Prédiction des événements rares : application à la prédiction du clic.

Slimane Makhlouf*, Avner Bar-Hen**, François-Xavier Jollois***

*Velvet Consulting
slimane.makhlouf@hotmail.fr,
**CNAM
avner.bar-hen@lecnam.net
***LIPADE, Université Paris Cité
francois-xavier.jollois@u-paris.fr

Résumé. Lors de l'affichage de publicités en ligne, un système d'enchères en temps réel est mis en place pour choisir l'annonceur. Du côté de ce dernier, à partir d'informations sommaires (majoritairement catégorielles), le but est de prévoir si l'utilisateur va cliquer sur l'affichage ou non (et donc de choisir d'enchérir ou non). Les clics étant rares, de l'ordre de 1 pour 1000 en moyenne, il y a un fort déséquilibre entre les deux classes (clic ou non-clic). Dans ce cadre, les modèles de prédiction habituels sont biaisés, puisqu'ils ne prennent pas en compte ce déséquilibre. Les mesures d'évaluation, telles que AUC_{ROC}, sont également défectueuses lorsqu'elles sont appliquées à ces événements rares. Dans cet article, nous étudions ces biais dans le cadre des enchères temps réel et proposons une nouvelle mesure d'évaluation. Nous concentrons notre analyse sur la prédiction du clic ($pCTR$), en considérant des algorithmes de prédiction linéaires (régression logistique) et non linéaires (Deep Factorization Machine). Nous évaluons la performance des modèles à l'aide d'une fonction d'évaluation probabiliste spécifique incluant les coûts liés à l'enchère, et nous la comparons à plusieurs mesures d'évaluation classiques. Cette mesure met en évidence les limites des métriques classiques dans les problèmes de prédiction fortement déséquilibrés. Elle permet ainsi une meilleure évaluation des modèles spécifiques à la prédiction du clic et fournit également des indications quant à la rentabilité de la campagne d'affichage.

1 Introduction

Ces dernières années, le marché de la publicité en ligne a connu une croissance exponentielle. Dans le même temps, les méthodes d'attribution de ces publicités ont beaucoup évolué. Le real-time bidding (RTB) est le principal moyen pour les annonceurs d'acheter et d'afficher leurs publicités en ligne. Les algorithmes RTB sont conçus pour optimiser les transactions afin que les éditeurs reçoivent le meilleur prix par affichage remporté et que les annonceurs puissent diffuser leurs annonces auprès du public le plus pertinent au moindre coût. Chaque emplacement publicitaire disponible est mis aux enchères sur une plateforme AdExchange où chaque

annonceur peut placer des offres par le biais d'un algorithme en temps réel et finalement afficher ses annonces.

Les enchères pour les affichages publicitaires sont calculées sur la base de critères tels que la taille et la position d'une publicité, des détails sur le site web. Pour maximiser le revenu (*i.e. le nombre de clics*) d'une campagne, un enchérisseur doit estimer pour chaque demande d'enchère, la probabilité de clic (pCTR), également appelée réponse de l'utilisateur. Cette estimation de l'utilité relève de la prédiction d'événements rares car l'affichage d'une annonce ne conduit que très rarement à un clic : moins de $0,1\%$ des annonces sont effectivement cliquées. Les algorithmes d'apprentissage automatique nécessitent généralement une proportion raisonnable d'événements, c'est-à-dire de cas d'intérêt que l'on veut apprendre à prédire. Par conséquent, lorsque les événements sont rares, les modèles de classification classiques sont biaisés et de nombreux travaux ont été réalisés pour étudier et corriger ces biais (Tomz et al., 2003; Van Den Eeckhaut et al., 2006; Weiss et Hirsh, 1998; Ranjan, 2020). Le principal problème dans le contexte des événements rares est le déséquilibre des classes : le très faible rapport entre les événements et les non-événements nécessite de rassembler de très grands ensembles de données pour avoir un nombre suffisant d'événements. King et Zeng (2001) ont proposé une version débiaisée de la régression logistique en corrigeant ses prédicteurs $\hat{\beta}$ avec le biais b de l'estimation du maximum de vraisemblance (EMV) (McCullagh et Nelder, 1989; Yang et al., 2015). Cette correction est axée sur les ensembles de données de petite taille. Puisque les EMV pour la régression logistique sont asymptotiquement sans biais, moins nous avons d'événements, plus le biais est élevé (Leitgöb, 2013). Firth (1993) et Tomz et al. (2003) ont proposé l'estimation par maximum de vraisemblance pénalisée (EMVP). Leitgöb (2013) montre que cette méthode corrige le biais même pour les très petits ensembles de données alors que la méthode de King a tendance à sur-corriger les estimations lorsque l'ensemble de données devient plus petit.

