Rapport de stage

Optimisation des stratégies publicitaires par la data

Karim ASSAAD

Le 24 /09/2016

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon stage et qui m'ont aidé lors de la rédaction de ce rapport.

Je tiens à remercier vivement mon maitre de stage, Mr Antoine BSAIBES, Consultant Data et Data Miner au sein de Weborama, pour son accueil, le temps passé ensemble et le partage de son expertise au quotidien. Grâce à sa confiance j'ai pu m`impliquer totalement dans les missions que j`ai reçu. Il fut d'une aide précieuse dans les moments les plus délicats.

Je remercie également toute l'équipe Data pour leur accueil, leur esprit d'équipe et en particulier Mr Fred Grelier le Data Officier de l'équipe, Mr Xavier De Colombel le Stratégiste de l'équipe, qui m'ont beaucoup aidé à comprendre les problématiques.

Table des matières

Chapitre 1

Introduction

Actuellement, le marketing digital occupe une position primordiale dans tous les secteurs. Il constitue une innovation dans le monde des nouvelles technologies et ne cesse d'évoluer. De ce fait, une bonne maitrise de cette discipline est très avantageuse pour les entreprises.

Dans ce contexte, je présente mon stage de deuxième année établi au sein de Weborama qui s'intéresse à la réalisation des projets au profits du marketing en utilisant la méthode de la classification, les modèles de prédiction et les analyses de données.

En cohésion avec tout ceci, je me suis fixé la tâche d'accomplir à perfection la réalisation des projets qui m'ont été adressés. Dans ce but, j'ai assimilé le principe de la classification, étudié les données disponibles et essayé d'améliorer les performances des modèles utilisés dans la prédiction des scores.

Dans ce rapport je vous fais part de l'avancement de mon travail selon le déroulement suivant :

Je commencerais dans la première partie par présenter le cadre du stage et quelques notions impérieuses à la compréhension du domaine du data mining

Dans une seconde partie, je décrirai les taches effectue tout au long de la durée du stage et dans la dernière partie j'établirai un bilan sur ma contribution personnelle aux différents projets.

