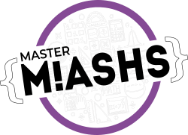
Une image contenant texte

Description générée automatiquement



**Modélisation du comportement utilisateur dans la publicité mobile**

**Projet tutoré : 2019 - 2020**

***Encadré par:***

* *Faustine BOUSQUET*
* *El hassan DAOUDI*
* *Samy ZAROUR*

***Par:***

* *Joseph AKA BROU*
* *Marwa ELATRACHE*
* *Caroline MARTIN*
* *Tharshika NAGARATNAM*
* *Omar SECK*

# Remerciements

Avant tout développement sur cette étude, il nous apparaît opportun de commencer ce rapport par des remerciements à ceux qui nous ont beaucoup appris mais également ceux qui nous ont suivis tout au long de notre projet.

Nous tenons donc à remercier l’équipe pédagogique, les intervenants professionnels ainsi que le secrétariat du département MIASHS de l’université Paul Valéry pour cette année enrichissante.

Nous voudrions également remercier nos encadrants de projet, Faustine BOUSQUET, El Hassan DAOUDI et Samy ZAROUR, pour leurs précieux conseils et avis. Tout au long de notre projet, leurs remarques constructives nous ont permis d’avancer dans notre travail et de consolider nos acquis.

# Sommaire

[Remerciements 2](#_Toc27321755)

[Sommaire 3](#_Toc27321756)

[Glossaire 4](#_Toc27321757)

[1. Introduction 5](#_Toc27321758)

[1.1 Contexte du projet 5](#_Toc27321759)

[1.2 Présentation de l’entreprise TabMo 5](#_Toc27321760)

[1.3 Présentation des objectifs et de la finalité 6](#_Toc27321761)

[2 Jeux de données 7](#_Toc27321762)

[2.1 Collecte de données 7](#_Toc27321763)

[2.2 Découverte de données 8](#_Toc27321764)

[2.3 Nettoyage et validation de données 9](#_Toc27321765)

[2.4 Transformation et enrichissement de données 10](#_Toc27321766)

[2.5 Stockage de données 10](#_Toc27321767)

[3 Exploration de données 11](#_Toc27321768)

[3.1 Analyse descriptive 11](#_Toc27321769)

[3.1.1 Axe User 11](#_Toc27321770)

[3.1.2 Axe Creative 11](#_Toc27321771)

[3.2 Analyse multidimensionnelle des données 11](#_Toc27321772)

[3.3 Vers un modèle de prédictif 11](#_Toc27321773)

[BIBLIOGRAPHIE 12](#_Toc27321774)

[Annexe 12](#_Toc27321775)

# Glossaire

SSP : Serveur Side Platform

DSP : Demande Side Platform

RTB : Real Time Bidding

AdRequest : Requête envoyé par le téléphone utilisateur au SSP

BidRequest : Requête envoyé par le SSP au DSP

BidResponse : Réponse du DSP pour donner suite au SSP

Impression : L’affichage d’une publicité après la Win d’une enchère.