En ce qui concerne l'évaluation de la classification, l'une des métriques les plus utilisées est la AUC_{ROC} (Area under the ROC curve). Elle mesure l'aire sous la courbe ROC, qui est la courbe représentant le taux de vrais positifs (TPR), également appelé rappel (ou sensibilité en classification binaire) par rapport au taux de faux positifs (FPR), définis comme suit :

$$TPR = \frac{TP}{TP + FN} \text{ , et } FPR = \frac{FP}{FP + TN}$$

pour tous les seuils de séparation. TP, FN, FP et TN sont définis comme :

— Faux positif (FP) : Prédire un clic quand il n'y en a pas.
— Faux négatif (FN) : Ne pas prédire un clic alors qu'il y en a un.
— Vrai positif (TP) : Prédire un clic alors qu'il y en a un.
— Vrai négatif (TN) : Ne pas prédire de clic et qu'il n'y en ait pas.

L'aire sous la courbe ROC mesure la qualité d'un modèle (Narkhede, 2018; Hilden, 1991). Cependant, l'AUC_{ROC} accorde une importance égale aux faux positifs (FP) et aux faux négatifs (FN) : en RTB, prédire un clic qui ne se produit pas réellement a un coût différent de celui de ne pas prédire un clic lorsqu'il y en a un. Un effet secondaire direct de cela concerne la comparaison des modèles : deux modèles de prédiction du CTR peuvent avoir le même AUC_{ROC} mais conduire à des performances d'enchères différentes. Drummond et Holte (2004) montrent comment l'AUC_{ROC} échoue à prendre en compte les coûts. Dans des contextes d'événements rares tels que les risques géologiques, les crises et le RTB, les événements et les non-

événements ont des répercussions différentes. Il est donc souhaitable d'utiliser une fonction d'évaluation sensible aux coûts.

De plus, la courbe ROC étant calculée pour chaque seuil possible, ses régions extrêmes sont prises en compte dans le AUC_{ROC} alors qu'elles ne sont pas pertinentes dans le contexte de la classification. Le AUC_{ROC} ne peut également prendre en compte que les classements et non les probabilités absolues (Lobo et al., 2008) alors que la plupart des algorithmes d'optimisation des offres s'appuient sur le pCTR pour proposer des offres. Une autre métrique courante pour évaluer les modèles d'événements rares est l'aire sous la courbe de précision/rappel (AUC_{PR}). Elle mesure l'aire sous la courbe représentant la précision ($P = \frac{TP}{(TP+FP)}$) en fonction du rappel (TPR). Contrairement à l'AUC_{ROC}, elle ne tient pas compte du taux de faux positifs (FPR) qui tend à être très faible dans les applications d'événements rares en raison de la grande proportion d'exemples négatifs. Elle s'est avérée efficace dans le contexte des événements rares (Saito et Rehmsmeier, 2015; Davis et Goadrich, 2006; Sofaer et al., 2019). Sofaer et al. (2019) comparent AUC_{ROC} et AUC_{PR} montrant que pour les événements rares, AUC_{PR} est moins enclin à surestimer la performance de classification des modèles en négligeant le taux de vrais négatifs. Néanmoins, AUC_{PR} calcule toujours un score général en considérant tous les seuils, y compris les régions extrêmes, et est donc biaisé.

La prédiction du taux de clics a déjà été bien étudiée par la communauté de recherche sur les systèmes de recommandation, dans des tâches telles que les problèmes de recommandation top-k (Covington et al., 2016; Khabbaz et Lakshmanan, 2011; Wang et al., 2019; Xiao et al., 2020). Cependant, les paramètres de l'environnement d'enchères en temps réel impliquent que les demandes d'annonces soient traitées de manière séquentielle et non globale. Cette différence crée des problèmes de démarrage à froid en raison de la grande variété de visiteurs de sites Web, de positions d'annonces, de sites Web, etc. Il est également important de souligner les problèmes de confidentialité qui sont actuellement abordés par les acteurs de l'industrie tels que Google avec la politique cookieless et PrivacySandbox [1] mais aussi par les législateurs européens avec les lois européennes sur la confidentialité. Cela empêchera les algorithmes d'enchères en temps réel d'utiliser les modèles développés à des fins de recommandation. Une autre différence concerne les coûts, car dans la plupart des cas d'utilisation de la recommandation, on ne paie généralement pas pour faire une recommandation, donc le coût d'un clic mal prédit dans les enchères en temps réel est beaucoup plus élevé que dans les systèmes de recommandation (Shen et al., 2022).