Chapitre 2

Weborama

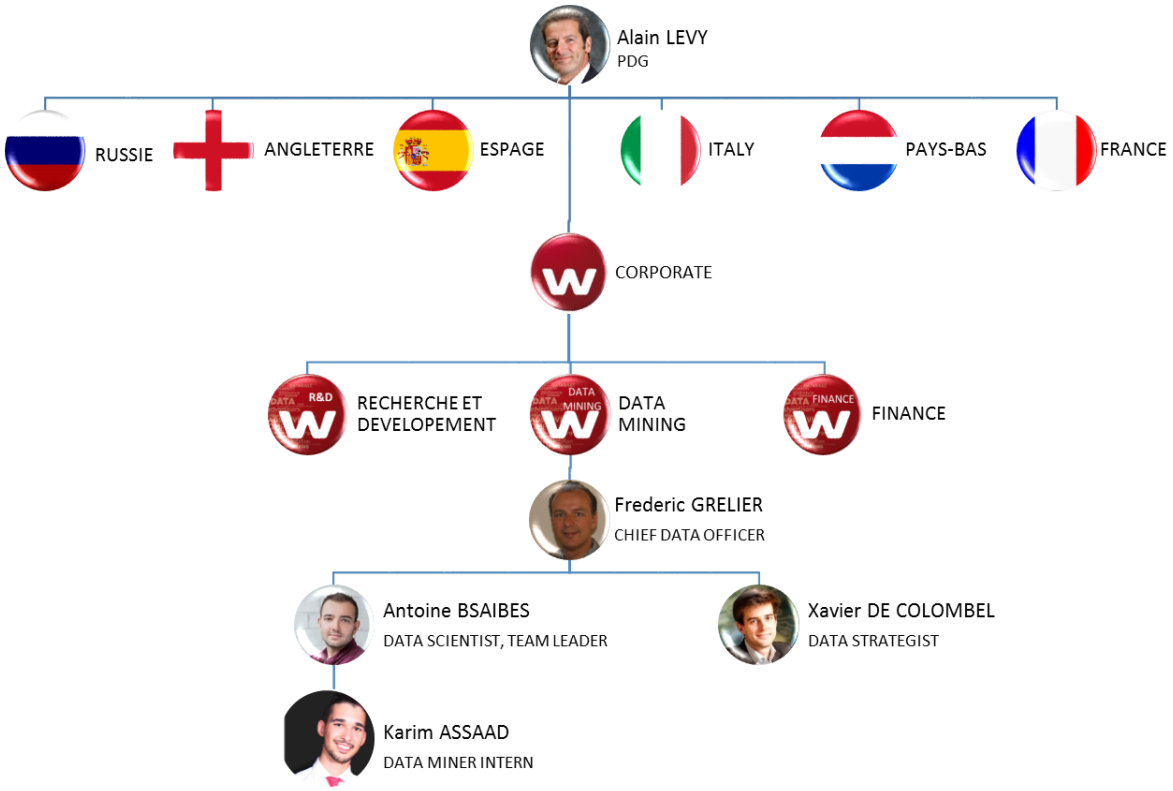
Présentation

Weborama est une entreprise Française créée en 1998, spécialisée dans la collecte des données marketing et la diffusion des campagnes publicitaires en ligne. La société développe des bases de données de profils et des solutions technologiques pour le marketing digital. À lors actuel, elle est présente dans plusieurs pays (notamment en Espagne). La société est membre de l'Interactive Advertising Bureau (IAB) et participe de façon active aux travaux de normalisation et d'autorégulation du marché publicitaire en France et en Europe.

Weborama est une Audience Driven Advertising Platform qui s'adapte parfaitement au nouvel écosystème publicitaire numérique et se traduit par une approche originale de la transformation de données brutes en valeur marketing.

Depuis 16 ans, l`entreprise a développé des actifs technologiques exceptionnels dans le domaine d'adserving et de la science de la data. Aujourd'hui, la société allie cette intelligence à sa base de données européenne de 400 millions de profils marketing et atteint ainsi un chiffre d'affaires de 25,8 millions d'euros en 2013.

Plus de 300 clients annonceurs, agences et éditeurs utilisent ses solutions innovantes pour piloter et optimiser leurs investissements en ligne en France, en Espagne, en Italie, au Portugal, aux Pays-Bas, à l'Angleterre et en Russie.



Contexte

La méthode Programmatique d'achat publicitaire fait référence à l'utilisation des logiciels pour acheter la publicité digitale qui remplace la méthode traditionnelle qui consiste en la négociation humaine et l'insertion des ordres manuellement.

Méthode classique

- Transaction de gré à gré
- Prix et volumes garantis
- Transparence de la marque média achetée
- Processus d'achat manuel

