Groupe de statistique appliquée - Note de synthèse

Construction d'un modèle de prédiction de clic sur des publicités en ligne avec contraintes sur la quantité de données et le temps d'évaluation

Vincent NGUYEN, Baptiste PASQUIER, Alix PLAMONT & Théo PORTALIER
Octobre 2020 - Mai 2021

Résumé

Dans le cadre de notre deuxième année à l'ENSAE Paris, nous avons effectué notre projet de statistique appliquée auprès de l'entreprise Criteo, référence mondiale du reciblage publicitaire, sous la supervision de Louis PRUVOT-CAPRIOLI et Hallvard GESLIN. Le cœur de métier de l'entreprise consiste à acheter des emplacements publicitaires via des systèmes d'enchères pour le compte d'une firme cliente (Fnac, La Redoute...), laquelle rémunère ensuite Criteo en fonction du nombre de clics des potentiels consommateurs sur l'affichage publicitaire. Ainsi, dans la mesure où ses gains dépendent du nombre de clics, Criteo supporte l'intégralité du risque ; c'est pourquoi il lui est essentiel de mettre en place en amont des algorithmes de prédiction de clic afin de n'acheter que des affichages ayant une haute probabilité de donner lieu à un clic de la part de l'utilisateur. Par ailleurs, les algorithmes prédictifs utilisés doivent présenter des temps de prédiction extrêmement réduits dans la mesure où les enchères se font en quelques millisecondes seulement.

Notre mission a donc consisté à tester différents modèles de *Machine Learning* permettant de prédire efficacement ces clics à partir d'une base de données fournie par l'entreprise et issue de l'enregistrement de certaines données de navigation internet d'utilisateurs, ou encore de données recueillies par des applications pour mobile. Il est rapidement apparu que certaines caractéristiques liées à l'utilisateur ou à l'affichage publicitaire tendent à augmenter la probabilité des clics, ce qui nous a permis de dégager quelques premières tendances explicatives. Nous avons ensuite comparé les performances des modèles implémentés en définissant préalablement une métrique de référence, autrement dit un critère de mesure de ces performances cohérent avec les impératifs stratégiques de Criteo. Nous avons enfin examiné l'évolution de ces performances sous contrainte temporelle, en regardant par exemple à quel point il est possible de réduire les données d'apprentissage (c'est-à-dire les exemples labellisés sur lesquels se basent les prédictions) tout en maintenant un bon niveau de performance. Finalement, nous avons pu dégager deux modèles offrant un bon compromis entre la performance et le temps d'évaluation d'une part et la robustesse aux contraintes d'autre part.

Dans une optique de vulgarisation scientifique, nous présenterons brièvement dans ce qui suit les résultats majeurs issus de nos travaux ainsi que les concepts fondamentaux auxquels nous avons eu recours.

1 Quelles sont les données utilisées ?

Chaque ligne de notre base de données (que nous nommerons par la suite observation) correspond à l'affichage effectif d'une publicité sur l'écran d'un appareil (ordinateur, téléphone, tablette...) pour un utilisateur donné. Ces observations comportent 47 variables, qui donnent des renseignements à la fois sur la publicité et l'enchère sous-jacente, mais aussi sur l'utilisateur et les spécificités de la campagne publicitaire. Par exemple, des informations portant sur le degré d'engagement de l'utilisateur, la taille et le coût de l'affichage publicitaire ou encore la performance de la campagne sont disponibles, de même que des renseignements quant à l'historique d'activité de l'utilisateur (nombre d'achats qu'il a déjà effectués auprès du site annonceur, nombre de jours écoulés depuis son dernier clic sur une publicité...). Les données à notre disposition couvrent temporellement une période d'environ une semaine pour un total de 2 135 241 observations.

Cette base nous a été transmise sous la forme de deux fichiers distincts : l'un dédié à l'entraînement de nos modèles statistiques (87% des données), l'autre dédié aux tests (13% des données) et grâce auquel nous avons comparé les performances de nos modèles. Nous dénombrons en moyenne 1.6 affichage par utilisateur dans la base d'entraînement, tandis que les publicités proviennent de 10 annonceurs distincts et s'inscrivent dans 78 campagnes publicitaires différentes. Le jeu de données est sans surprise extrêmement déséquilibré, avec seulement 6% d'affichages publicitaires donnant lieu à un clic. Afin de faciliter notre travail ultérieur, nous avons également procédé à quelques modifications sur les bases initiales, en remplaçant notamment les valeurs manquantes, très nombreuses pour les variables associées à des nombres d'affichages ou de ventes, par des 0, ou en recodant les modalités de certaines variables pour plus de clarté.