Dans la suite de cet article nous présentons d'abord les modèles de prédiction d'événements rares. Nous utilisons la version pondérée de la régression logistique et l'appliquons à la prédiction du taux de clics, ce qui n'a pas encore été fait à notre connaissance, et démontrons la supériorité de cette approche. Nous présentons la tâche de prédiction de la réponse utilisateur et la manière dont elle s'inscrit dans le cadre des événements rares tout en présentant des défis spécifiques. Nous nous concentrons ensuite sur l'évaluation de la performance de la prédiction des événements rares, en discutant des angles morts des métriques habituelles telles que AUC_{ROC} et AUC Precision Recall (AUC_{PR}). Nous introduisons une fonction de valeur sensible aux coûts pour aider à surmonter ces biais et fournir des informations utiles à l'utilisateur final, en particulier en termes de bénéfices futurs d'une campagne.

1. https://privacysandbox.com/

2 Prédiction et évaluation du taux de clics

Cette section se concentre sur la prédiction du taux de clics (pCTR), en analysant les performances de prédiction pour les modèles d'événements rares. Les performances sont évaluées à l'aide de plusieurs métriques mettant en évidence les faiblesses des métriques habituelles dans des contextes de données fortement déséquilibrées. Les avantages et les inconvénients de chaque approche sont mis en évidence.

Dans la suite de cet article nous désignons les données d'entrées à n échantillons par $X = \{x_1, \ldots, x_n\}$.

2.1 Modélisation d'événements rares

Dans notre étude sur la prédiction d'événements rares, nous considérons trois modèles que nous présentons dans cette section : La régression logistique (LR) classique comme modèle de base, une version corrigée de LR ainsi que Deep Factorization Machine (DFM), un modèle d'apprentissage profond développé pour la prédiction du clic (Guo et al., 2017).

2.1.1 Régression logistique (LR)

Pour modéliser la probabilité d'un événement donné, la régression logistique (Hosmer Jr et al., 2013) est une méthode très connue, notamment dans la communauté de l'apprentissage automatique. Elle estime les paramètres θ d'un modèle logistique exprimé par :

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

où $x = \{x^1, x^2, \ldots, x^T\}$. x étant un échantillon de données de T variables, $h_\theta(x)$ exprime, étant donné θ, les paramètres du modèle, la probabilité que x appartienne à la classe positive. Pour estimer les paramètres θ, nous utilisons la descente de gradient stochastique avec la logistic loss (Logloss) définie comme l'inverse de la vraisemblance :

$$L(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x)) \tag{1}$$

Bien que la régression logistique puisse être assez puissante et rapide pour modéliser des combinaisons linéaires, la prise en compte des interactions d'ordre supérieur $x^{(t)}x^{(t')}$ entre les caractéristiques t, t' de l'échantillon d'apprentissage $\{x\}$ n'est pas triviale. Une telle combinaison de caractéristiques peut être traitée manuellement avec l'ingénierie des caractéristiques au prix de lourds efforts puisque le nombre d'interactions par paire augmente quadratiquement avec le nombre de caractéristiques.

2.1.2 Régression logistique pondérée (wLR)

Le modèle de régression logistique pondérée utilisé dans cet article ajoute un poids à l'erreur de chaque échantillon par le poids de sa classe. Cela équivaut à une perte d'entropie croisée pondérée :

$$L(h_\theta(x), y) = -y c_0 \log(h_\theta(x)) - (1 - y) c_1 \log(1 - h_\theta(x))$$

où :

$$c_0 = \frac{n}{n_{classes} * n_{events}} \text{ et } c_1 = \frac{n}{n_{classes} * n_{non-events}}$$

avec n le nombre total d'échantillons, $n_{classes} = 2$ dans le cas de la classification binaire, n_{events} et $n_{non-events}$ respectivement le nombre d'événements, dans notre cas, les clics et de non-événements (absence de clic) dans l'ensemble de données. Le but de cette correction est d'égaliser l'importance de chaque classe en multipliant la perte de chaque échantillon par l'inverse de la prévalence de sa classe. Elle diminue donc fortement le poids des échantillons de la classe majoritaire et augmente celui des échantillons de la classe minoritaire. wLR fournit des estimations non biaisées de θ mais le problème d'estimation d'interactions entre variables est toujours présent.