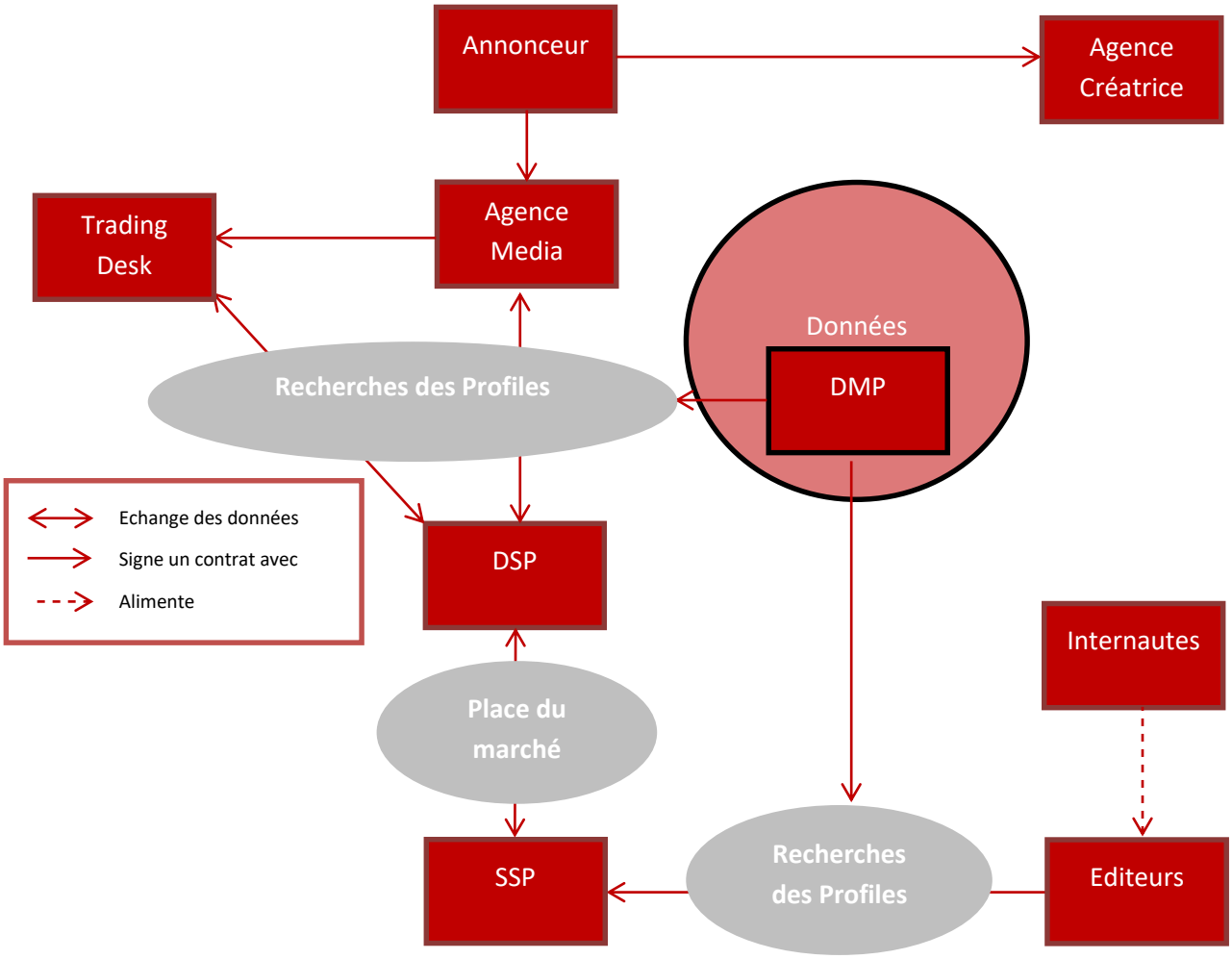
Nouvelle méthode

- Transaction aux enchères
- Prix et volumes variables
- Achat à l'aveugle ou avec une transparence limitée
- Processus d'achat automatisé

C'est l'utilisation des machines pour acheter les publicités, en effet cet écosystème permet de fiabiliser les systèmes d`achats, de les rendre moins chers et plus efficaces puisque les êtres humains ont besoin de se reposer contrairement aux machines.

Le marché de la programmatique est en croissance, il atteint actuellement 20 milliard d'euro à l`échelle internationale dont 1 milliard en France. Il est attendu qu`il atteigne les 37 milliard d'euro en 2019.

Le schéma suivant illustre cet écosystème



Annonceurs : Sont présentés par les clients qui ont un besoin d`advertising

Agences Créatrice : Responsable de la création de la publicité.

Agence Media : C`est l`agence qui conseille les annonceurs sur les choix de media planning et qui joue un rôle d` intermédiaire dans les procédures d'achat d'espaces publicitaires.

Trading Desk : C'est un acheteur médiatique massive et re-vendeur, qui fonctionne comme une unité de travail indépendant dans un grand achat préoccupation des médias.

DSP : Acronyme de Demande Side Platform, c'est un service permettant aux annonceurs, trading desk et agences d'optimiser leurs achats d'espaces publicitaires display. Les plateformes DSP les plus avancées fonctionnent en temps réels. Lorsqu'une campagne est programmée et définie à travers ses critères de ciblage par un acheteur, la plateforme d'optimisation recherche les impressions disponibles au meilleur coût.

SSP : Acronyme de Sell Side Platform, C'est une plate-forme technologique qui permet aux éditeurs de gérer leur inventaire publicitaire d'impression et de maximiser les revenus des médias numériques.

Editeurs : C'est les sites ou la publicité va être publiée.

Internautes : Les utilisateurs d'internet qui sont caractérisés par des cookies.

DMP : Acronyme de Data Management Plateforme, elle répond à un enjeu crucial des entreprises en leurs permettant de mieux connaître les prospects, de personnaliser les messages ou encore d'optimiser les campagnes et de réduire dépenses marketing.

1st, 2nd et 3rd party

- Les First Party : C'est des données "gratuites" puisque celles-ci sont collectées par l'entreprise sur ses médias propres (website, campagnes media, CRM, réseaux sociaux…) et alimentent des bases de données Cloud, les DMP pour Data Management Platform.