Win : Enchère gagnée

Clic : Le fait qu’un utilisateur clic sur la publicité après l’impression

CPM : Coût par mille

Banner : Publicité de type image

UI : User interface

# Introduction

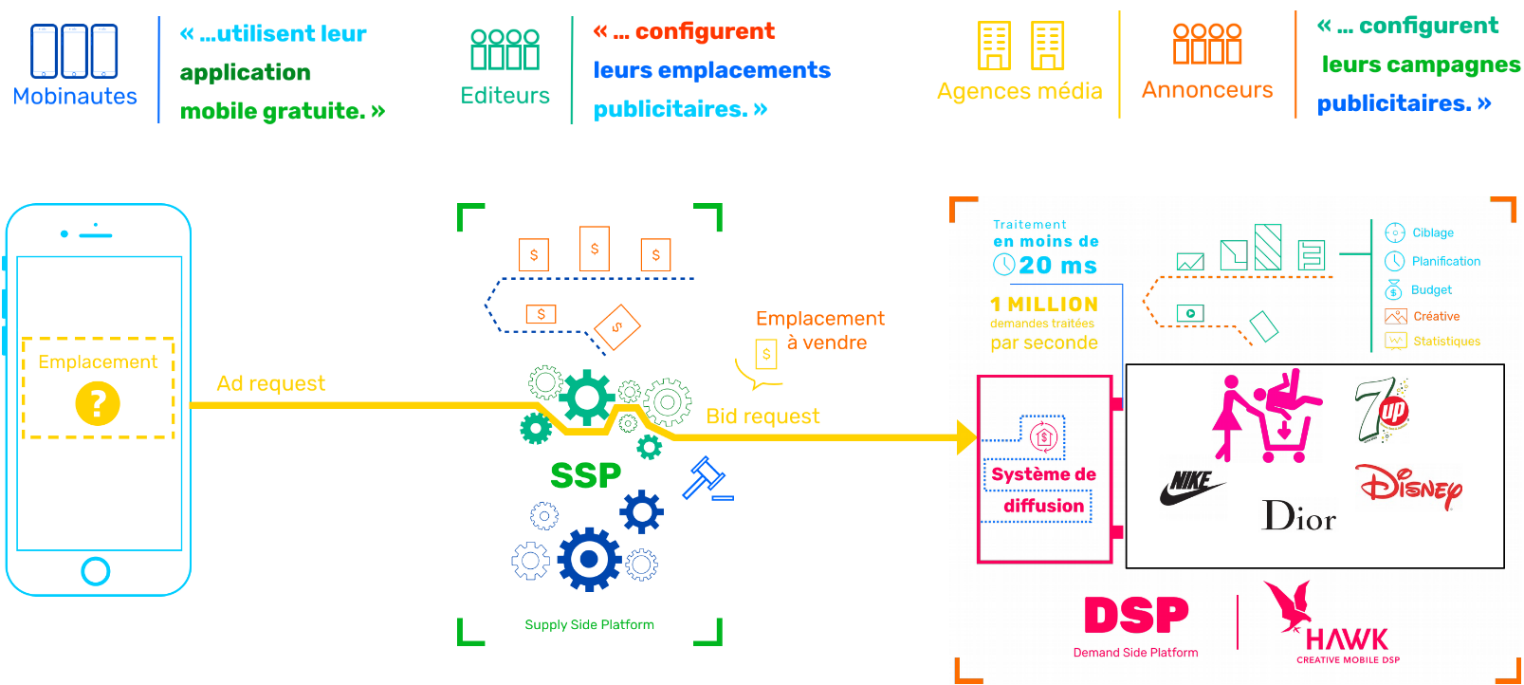
## **Contexte du projet**

Dans le cadre de notre Master 1 MIASHS, nous devons réaliser un TER (Travail d’Etude et de Recherche). Pour ce travail, le sujet qui a suscité notre intérêt est celui traitant de la modélisation du comportement utilisateur dans la publicité mobile. En effet, le mobile est aujourd’hui devenu un puissant instrument de marketing.

Par son intermédiaire, les annonceurs peuvent toucher chaque consommateur individuellement, de façon ciblée, à n’importe quel moment et où qu’il se trouve. Cela est d’autant plus vrai pour notre génération, celle des réseaux sociaux, sur laquelle les annonceurs ont aujourd’hui accès à un panel d’outils relativement large pour transmettre un message personnalisé.

C’est donc ce sujet d’étude proposé par la société TabMo sur lequel nous nous sommes lancés depuis le mois de septembre. A travers cette étude, notre intention est dans un premier temps d’expliquer ce qui conduit l’utilisateur à s’intéresser à une publicité.

## **Présentation de l’entreprise TabMo**

TabMo Labs a été créé en fin 2013. L’activité de l’entreprise est l’achat d’espaces publicitaires sur tablette et mobile pour le compte d’annonceurs et d’agences media. Le marché étant en pleine expansion, ils décident début 2015 de prendre un tournant technologique et ainsi de développer leur propre solution, un DSP et son écosystème, adapté à la spécificité du programmatique mobile et ainsi rendre l’expérience utilisateur plus complète. S’ouvre alors TabMo Labs sur Montpellier, l’équipe en R&D en début 2015.