2.1.3 Deep Factorization Machine (DFM)

Le modèle DFM a été initialement introduit par Guo et al. (2017) pour la prédiction du taux de clics. La DFM est basée sur deux composants : une machine à factoriser (FM) (Rendle, 2010) qui apprend les interactions de caractéristiques linéaires et par paire et une partie de réseau neuronal profond (DNN) qui cherche à modéliser les interactions d'ordre supérieur. Le modèle FM est défini par :

$$y_{FM}(x) = w_0 + \sum_{i=1}^{t} w_i x_i + \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} \langle m_i, m_j \rangle x_i x_j$$

$$\text{avec } \langle m_i, m_j \rangle = \sum_{f=1}^{t} m_{i,f} \cdot m_{j,f}$$

Où la première partie est un modèle linéaire classique avec w_0 le biais et w_i les poids d'ordre premier. Le dernier terme peut être réécrit :

$$\sum_{i=1}^{t-1} \sum_{j=i+1}^{t} \langle m_i, m_j \rangle x_i x_j = \frac{1}{2} \sum_{f=1}^{k} \left[\left(\sum_{i=1}^{t} m_{i,f} x_i \right)^2 - \sum_{i=1}^{t} m_{i,f}^2 x_i^2 \right]$$

où $m_i, m_j \in M \in \mathbb{R}^{T \times k}$, T est le nombre de variables dans l'échantillon d'apprentissage et k est un hyperparamètre contrôlant la dimension de la représentation vectorielle qui doit être choisi par validation croisée. Cette représentation réduit considérablement la complexité des données catégorielles en codage disjonctif, car seules les entrées non nulles sont calculées (Rendle, 2010). Cette représentation vectorielle peut être vue comme une modélisation des interactions entre les caractéristiques sans avoir à modéliser explicitement les interactions par paire. Les paramètres sont initialisés de manière aléatoire et optimisés en minimisant la perte d'entropie croisée avec la descente de gradient stochastique. La sortie du DFM est donnée par $y_{DFM} = sigmoid(y_{FM} + y_{DNN})$ où y_{DNN} est la sigmoïde de la sortie du DNN. Comme indiqué dans l'article original (Guo et al., 2017), l'un des avantages du modèle DFM est que les deux composants sont alimentés par la même entrée, qui est la sortie d'une couche d'embedding : chaque caractéristique en codage disjonctif est transformée en une représentation

de longueur fixe. L'inconvénient potentiel de cette approche est qu'elle ajoute une complexité d'apprentissage au modèle et rend donc la convergence plus difficile. Les modèles d'apprentissage profond sont également connus pour être très sensibles aux hyperparamètres et nécessitent généralement beaucoup d'ajustements.

2.2 Évaluation du modèle d'événements rares

Formellement, la prédiction du taux de clics (CTR) consiste à prédire la probabilité d'un clic $p(x)$ étant donné les données de demande d'affichage x. Ces données d'entrée contiennent des informations sur le visiteur de la page Web (navigateur, appareil, région de connexion, ...) et sur l'emplacement publicitaire lui-même (taille de l'emplacement, type de publicité : vidéo, pop-up, ...). L'utilisation de ces données pour prédire si un clic va se produire est cruciale pour l'enchérisseur car elle fournit l'utilité espérée $u(x) = p(x)v - c(x)$ avec v la valeur que l'annonceur attribue à un clic et $c(x)$ le prix à payer pour l'enchère gagnée. Cette valeur dépend fortement du coût par clic (CPC), également appelé paiement par clic souvent utilisée comme système de tarification. Pour qu'un modèle soit rentable, v doit être supérieur au CPC ($CPC_{moyen} \approx \$2 \approx 12.73CN$)[2]. Dans le présent document, les prix sont exprimés en yuans x 1000, conformément à la littérature et à la métrique du coût par mille (CPM), définissant le coût moyen de mille affichages remportés. C'est l'indicateur de performance couramment utilisé dans le secteur.

Dans un contexte de prédiction du CTR pour les enchères en temps réel, un faux positif signifie la prédiction d'un clic sur une demande d'enchère qui n'entraînera pas un clic. Un tel échantillon mal classé conduira l'algorithme d'optimisation des enchères à surenchérir et à dépenser inutilement du budget. Au contraire, un faux négatif signifie que l'on ne prédit aucun clic alors qu'il y en aura un, ce qui incitera le composant d'optimisation des enchères à ne pas enchérir et donc à manquer ce clic. En comparant ces deux cas, il est évident que les faux positifs doivent être fortement pénalisés dans l'évaluation d'un modèle de prédiction du CTR, ce que AUC_{ROC} ne permet pas. Pour refléter ces différents coûts, nous proposons une fonction de valeur qui prend en compte la valeur associée à un clic v par l'annonceur. Elle donne également des valeurs différentes pour les faux négatifs et les faux positifs. Les algorithmes d'optimisation des enchères se basant généralement sur les probabilités de clics prédites et non sur des prédictions binarisées (Zhang et al., 2014a; Lee et al., 2013), nous définissons la fonction de valeur de manière probabiliste comme suit :

$$\begin{aligned} V(x,v) &= \underbrace{-\left(\sum_i^n c_i p(x_i)(1-y_i)\right)}_{\text{coût FP}} + \underbrace{\left(\sum_i^n (v-c_i)p(x_i) \times y_i\right)}_{\text{coût TP}} \\ &= \sum_i^n \left(-c_i p(x_i) + v p(x_i) y_i\right) \end{aligned}$$

Où $y_i \in \{0,1\}$ est une variable binaire correspondant à l'existence ou non d'un clic. La fonction présentée n'associe pas de coût aux faux négatifs car en application réelle, le fonctionnement des enchères en temps réel ne donne pas à accès à l'information du clic y_i si l'enchère n'est pas remportée. De la même manière, les vrais positifs sont considérés comme coûts nuls.