2 Quelles sont les caractéristiques susceptibles d'expliquer un clic ?

Dans le double but de présélectionner des variables et de vérifier ultérieurement la cohérence de nos modèles de prédiction, nous avons entrepris d'analyser les liens entre les différentes variables à notre disposition, et notamment leurs liens avec la variable à expliquer : l'indicatrice du clic.

Dans un premier temps, l'étude des corrélations

entre les différentes caractéristiques quantitatives de l'individu ou de la publicité à notre disposition (engagement marketing, nombre de jours écoulés depuis le dernier clic, coût de l'emplacement publicitaire...) a mis en exergue des blocs de variables très corrélées entre elles (coefficient absolu supérieur à 0.8). Afin de réduire le nombre de variables, nous avons fait le choix de n'en conserver qu'une seule par bloc. En outre, nous avons choisi d'écarter les variables comprenant trop de valeurs manquantes irremplaçables ou de modalités concernant un nombre trop faible d'observations. Nous avons dégagé de cette première étape une présélection de 13 variables quantitatives et 9 variables catégorielles.

Dans un second temps, nous nous sommes concentrés sur les liens spécifiques entre chacune des variables et notre variable à prédire. Nous avons pour cela tracé le taux de clic observé dans la base en fonction des modalités de nos variables. Ces tracés nous ont confortés dans notre présélection de variables – nous laissant penser qu'elle contient un pouvoir explicatif intéressant du clic – et ont montré des liens positifs entre les variables présentées en figure 1 et le taux de clic.

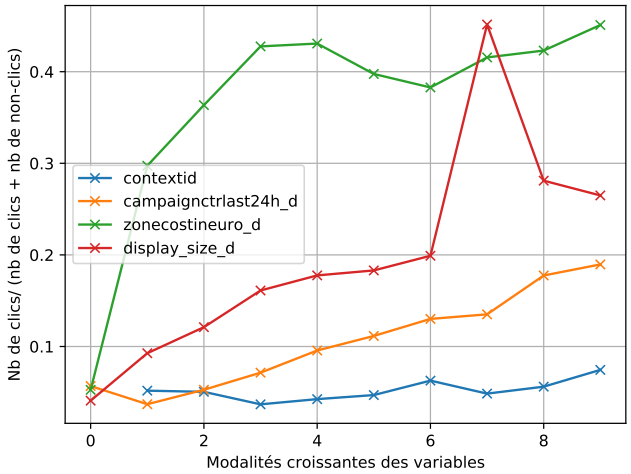

FIGURE 1 – *Tracé de la proportion de clics en fonction de quatre caractéristiques quantitatives observées (l'engagement marketing de l'utilisateur, la performance de la campagne publicitaire durant les dernières vingt-quatre heures, le coût de l'emplacement publicitaire et la taille de l'affichage).*

En complément de ces statistiques multivariées, nous avons entrepris une technique de réduction du nombre de variables quantitatives. L'objectif de la méthode utilisée, appelée analyse en composantes principales (ACP), est de créer à partir des variables originelles de nouvelles variables décorrélées contenant chacune une part décroissante de l'information disponible (la première variable créée a le plus fort pouvoir explicatif ; la dernière le plus faible). L'ACP produit donc une synthèse de nos variables en réduisant la dimension : cela permet de réaliser un nuage des points des observations dans le plan ou dans l'espace. Le plan permettant de mieux discriminer les clics des non-clics est donné en figure 2.