En tant que DSP, TabMo est en charge de la diffusion de campagnes publicitaires pour des annonceurs et agences media. Ces campagnes se composent d’annonces publicitaires, qui doivent être diffusées selon des critères souvent très spécifiques (géographie, liste d’applications, réseaux…). Les campagnes doivent de plus atteindre certains objectifs en termes de visibilité, ou d’actions des mobinautes tels que l’affichage, le clic, ou l’achat du produit. Le bon déroulement des diffusions des publicités est la mission d’un DSP.

Pour mener à bien cette mission, TabMo est en relation avec divers SSP. La mission de ces SSP TabMo consiste à envoyer une BidRequest lorsqu’un emplacement publicitaire est disponible. La disponibilité de cet emplacement se traduit elle-même par une AdRequest qui est envoyée depuis de le mobile de l’utilisateur vers le SSP. Ainsi, tout comme d’autres DSP, TabMo reçoit près de 500 000 BidRequest par seconde. Pour départager les DSP, et donc déterminer lequel va pouvoir afficher la publicité de son annonceur, une enchère à lieue. Pour chaque enchère, TabMo a au préalable le montant que l’annonceur souhaite mettre pour afficher sa publicité. En fonction des montants qui lui sont alloués, TabMo envoie environ 500 BidResponses qui elles-mêmes donnent approximativement à 70 wins par seconde.

## **Présentation des objectifs et de la finalité**

Lorsqu’une enchère est gagnée, TabMo a la possibilité de diffuser une annonce sur l’espace publicitaire qui vient d’être remporté. Cela représente près de 14% des BidRequest que reçoit TabMo.

Lorsqu’une publicité a été affichée sur le mobile de l’utilisateur, on parle d’impression. A cet instant, l’utilisateur qui reçoit l’affiche publicitaire, sous forme de vidéo, d’image ou audio, est libre de la **regarder**, de la **visionner**, de **l’écouter** et surtout de **cliquer** sur la publicité si celle – ci l’intéresse.

Ce sont justement ces phénomènes que nous avons voulus comprendre dans le but de les expliquer et de les prédire. En effet, il est intéressant pour TabMo d’anticiper et savoir si l’utilisateur sera intéressé ou non par le contenu de la publicité. En fonction de cette information TabMo pourrait optimiser ses enchères afin de réduire les coûts pour ses annonceurs.

Au travers ce rapport, vous aurez donc l’opportunité d’apprécier tous les outils et méthodes que nous avons mis en place afin de répondre à cette problématique. Tout au long de votre lecture, vous pourrez également voir comment nous nous sommes organisés afin de synchroniser nos différents travaux et résultats. Enfin, nous vous ferons part de nos difficultés rencontrées puis nous conclurons ce rapport en vous faisant part de nos axes d’améliorations.

# Jeux de données

## **Collecte de données**

Lorsque l’on ouvre une application mobile, il se peut que l’on observe des publicités. Ces publicités sont le résultat d’un système complexe qui tourne en arrière-plan. On parle de RTB, l’achat d’espace publicitaire en temps réel. Le principe est simple, un annonceur souhaite que sa publicité soit affichée quand un internaute cherche le mot clé « XYZ », il va donc allouer un budget (une enchère) pour « acheter » ce mot clé.