2. https://www.wordstream.com/cost-per-click

3 Application

3.1 Paramètres expérimentaux

Pour réaliser nos expériences, nous utilisons le jeu de données Ipinyou (Zhang et al., 2014b) qui est largement utilisé dans la littérature sur l'optimisation des enchères et la prédiction du CTR (Zhang et al., 2014b; Huang et al., 2020; Qu et al., 2016; Ren et al., 2018). Il a été initialement publié par Ipinyou, une société de publicité leader sur le marché chinois, pour son concours international d'enchères en temps réel en 2013 et reste un ensemble de données de référence dans la communauté de recherche sur la prédiction du CTR (Huang et al., 2020; Liu et al., 2020b) et l'optimisation des enchères (Huang et al., 2020; Liu et al., 2020a). Nous utilisons une version modifiée de l'ensemble de données original [3]. Ce jeu de données modifié est organisé en neuf campagnes, chacune correspondant à un annonceur différent. Nous construisons un ensemble de données avec environ 500 caractéristiques binaires (en fonction du nombre de modalités dans chaque campagne) obtenues par codage disjonctif des 12 variables sélectionnées : weekday, hour, region, city, adexchange, slotwidth, slotheight, slotvisibility, slotformat, creative, os, browser. Notons que toutes ces caractéristiques sont catégorielles, et que l'ensemble de données en codage disjonctif est très épars. Cette sélection de variables a été faite en écartant celles ayant un nombre élevé de modalités, parfois autant que d'exemples dans les données, ce qui les rend inutiles pour un modèle de prédiction et aurait créé un ensemble de données encore plus clairsemé. Nous avons également divisé les usertags qui étaient à l'origine une liste de tags caractérisant le visiteur selon la segmentation interne d'Ipinyou et les avons également encodés. Les résultats présentés sont obtenus par validation croisée 10 fois. Nous avons divisé le jeu de données en 10 folds avec un ratio entraînement/test fixé à 90%. Comme les données des campagnes concernent différents annonceurs, les probabilités de clics et les annonces sont très différentes suivant que les annonces concernent des voitures ou des vêtements par exemple. Nous entraînons donc un modèle différent par campagne. Suivant King et Zeng (2001), nous sous-échantillonnons les non-événements pour obtenir un ratio de 1 événement 10 pour non-événements. Pour chaque fold, nous sélectionnons tous les événements (c.-à-d. les clics) et échantillonnons aléatoirement 10 fois plus de non-événements. Cette procédure ne dégrade que très peu les performances du modèle en accord avec Wang (2020).

Pour les régressions logistiques, nous utilisons l'implémentation scikit-learn [4]. Pour la version pondérée de LR, nous utilisons le paramètre $class\_weight = "balanced"$ qui est déjà implémenté pour le déséquilibre de classes.

Notre DFM est écrit en Pytorch [5]. Nous avons empiriquement fixé la taille du DNN à (400, 400, 400), ce qui correspond également à l'implémentation par Guo et al. (2017). Nous fixons la dimension de la représentation latente à $k = 6$. Le modèle est entraîné pendant 500 itérations sur chaque fold et 20% des individus des folds sont utilisés pour la validation.

Pour les mesures d'évaluation, nous utilisons les implémentations de scikit-learn pour la perte d'entropie croisée, également appelée logistic loss, et AUC_{ROC}. Pour AUC_{PR}, nous

3. https://github.com/wnzhang/make-ipinyou-data
4. https://scikit-learn.org/
5. https://pytorch.org/

utilisons le *average_precision_score*[6] qui approxime l'aire sous la courbe de rappel de précision et est défini par : $AUC_{PR} = \sum_n (R_n - R_{n-1}) P_n$ où P_n et R_n sont la précision et le rappel au n-ième seuil.