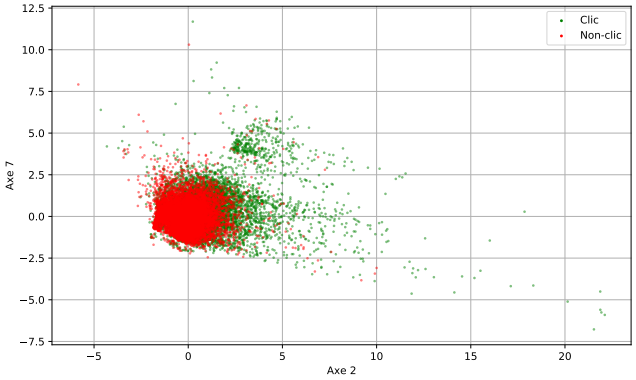

FIGURE 2 – *Nuage de points des observations dans le plan formé par les deuxième et septième variables synthétiques obtenues par ACP sur les variables quantitatives présélectionnées. Un nombre égal de points correspondant à un clic (en vert) et à un non-clic (en rouge) est représenté.*

En étudiant les corrélations des variables originelles avec les nouvelles variables synthétiques, on constate à nouveau que les variables évoquées dans la figure 1 ont un fort pouvoir discriminant – les autres variables ont un effet plus mitigé qui est plus difficile à capter avec cette méthode. On s'attend donc à ce que ces variables soient utiles et importantes pour nos modèles de prédiction.

3 Quels sont les modèles de prédiction de clic utilisés et comment les construit-on ?

Afin de réaliser nos prédictions, nous avons eu recours à trois modèles de classification binaire récurrents en *Machine Learning*, à savoir la régression logistique, les forêts aléatoires et le modèle de *gradient boosting* XGBoost, dont nous présentons les fondements théoriques en annexe de notre rapport.

Néanmoins, avant d'implémenter nos modèles, il nous a fallu procéder à une phase de sélection de variables explicatives afin de réduire le volume de calculs à réaliser. Après avoir testé différentes méthodes, nous avons jeté notre dévolu sur la Recursive Feature Elimination with Cross-Validation (RFECV), laquelle permet de sélectionner les meilleures variables pour un modèle donné à partir d'une métrique de performance que nous définirons dans la prochaine section. Nous avons ainsi pu compléter les premiers résultats issus de l'analyse descriptive en réduisant notre liste de variables à une dizaine environ pour chaque modèle.

Par ailleurs, nous faisons face ici à un jeu de données extrêmement déséquilibré, au sein duquel la classe minoritaire, associée au clic, représente à peine

6% des observations de la base. Ce type de déséquilibre, au-delà du ciblage publicitaire, est également présent dans des problématiques telles que la détection de fraudes. Les modèles que nous avons utilisés tendent alors à classer naïvement une immense majorité des observations dans la classe majoritaire, autrement dit à prédire massivement des non-clics, et présentent ainsi une faible performance pour les métriques pertinentes d'un point de vue métier. C'est la raison pour laquelle nous avons fait appel à des techniques de sur-échantillonnage, offrant la possibilité de rééquilibrer notre jeu de données.

4 Comment mesurons-nous la performance de ces modèles ?

Notre objectif est de construire un modèle prédictif suffisamment efficace en termes de performances brutes et assurant à l'entreprise une rentabilité suffisante. Il nous est donc apparu essentiel de définir une métrique de référence à même de mesurer ces performances et de comparer nos différents modèles.

Pour une entreprise comme Criteo, rémunérée au clic, prédire un faux positif (autrement dit prédire à tort un clic) revient à perdre le coût de l'emplacement publicitaire acheté, mais demeure moins coûteux que la prédiction d'un faux négatif, c'est-à-dire la prédiction qui s'avère finalement erronée d'un non-clic. En effet, Criteo perd alors très vraisemblablement la rémunération du clic de l'utilisateur, alors même qu'il s'agit de sa seule source de revenu. La doctrine consiste donc à éviter de prédire des non-clics qui seraient en réalité des clics, quitte à prédire trop de clics. Pour cela, il apparaîtrait naturel de prêter une grande attention au *recall*, qui indique la proportion de clics bien prédits parmi l'ensemble des clics effectifs. Néanmoins, prédire un clic à tort représente tout de même un coût non négligeable pour l'entreprise, c'est pourquoi il peut aussi être intéressant de s'appuyer sur la *precision*, qui comme son nom l'indique confère plus d'importance à la véracité de la prédiction (proportion de clics bien prédits parmi l'ensemble des clics prédits).