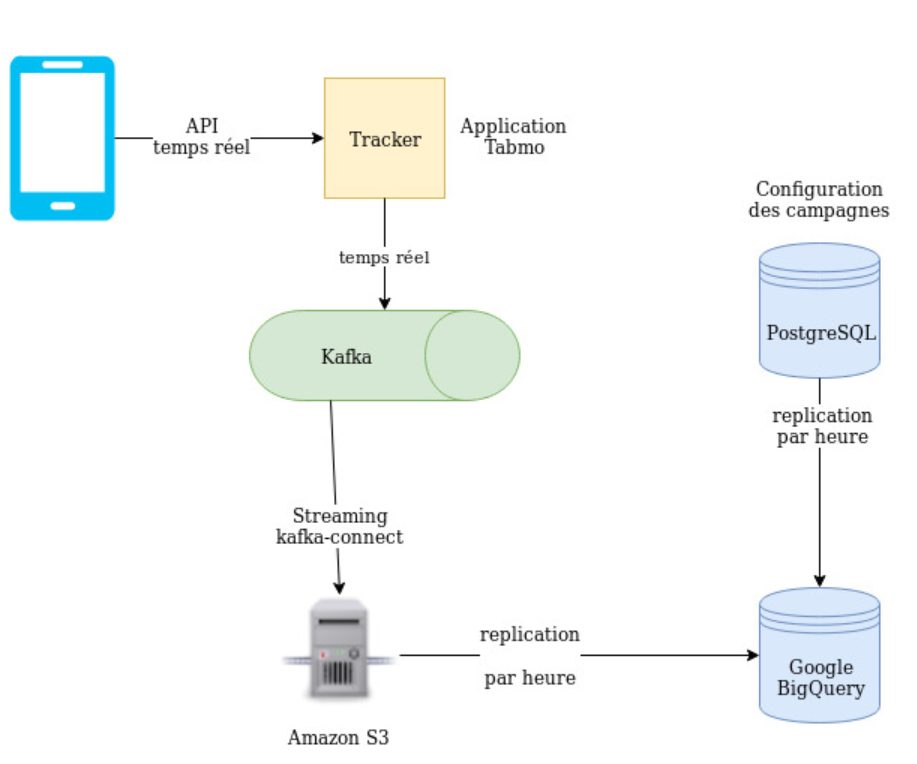
Au début le RTB a surtout été vu par les annonceurs comme une opportunité de monétiser les « invendus », autrement dit les espaces que l’on n’était pas parvenu à vendre directement à un annonceur spécifique. Pour les annonceurs le RTB était la promesse d’acheter de l’espace moins cher mais par définition moins qualitatif.

Les choses ont rapidement évolué avec le développement de nouveaux acteurs et outils comme les SSP et DSP comme TabMo qui permettent d’optimiser l’achat des espaces publicitaires. Au cours de leur échange, le SSP et le DSP échange des informations. Durant cet échange qui dure en moyenne 200 millisecondes, plusieurs systèmes de gestion de données entre en jeu.

En effet, TabMo dispose d’un système de gestion de données « Kafka » qui permet de stocker les données à la chaine sur plusieurs machines. Le travail est parallélisé et cela permet d’optimiser l’accès à la donnée. Parmi les données enregistrées, on retrouve :

* Les infos sur les enchères (500 par seconde)
* Les infos sur les BidRequest (500 000 par seconde)

En plus de ces données, on retrouve également des données brutes partitionnées par année, mois, jour et heure dans un Data Lake Amazon S3. Les données de campagnes des annonceurs (Profil visés, date de début, date de fin, …) sont également stockées. Les données concernant les enchères gagnées sont quant à elles stockées dans un datawarehouse Google BigQuery.

BigQuery est un entrepôt de données cloud sans serveur, hautement évolutif et économique, qui est doté du service d'analyse en mémoire BI Engine et intègre des fonctionnalités de machine learning. Dans le cadre de notre TER, c’est sur les données de ce datawarehouse que nous avons eu à travailler. Ainsi, nous avons récupérer un dataframe issu d’une requête BigQuery de Samy Zarour, l’un de nos encadrant TER.

## **Découverte de données**

Le projet s’est déroulé en plusieurs étapes :

* Première étape : comprendre les problématiques métier, identifier les données disponibles et commencer à extraire des données.
* Deuxième étape : explorer les premières données disponibles, identifier des problématiques auxquelles il est possible de répondre avec ces données, identifier leurs limites, mettre en place un plan d’action
* Troisième étape : répondre au plan d’actions défini pour donner suite à la réunion d’avancement du X Novembre 2019. Le plan d’action était composé de deux axes d’études : Un premier axe orienté utilisateur puis un second orienté sur le type de la publicité.

Pour ce projet, nous avons rapidement pris l’initiative de nous créer une espace de travail organisé, composé d’un Trello pour gérer les différentes tâches à mener. D’un GitHub pour nous partager le code, ainsi qu’un agenda avec des réunions régulièrement planifiées avec nos responsable de TER.