3.2 Résultats

Campagne	$N_{sous-echatillone}$	$N_{origine}$	dimension	Clics	Algorithme	Logloss	AUC_{ROC}	AUC_{PR}
					wLR	0.22	**0.94**	**0.81**
1458	26 994	3 083 056	525	2 454	LR	**0.11**	**0.94**	**0.81**
					DFM	0.39	0.60	0.12
					wLR	0.62	0.62	**0.19**
2259	3 080	835 556	191	280	LR	**0.31**	**0.63**	0.18
					DFM	0.48	0.50	0.11
					wLR	0.56	**0.62**	0.15
2261	2 277	687 617	577	207	LR	**0.31**	**0.62**	**0.17**
					DFM	0.55	0.54	0.10
					wLR	0.63	0.58	0.15
2821	9 273	1 322 561	550	843	LR	**0.31**	**0.60**	**0.16**
					DFM	0.44	0.51	0.10
					wLR	0.64	0.60	0.13
2997	15 246	312 437	511	1 386	LR	**0.29**	**0.61**	**0.14**
					DFM	**0.29**	0.50	0.08
					wLR	0.27	0.92	0.76
3358	14 938	1 742 104	541	1 358	LR	**0.14**	**0.93**	**0.77**
					DFM	0.40	0.62	0.18
					wLR	0.58	0.68	0.29
3386	22 836	2 847 802	533	2 076	LR	**0.27**	**0.69**	**0.30**
					DFM	0.39	0.58	0.12
					wLR	0.32	**0.91**	0.70
3427	21 186	2 593 765	550	1 926	LR	**0.15**	0.90	**0.75**
					DFM	0.36	0.56	0.11
					wLR	0.44	**0.85**	0.38
3476	11 297	1 970 360	533	1 027	LR	**0.20**	**0.85**	**0.59**
					DFM	0.60	0.52	0.10

TAB. 1 – *Caractéristiques et performances pour chaque campagne. Les performances sont données dans différentes métriques pour chaque modèle. La perte logistique est définie dans l'équation 1*

D'après le tableau 1, la régression logistique classique donne les meilleurs résultats dans chaque campagne.

Nous présentons dans la figure 1 la fonction de valeur pénalisée V en fonction du v. En fixant la valeur que l'annonceur associe à un clic, on obtient une borne inférieure de cette valeur pour qu'une campagne soit rentable. Ces résultats montrent un résultat totalement différent concernant le classement des modèles : la version équilibrée de la régression logistique surpasse les deux autres modèles. Cela confirme les biais des mesures d'évaluation classiques qui ne reflètent pas les coûts associés aux enchères. La fonction de valeur proposée est linéaire et nous voyons que la régression logistique pondérée a la pente la plus élevée, ce qui signifie que c'est le modèle le plus rentable quelle que soit la valeur v. À des fins de comparaison, nous ajoutons l'Oracle qui est le meilleur prédicteur de CTR possible et qui n'a donc que des vrais positifs. La fonction de valeur de l'Oracle est toujours plus élevée que les algorithmes considérés. Cela confirme que la fonction d'évaluation proposée reflète bien les performances de prédiction du CTR et que nous pouvons nous y fier pour choisir le meilleur modèle, c'est-à-dire wLR dans nos expériences. En plus de permettre une meilleure évaluation des modèles,

6. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

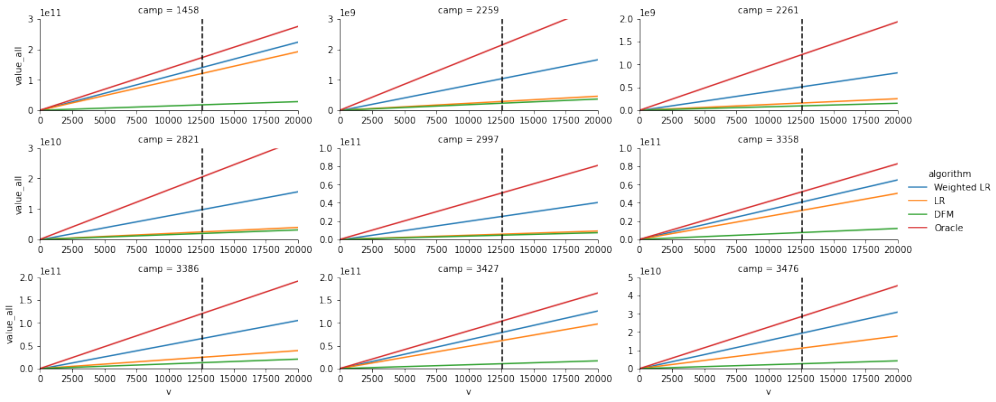

FIG. 1 – *Fonction de valeur pour chaque campagne avec en abscisse la valeur v associée à un clic et V, la valeur de la fonction d'évaluation proposée en ordonné. La ligne verticale en pointillé représente le CPC moyen du marché*

