche bayésienne non paramétrique pour la sélection des variables. Nous avons défini UMODL-FS, une méthode de sélection de variables pour l'uplift. Les expériences démontrent que UMODL-FS élimine correctement les variables non pertinentes et surpasse clairement les méthodes de pointe en fournissant des modèles de d'uplift avec le qini le plus élevé et le plus stable. La méthode est sans paramètre, ce qui la rend facile à utiliser. Ce travail ouvre plusieurs perspectives. Il est prometteur d'étudier cette approche dans le cas de traitements multiples et de résultats multiples. D'autre part, comme les arbres de décision sont construits sur des variables discrétisées, cette approche peut être étudiée pour développer des algorithmes de modélisation de l'uplift fondés sur des arbres.

3. https ://doi.org/10.5281/zenodo.3653141



(a) Zenodo with two-model approach

(b) Zenodo with X-Learner

(c) Criteo with two-model approach

(d) Criteo with X-Learner

FIG. 2 – Moyenne de qini et sa variance en fonction du nombre de variables de bruit ajoutées. L'axe des X indique le nombre total de variables bruit ajoutées. L'axe Y représente les valeurs de qini obtenues par les modèles d'uplift.

Références

Boullé, M. (2006). MODL : A bayes optimal discretization method for continuous attributes. *Mach. Learn. 65*(1), 131–165.

Devriendt, F., J. Van Belle, T. Guns, et W. Verbeke (2020). Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Diemert, E., A. Betlei, C. Renaudin, et M.-R. Amini (2018). A Large Scale Benchmark for Uplift Modeling. In *KDD*, London, United Kingdom.

Grünwald, P. (2007). *The minimum description length principle*. Adaptive computation and machine learning. MIT Press.

Gutierrez, P. et J.-Y. Gérardy (2016). Causal inference and uplift modelling : A review of the literature. In *PAPIs*.

Guyon, I. et A. Elisseeff (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

Hitsch, G. J. et S. Misra (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. *Randomized Social Experiments eJournal*.

Hu, J. (2022). Customer feature selection from high-dimensional bank direct marketing data for uplift modeling. *Journal of Marketing Analytics*, 1–12.

Jacob, D. (2021). Cate meets ml. *Digital Finance 3*(2), 99–148.

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

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*.

Rafla, M., N. Voisine, B. Crémilleux, et M. Boullé (2022). A non-parametric bayesian approach for uplift discretization and feature selection. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rzepakowski, P. et S. Jaroszewicz (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowl. Inf. Syst. 32*(2), 303–327.

Zhao, Y., X. Fang, et D. Simchi-Levi (2017). Uplift modeling with multiple treatments and general response types. In N. V. Chawla et W. Wang (Eds.), *SIAM Int. Conf. on Data Mining, Houston, Texas, USA, April 27-29, 2017*, pp. 588–596. SIAM.

Zhao, Z., Y. Zhang, T. Harinen, et M. Yung (2020). Feature selection methods for uplift modeling. *CoRR abs/2005.03447*.

Summary

Uplift modeling aims to estimate the incremental impact of a treatment, such as a marketing campaign or a drug, on an individual's outcome. Bank or Telecom uplift data often have hundreds to thousands of features. In such situations, detection of irrelevant features is an essential step to reduce computational time and increase model performance. We present a parameter-free feature selection method for uplift modeling founded on a Bayesian approach. we describe a parameter-free feature selection method for uplift. Experiments show that the new method both removes irrelevant features and achieves better performances than state of the art methods.