Ces différentes réunions nous ont permis de récupérer de précieuses informations. Parmi ces informations, nous avons eu un dictionnaire de données que vous retrouverez en page d’annexe \_. Ce dictionnaire complété des explications de nos tuteurs nous a permis cerner l’utilité de chaque variable.

Dans un premier temps, nous avons reçu un jeu de données traitant des flux publicitaires sur une journée. Par l’intermédiaire de ce jeu de données nous avons pu avoir un aperçu des données concernant les 45 variables présentent. L’illustration \_ en annexe est un exemple de tuple du jeu de données.

Le tuple que l’on observe sur cette illustration concerne une enchère qui a été gagnée. Pour cette enchère, nous disposons de diverses informations que nous avons pu catégoriser. Parmi ces informations, nous retrouvons :

* **Des informations sur l’enchère en elle-même**

Comme par exemple l’heure de l’enchère, son identifiant, le prix auquel elle a été remportée

* **Des informations sur la publicité qui a fait suite à l’enchère**

Dans ces informations, nous disposons de l’identifiant de la publicité, l’identifiant de son annonceur, le type de publicité (image, vidéo, …)

* **Des informations sur le site qui a hébergé la publicité**

Nous avons le nom du site, nous savons également si c’était sur une application où une recherche web. Nous disposons également d’une variable IAB qui nous permet de retrouver le type du site (Musique, réseau social, shopping, cuisine, …).

* **Des informations sur le mobile de l’utilisateur**

Parmi ces informations on retrouve le système d’exploitation, la marque du modèle, ou encore le langage du mobile.

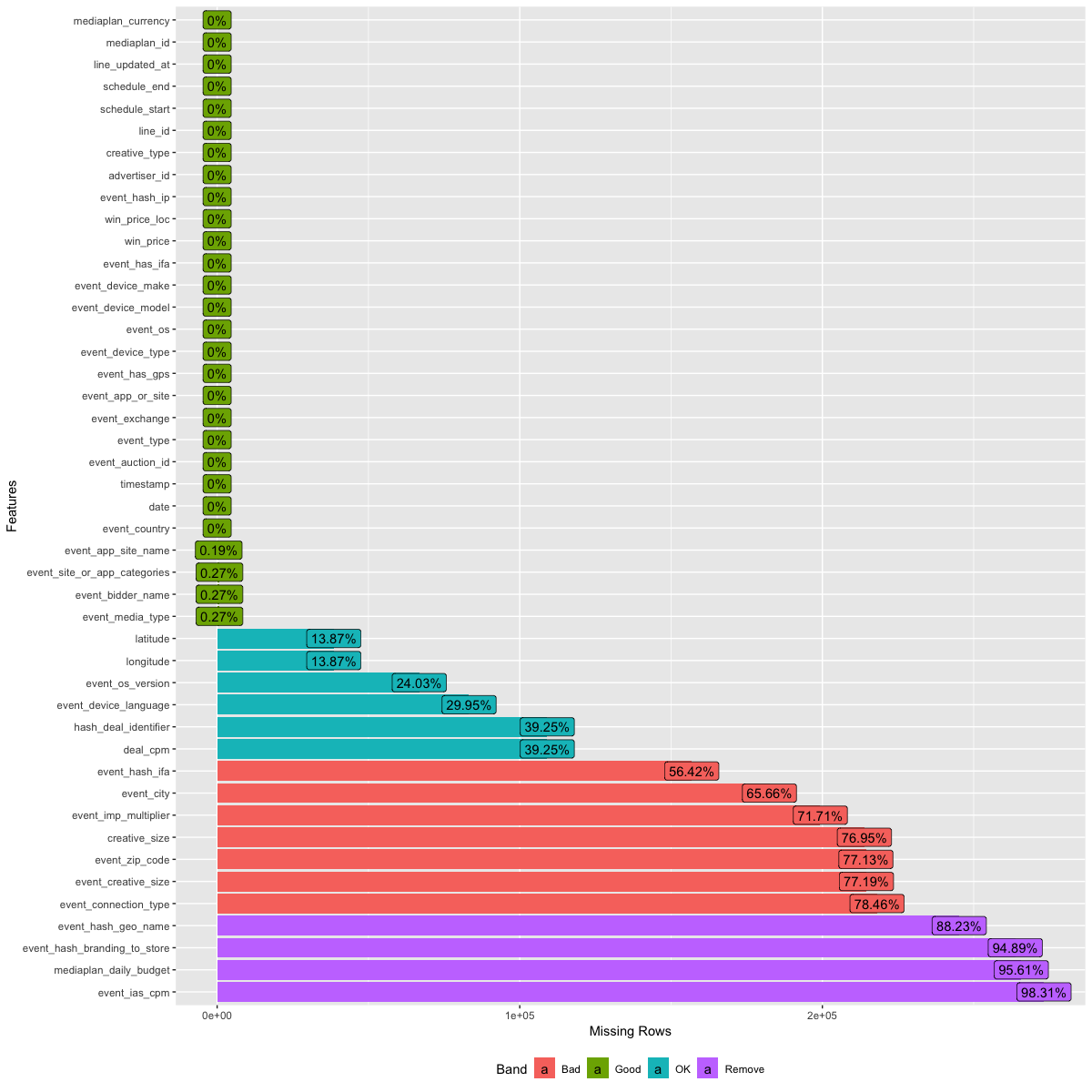
* **Des informations sur la localisation du mobile**