la fonction proposée fournit des indications utiles d'un point de vue appliqué : elle donne à l'annonceur le CPC le plus élevé qu'il pourrait accepter de l'éditeur pour que sa campagne soit rentable. Cette caractéristique confirme les avantages de l'utilisation de la fonction proposée pour évaluer et classer les modèles de prédiction du taux de clics. En examinant les différences de performance entre les campagnes, nous constatons une corrélation entre la performance de prédiction et le nombre d'événements dans la campagne, surtout en terme de V : les campagnes 2259 et 2261 ont beaucoup moins d'événements et aussi les valeurs les plus faibles en termes de fonction de valeur. Cette remarque vaut pour les autres métriques du tableau 1 mais seulement à un degré moindre.

Concernant le modèle d'apprentissage profond DFM, les résultats montrent des performances médiocres, quelle que soit la métrique d'évaluation. Bien qu'il ait été développé pour cette tâche spécifique, ce modèle semble être très spécifique aux données et n'est pas compétitif. Dans notre cas, il ne vaut pas le temps et la puissance de traitement supplémentaires qu'il requiert par rapport à la régression logistique pondérée.

4 Conclusions

Les enchères en temps réel sont un domaine actif et des améliorations sont nécessaires pour optimiser les stratégies d'enchères. La prédiction du taux de clics est un élément essentiel des algorithmes d'enchères.

Nous avons étudié les fonctions d'évaluation pour les événements rares et proposé une fonction d'évaluation spécifique à la prédiction du taux de clics, tenant compte des coûts. Nous montrons que la logloss, AUC_{ROC} mais aussi AUC_{PR} ne tiennent pas compte de la nature déséquilibrée des données, faisant de ces métriques, de mauvais critères de sélection de modèle de prédiction d'événements rares. Notre proposition corrige ce biais, et, en considérant les coûts et les valeurs des clics, permet également d'évaluer le rendement potentiel d'un modèle de prédiction du taux de clics.

Nous avons proposé la régression logistique pondérée pour la prédiction d'événements rares appliquée à la prédiction du taux de clics et avons comparé ses performances avec une base de régression logistique classique et un modèle d'apprentissage profond. Cela conduit à une amélioration significative des prédictions du taux de clics. L'approche par régression logistique est d'autant plus intéressante dans le cas de la prédiction du clic qu'elle peut être appliqué à des données éparses de haute dimension (Genkin et al., 2007; Lin et al., 2007) ce qui est le cas des données de RTB. À cela s'ajoute l'avantage de rapidité de calcul de cette approche correspondant au besoin de temps réel des enchères en temps réel. D'un point de vue économique, l'approche proposée de régression logistique pondérée est la plus rentable, sur la base de la fonction de valeur tenant compte des coûts.

Dans cet article, cependant, nous avons couvert le problème de la prédiction de la réponse de l'utilisateur sans tenir compte de la partie d'optimisation des enchères en temps réel. Nous travaillons à l'unification de ces deux parties en intégrant la fonction de valeur présentée comme fonction objective dans un algorithme d'optimisation des offres et nous le présenterons dans un travail futur.

Références

Covington, P., J. Adams, et E. Sargin (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, New York, NY, USA, pp. 191–198. Association for Computing Machinery.

Davis, J. et M. Goadrich (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA, pp. 233–240. Association for Computing Machinery.

Drummond, C. et R. C. Holte (2004). What roc curves can't do (and cost curves can). In *ROCAI*, pp. 19–26. Citeseer.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Genkin, A., D. D. Lewis, et D. Madigan (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics 49*(3), 291–304.

Guo, H., R. Tang, Y. Ye, Z. Li, et X. He (2017). Deepfm : A factorization-machine based neural network for ctr prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 1725–1731. AAAI Press.

Hilden, J. (1991). The area under the roc curve and its competitors. *Medical Decision Making 11*(2), 95–101.

Hosmer Jr, D. W., S. Lemeshow, et R. X. Sturdivant (2013). *Applied logistic regression*, Volume 398. John Wiley & Sons.

Huang, G., Q. Chen, et C. Deng (2020). A new click-through rates prediction model based on deepcross network.

Khabbaz, M. et L. V. S. Lakshmanan (2011). Toprecs : Top-k algorithms for item-based collaborative filtering. In *EDBT/ICDT '11*.

King, G. et L. Zeng (2001). Logistic regression in rare events data. *Political Analysis 9*, 137–163.

Lee, K.-C., A. Jalali, et A. Dasdan (2013). Real time bid optimization with smooth budget delivery in online advertising. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, New York, NY, USA. Association for Computing Machinery.

Leitgöb, H. (2013). The problem of modeling rare events in ml-based logistic regression.