Afin de tenir compte de ces deux enjeux, nous avons finalement opté pour un score hybride, en l'occurrence le score F_3, permettant de considérer ces deux métriques tout en assignant trois fois plus de poids au *recall* qu'à la *precision*, et ainsi de répondre aux impératifs économiques de l'entreprise. Dans la comparaison des résultats issus des différents modèles qui va suivre, nous chercherons donc à maximiser ce score, lequel constitue désormais notre métrique de référence.

Par ailleurs, pour répondre aux exigences des enchères, il nous faut aussi minimiser le temps de prédiction de nos algorithmes : l'idée est donc d'obtenir le meilleur compromis entre le score de performance d'une part et le temps de prédiction d'autre part.

5 Quel est le modèle le plus performant ?

Les trois modèles que nous avons utilisés possèdent de nombreux hyperparamètres, qu'il faut judicieusement choisir afin d'optimiser les performances et donc de maximiser le score F_3. Ce choix a été effectué par l'intermédiaire d'une méthode de *grid search* combinée à une validation croisée, laquelle consiste à parcourir un ensemble de combinaisons de paramètres puis à sélectionner la meilleure combinaison selon le score F_3. Ainsi, pour chacun des modèles, nous avons mis en œuvre plusieurs fois cette méthode, avec utilisation ou non des variables sélectionnées par la méthode RFECV et d'une méthode de sur-échantillonnage.

Nous avons alors pu comparer les performances des meilleures combinaisons de paramètres obtenues pour chaque modèle et chacune des leurs variations (RFECV et sur-échantillonnage), ce qui nous a permis de sélectionner une variation par modèle selon un compromis entre le score F_3 et les temps de prédiction et d'apprentissage. Dans les trois cas, il s'agit du modèle avec application de la RFECV et sans recours à une méthode de sur-échantillonnage.

Pour améliorer la performance de nos algorithmes, nous avons ensuite ajusté le seuil de décision. En effet, les modèles de classification permettent ici de calculer la probabilité d'un clic pour l'observation considérée, le clic étant prédit dès lors que la probabilité excède 0.5, valeur fixée par défaut. Néanmoins, rien n'indique que ce seuil est optimal du point de vue de la métrique qui nous intéresse, c'est pourquoi nous avons donc tracé, pour la meilleure variation de chaque modèle, l'évolution du score F_3 en fonction du seuil de décision. Il est ainsi apparu que la régression logistique et le modèle XGBoost proposent des performances maximales pour des seuils proches de 0.5, tandis que seul un ajustement du seuil à 0.64 pour le modèle Random Forest semble augmenter significativement la performance, comme en témoigne la figure 3 ci-dessous.

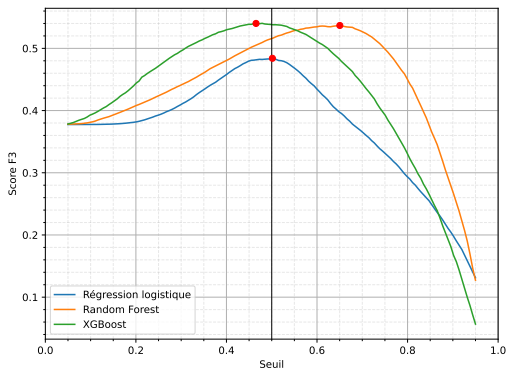

FIGURE 3 – *Représentation du score F_3 en fonction du seuil de décision pour chacun des trois modèles.*

Une fois les meilleures combinaisons de paramètres déterminées et les seuils ajustés, nous avons pu comparer les performances des trois modèles sur la base de test. Le modèle Random Forest ne nous a pas paru pertinent dans la mesure où il n'apporte pas un gain de performance significatif à même de compenser un temps de prédiction très long, à la différence des deux autres modèles, intéressants pour notre étude : la régression logistique offre une performance décente avec un score F_3 de 0.486 et un temps de prédiction très court de 0.01 seconde, complétés par un temps d'apprentissage très économe, tandis que le modèle XGBoost présente une bonne performance (0.535 pour le score F_3) avec un temps de prédiction relativement réduit de 0.06 seconde. Ces deux spécifications semblent donc répondre aux attentes de l'entreprise Criteo, tant sur le plan de la performance brute que des contraintes temporelles inhérentes au système d'enchères.