Le pays, la ville, le code postal, la longitude, la latitude.

Comme il a été mentionné, chaque enchère du jeu de données correspond à une enchère qui a conduit à une impression. En revanche, on retrouve très peu d’enchère qui ont mené à un clic. De ce fait, nous avons par la suite, demandé des jeux de données avec davantage de données sur plusieurs jours et plusieurs semaines. Il est important de préciser que les différents jeux de données que nous avons reçus avaient exactement la même architecture.

## **Nettoyage et validation de données**

Au cours de notre étude descriptive, nous avons pu observer un certain nombre d’erreurs, de données peu cohérentes et peu qualitatives en termes d’informations. Par exemple, nous avions observé des doublons et des variables avec des valeurs manques. Par ailleurs, nous avons également constaté que l’interprétation d’une publicité de type image n’était pas identique à celle d’une publicité vidéo ou audio. Pour enter dans les détails, une image est juste « vue » tandis qu’une vidéo peut être regardé au premier quartile, au second, voir jusqu’à la fin. Aussi, dans le cadre de la première partie de notre projet nous avons pris l’initiative de nous focaliser sur les enchères ayant donné suite à des publicités de type banner.

Dans les faits, pour réaliser cette action de nettoyage, nous avons au préalable étudié le pourcentage de valeurs manquante de chaque variable. C’est cette information que nous avons synthétisée pour vous dans l’illustration ci - contre sur laquelle on peut voir le pourcentage d’informations manquantes pour les variables qui ont des données manquantes.

Comme on peut le voir, certaines variables de localisation comme le code postal(event\_zip\_code), la ville(event\_city), longitude, latitude, n’ont pas toutes la même qualité de renseignement. De ce fait, pour localiser une enchère, nous nous sommes contentés de ne conserver que les variables de longitude, latitude et la variable country qui n’a aucune valeur manquante.

Par ailleurs, certaines variables sont quasiment vides et donc inutilisables. Pour donner suite à nos réunions avec nos tuteurs, nous avons eu confirmations de pouvoir supprimer ces variables. Nous avons également supprimé les variables que nous avons jugé peu pertinentes pour la suite de notre étude. Parmi ces informations supprimées, on retrouve par exemple :