Lin, C.-J., R. C. Weng, et S. S. Keerthi (2007). Trust region newton methods for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, New York, NY, USA, pp. 561–568. Association for Computing Machinery.

Liu, B., N. Xue, H. Guo, R. Tang, S. Zafeiriou, X. He, et Z. Li (2020b). Autogroup : Automatic feature grouping for modelling explicit high-order feature interactions in ctr prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, New York, NY, USA, pp. 199–208. Association for Computing Machinery.

Liu, M., L. Jiaxing, Z. Hu, J. Liu, et X. Nie (2020a). A dynamic bidding strategy based on model-free reinforcement learning in display advertising. *IEEE Access 8*, 213587–213601.

Lobo, J., A. Jiménez-Valverde, et R. Real (2008). Auc : a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography 17*, 145–151.

McCullagh, P. et J. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Narkhede, S. (2018). Understanding auc-roc curve. *Towards Data Science 26*(1), 220–227.

Qu, Y., H. Cai, K. Ren, W. Zhang, Y. Yu, Y. Wen, et J. Wang (2016). Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1149–1154.

Ranjan, C. (2020). *Understanding Deep Learning : Application in Rare Event Prediction*. Connaissance Publishing.

Ren, K., W. Zhang, K. Chang, Y. Rong, Y. Yu, et J. Wang (2018). Bidding machine : Learning to bid for directly optimizing profits in display advertising. *IEEE Transactions on Knowledge and Data Engineering 30*(4), 645–659.

Rendle, S. (2010). Factorization machines. In *2010 IEEE International conference on data mining*, pp. 995–1000. IEEE.

Saito, T. et M. Rehmsmeier (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE 10*(3), 1–21.

Shen, Q., H. Wen, W. Tao, J. Zhang, F. Lv, Z. Chen, et Z. Li (2022). Deep interest highlight network for click-through rate prediction in trigger-induced recommendation. In *Proceedings of the ACM Web Conference 2022*, pp. 422–430.

Sofaer, H. R., J. A. Hoeting, et C. S. Jarnevich (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution 10*(4), 565–577.

Tomz, M., G. King, et L. Zeng (2003). Relogit : Rare events logistic regression. *Journal of statistical software 8*(1), 1–27.

Van Den Eeckhaut, M., T. Vanwalleghem, J. Poesen, G. Govers, G. Verstraeten, et L. Vande-

kerckhove (2006). Prediction of landslide susceptibility using rare events logistic regression : A case-study in the flemish ardennes (belgium). *Geomorphology 76*(3), 392–410.

Wang, H. (2020). Logistic regression for massive data with rare events. In H. D. III et A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 9829–9836. PMLR.

Wang, P., Y. Jiang, C. Xu, et X. Xie (2019). Overview of content-based click-through rate prediction challenge for video recommendation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, pp. 2593–2596. Association for Computing Machinery.

Weiss, G. M. et H. Hirsh (1998). Learning to predict rare events in event sequences. In *KDD*, Volume 98, pp. 359–363.

Xiao, Z., L. Yang, W. Jiang, Y. Wei, Y. Hu, et H. Wang (2020). Deep multi-interest network for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, New York, NY, USA, pp. 2265–2268. Association for Computing Machinery.

Yang, H., K. Ozbay, K. Xie, et B. Bartin (2015). Modeling crash risk of highway work zones with relatively short durations. In *In Transportation Research Board 94th Annual Meeting*.

Zhang, W., S. Yuan, et J. Wang (2014a). Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, New York, NY, USA, pp. 1077–1086. Association for Computing Machinery.

Zhang, W., S. Yuan, J. Wang, et X. Shen (2014b). Real-time bidding benchmarking with ipinyou dataset.

Summary

When online advertisements are displayed, a real-time bidding system is set up to choose the advertiser. On the advertiser's side, based on summary information (mostly categorical), the goal is to predict whether the user will click on the display or not (and thus choose whether to bid). Clicks being rare (1 per 1000 on average), there is a strong imbalance between the two classes (click and no-click). In this context, the usual prediction models are biased, since they do not take into account this imbalance. The usual evaluation measures, such as AUC_{ROC}, are also flawed when applied to these rare events. In this paper, we study these biases in the context of real-time auctions and propose a new evaluation measure. We focus our analysis on click prediction ($pCTR$), considering both linear (logistic regression) and non-linear (Deep Factorization Machine) prediction algorithms. We evaluate the performance of the models using a specific probabilistic evaluation function including the costs associated with the auction, and compare it to several classical evaluation measures. This measure highlights the limitations of classical metrics in highly unbalanced prediction problems. It thus allows for a better evaluation of specific click prediction models and also provides insights into the profitability of the display advertising campaign.