6 Quels sont les impacts des contraintes liées à la quantité de données et au temps d'évaluation ?

Après avoir isolé les meilleurs modèles en ajustant leurs paramètres et les seuils de décision, nous avons enfin cherché à mesurer l'influence de la réduction de la quantité des données du jeu d'entraînement sur la performance et les temps de calculs. Nous nous sommes donc intéressés à l'évolution du score F_3 en fonction du pourcentage d'observations, en sélectionnant à chaque fois aléatoirement un certain pourcentage de données. Il est bien évidemment apparu que la performance recule à mesure que le volume de données diminue (figure 4), mais dans une moindre mesure pour la régression logistique et le modèle XGBoost, alors que les forêts aléatoires présentent une chute vertigineuse de leur performance pour des pourcentages de données très faibles.

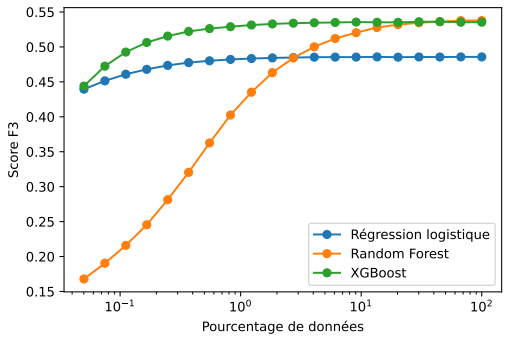

FIGURE 4 – *Représentation du score F_3 en fonction du pourcentage de données pour chacun des trois modèles (échelle logarithmique en abscisse).*

Nous avons ensuite intégré la dimension liée aux contraintes temporelles à notre analyse, en regardant à quel point il est possible de limiter les temps de prédiction et d'apprentissage des différents modèles sans trop altérer la performance brute et donc le score F_3. Les conclusions précédentes restent encore valables : la régression logistique et le modèle XGBoost sont plus robustes à une baisse drastique du temps d'apprentissage (figure 5), qui peut être abaissé respectivement de 3 à 0.1 seconde et de 10 à 1 seconde sans trop compromettre le score F_3. En ce qui concerne le temps de prédiction, dont la minimisation est fondamentale pour Criteo, nous avons pu observer que son ordre de grandeur reste le même quand la quantité de données augmente (figure 6), et qu'il dépend surtout de l'algorithme utilisé : 0.01 seconde pour la régression logistique, 0.1 seconde pour XGBoost et 1 seconde pour Random Forest. Ici aussi, les performances de la régression logistique et du modèle XGBoost semblent bien se maintenir pour des temps de prédiction plus faibles.

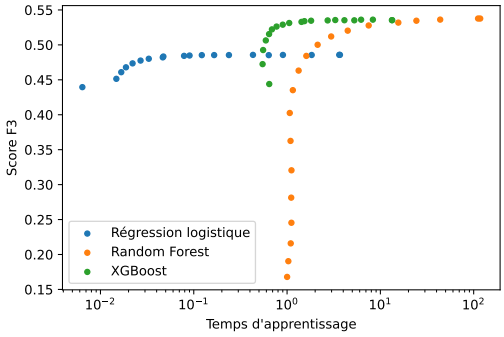

FIGURE 5 – *Représentation du score F_3 en fonction du temps d'apprentissage pour chacun des trois modèles (échelle logarithmique en abscisse).*

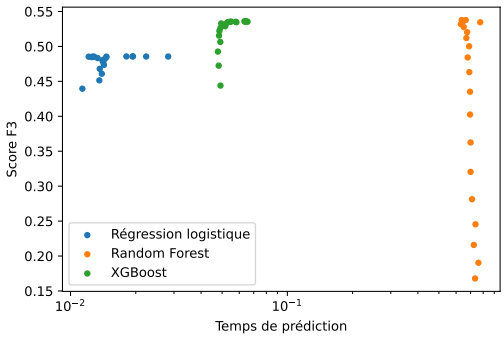

FIGURE 6 – *Représentation du score F_3 en fonction du temps de prédiction pour chacun des trois modèles (échelle logarithmique en abscisse).*

Dans les figures 5 et 6, les points sont associés aux différents pourcentages de données étudiés en figure 4. On lit donc en abscisse, pour chaque pourcentage de données, les temps d'apprentissage et de prédiction correspondants.