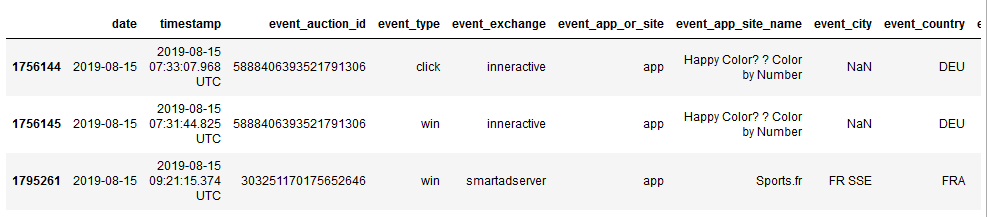
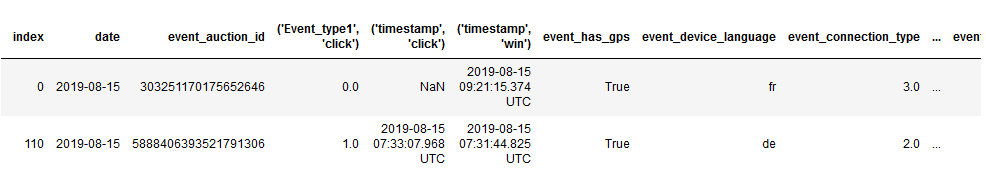
* La ville de l’utilisateur du mobile
* L’identifiant de l’annonceur
* L’identifiant de la campagne
* Le code postal
* La taille en pixel
* Le CPM
* La version du système d’exploitation

## **Transformation et enrichissement de données**

Par la suite, nous nous sommes également intéressés à chaque enchère. En effet, il n’est pas normal d’observer deux fois une même enchère (banner) dans le jeu de données. Pour être plus exacte, nous avons vu avec nos tuteurs qu’il n’existe qu’une condition pour qu’une enchère soit présente deux fois dans le jeu de données : qu’elle ait été cliquée. En effet, dans le cas où une impression a donné suite à un clic, une seconde ligne répétant les informations de l’enchère sera présente dans le jeu de données. Pour observer cela il nous suffit de nous intéresser à la valeur de la variable « event\_type » : si la modalité est « win » alors il aura les informations sur l’enchère et la publicité, si c’est « clic » alors nous avons ces mêmes informations avec en plus la date du clic.

Autrement dit, il est possible d’observer deux fois une enchère avec une première fois la modalité « event\_type » qui vaut « win » et une seconde fois qui vaut « clic ». En revanche, il n’est pas normal d’observer plus d’une fois un même numéro d’enchère avec à chaque fois « win » dans « event\_type ». En ce sens, nous avons éliminé du jeu de données toutes les enchères pour lesquels nous observions plusieurs win. Par la même occasion nous avons supprimé les enchères pour lesquels la date du clic arrivait avant la date d’impression

Pour synthétiser l’information, nous avons décidé de transposer la variable « event\_type » afin de n’obtenir qu’une seule ligne pour chaque enchère. Ainsi, pour chaque enchère, nous avons obtenu une seule ligne d’information à laquelle nous avons rajouté une variable qui nous informe si cette enchère traite d’une impression qui a donné suite à un clic ou non. Nous sommes donc passé d’un jeu du de données de l’illustration \_, à un jeu de données comme l’illustration \_.



En parallèle à cela, nous avons également binarisé la variable hash\_dealer. En effet, cette variable contient beaucoup de valeurs manquantes. Si dans un premier temps nous avons pensé à supprimer cette variable, nous avons dans un second temps compris l’intérêt qu’elle pourrait nous apporter grâce à nos tuteurs. En effet, la variable hash\_dealer est une variable qui contient un hash dans le cas où l’enchère résulte d’une enchère privée (partenariat entre TabMo et le site où il y a l’espace publicitaire) et une valeur manquante dans le cas contraire.

## **Stockage de données**

Pour des raisons de confidentialités évidentes, nous n’avons pas mis les données sur GitHub. En effet, les données que nous avons reçues de TabMo sont des données sensibles. De ce fait, comme toutes les entreprises travaillant avec les données utilisateurs, TabMo est soumis au RGPD. Ainsi, afin de capitaliser nos travaux et nos jeux de données, nous nous sommes servis d’un dépôt drive. Cela nous a permis de répartir les tâches et donc optimiser nos différents travaux. De plus, par l’intermédiaire du drive, nous avons également pu nous partager nos code Python (Colab) et donc centraliser les différents travaux qui avaient été réalisés par chacun.

# Exploration de données

## **Analyse descriptive**

Pour restituer nos différentes analyses, nous avons pris le soin de développer une application [**shiny**](https://marwa10.shinyapps.io/TER_pub/?fbclid=IwAR1_8fAaiqKdDCN-O0mjpNr04cf_Zlh1Z0KpSjB0penSH6DQDzetaJcJvgA). Cette application reporting nous a permis de proposer une plateforme à nos tuteurs afin qu’ils puissent voir l’avancée de nos travaux. Pour nos études descriptives, nous avons défini deux axes d’études, un premier orienté utilisateur, et un second orienté sur la publicité.

En ce qui concerne son aspect technique, notre application se décompose en deux parties.

* Une partie UI qui regroupe tous les éléments de mise en forme et d’affichage de l’interface utilisateur.
* Une partie Server où sont exécutés les codes R qui servent à produire les calculs, graphiques et représentations en générale.

### **Axe User**

Le travail consiste à explorer les données afin de répondre à ce type de question :

* Quel type d'utilisateur clique le plus ? (Ceux qui ont un iphone ? ceux qui naviguent la nuit ? ...)
* Existe-t-il un profil d'utilisateurs frauduleux ? (On peut définir une fraude comme une suite d'évènements trop rapprochés pour être générés par un humain)
* Les utilisateurs qui se connectent entre 20h et 23h sont-ils plus propices à cliquer ? A voir des vidéos ?

### **Axe Creative**

Ici, on se concentre sur le type de publicité.

* Quelles publicités génèrent une conversion ? Pour les pubs classiques, cela correspond à un event\_type de type click. Pour une vidéo on peut creuser et se demander plutôt est-ce qu'on peut savoir quel type de pub génère un évènement négatif (pause, mute...) ou un évènement positif (midpoint, resume...)
* Une pub diffusée aux heures de pointe a-t-elle plus de clic ?
* Une pub associée à un deal a-t-elle un taux de clic plus élevés que celles en open Auction ?

## **Analyse multidimensionnelle des données**

# Nos perspectives

Pour la suite de notre projet, nous envisageons de poursuivre notre étape de feature engineering. En effet, avant de nous lancer dans la conception de modèle, nous souhaitons optimiser la qualité de nos informations. Il s’agit de continuer à créer de nouvelles variables, regrouper des modalités, dichotomiser les variables ou encore agréger les données.

Une fois cette étape terminée, nous envisageons de mettre en place une méthode de machine learning dans l’objectif de prédire si oui ou non une enchère gagnée sera cliquée. Dans notre cas, il s’agit de mettre en place une méthode supervisée avec validation croisée. Pour cela, nous avons déjà des idées de méthodes que nous pourrions mettre en œuvre.

Pour commencer, nous envisageons des méthodes simples que nous allons pouvoir valider, refuser où améliorer. Pour cela, nous nous attendons à analyser les performances de ces modèles.

* Random Forrest
* Régression Logistique
* SVM

Un modèle va être considéré « bon » ou préférable à un autre modèle selon plusieurs critères potentiels. Le choix du meilleur modèle va donc dépendre de l’importance donnée à ces critères. En ce qui concerne, nos méthodes, les principaux critères sont :

* La matrice de confusion (Accuracy, Précision, Rappel, F1 - Score)
* La courbe ROC
* Du coût des différents types d’erreurs
* De la lisibilité du modèle (boîte blanche ou boîte noire)
* Du temps d’apprentissage et d’application du modèle
* De la pérennité et de la variabilité des variables explicatives dans le temps

Dans un second temps, nous souhaiterions nous projeter sur davantage de problématiques comme par exemple, déterminer si un clic est issu d’une ferme à clic. C’est en ce sens que nous allons continuer à travailler afin de vous présenter un TER abouti en fin d’année.

# 

# Bibliographie

* Document fourni par nos encadrants : « Introduction \_ TabMo, Programmatique & Données »



* TabMo : <https://labs.tabmo.io/tabmo-labs-lhistoire-d-une-équipe-r-d-montpelliéraine-e5ae3ca35106>
* WebMarketing : <https://www.webmarketing-com.com/2015/03/06/36022-programmatique-et-rtb-cest-quoi-la-difference>
* Cloud google: [https://cloud.google.com/bigquery/](file:///C:\Users\joaka\Documents\Master%20MIASHS%201\TER\cloud.google.com\bigquery\)

# Annexe