- Les Second Party : C'est des données qui viennent par l'intermédiaire d'un partenaire.

- Les Third Party : C'est des données vendues par des brokers et fournisseurs spécialisés. Ces données proviennent de sites qui commercialisent celles-ci, souvent des médias, réseaux sociaux ou des annonceurs eux-mêmes, ces sites et plateformes web jouissent d'un trafic dont les sessions sont nombreuses et longues pour collecter des informations pertinentes à propos des centres d'intérêts et thématiques qui intéressent chaque individu en ligne. Dans notre cas ces données proviennent de WCM, des sites reliés à la compagne publicitaire et les données achetées par l'annonceurs.

  Weborama a créé sa propre 3rd party qu'elle continue à développer et dont elle est propriétaire. Elles sont générées à partir du tracking de l'ensembles des URL (centaines de milliers par pays) où une probabilité sur une échelle de de 0 à 14 selon l'intensité de navigation est donnée pour chaque variable.

WCM

WCM (Weborama Campaign Manager) permet de créer les données, les collecter et de les convertir grâce à la diffusion et le tracking des campagnes publicitaires. WCM intègre également les fonctionnalités de trading desk via son module DSP. (WCM contient les données first party)

WAM

WAM (Weborama Audience Manager) est la DMP (Data Management Platform) de Weborama qui permet la création et la gestion des segments de ciblage marketing granulaires à partir des données 1st et 3rd party.

Problématique

Les Clients voulant bénéficier des nouvelles technologies de la data science de data pour optimiser leurs gains en terme d`advertising et de marketing, exige de réaliser des études approfondies sur un échantillon de plusieurs personnes afin de dégager leurs caractéristiques et propriétés spécifiques qui permettront par la suite de prévoir leurs comportements vis à vis certaines propositions ou spots publicitaires. Tout l`enjeux consiste en l`aptitude de la science de data de donner des prédictions exactes et précises et de permettre ainsi aux clients de bien mener leurs campagnes publicitaires et de mieux connaitre leurs cibles.

Chapitre 3

Rôle dans l'entreprise

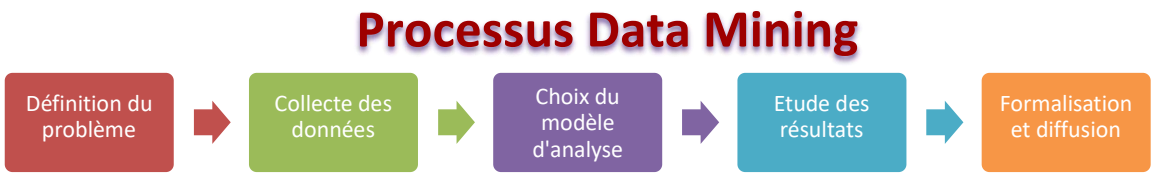
Description du stage

Durant le stage j`ai dû gérer les projets de deux clients principalement, dont le premier est leader dans le monde de l'automobile et le deuxième dans l'assurance.

Les activités effectueés durant mon passage à Weborama se résume comme suit :

- Création des scores spécifiques à la demande des clients.
- Segmentation des campagnes.
- Préparation des présentations sur PowerPoint.
- Recherche d`une solution qui permet d'activer les modèles et la classification.
- Conception d'une Library python qui facilite la démarche de la segmentation.

Description Data Mining



Le DATA MINING est un domaine reconnu mondialement et de plus en plus en croissance, il se pratiquait depuis 30 ans en faisant appel aux statistiques exploratoires et analyses de données qui sont considérées primordiales pour cette science.

Mais en réalité, ce n'est pas aussi simple que ça, le Data Mining introduit de nouveaux concepts qui sont loin d'être négligeables tels que les actuelles techniques d'analyse qui ne suivent plus la culture des statisticiens, mais sont en provenance de l'apprentissage automatique (intelligence artificielle, machine Learning), de la reconnaissance de formes (pattern recognition) et des bases de données.

L'extraction de connaissances est intégrée dans le schéma organisationnel de l'entreprise, ainsi, les données ne sont plus issues d'enquêtes ou de sondages mais proviennent d'entrepôts construits spécialement pour une exploitation qui a pour buts d`être analysée, ceci est appelé le DATA WAREHOUSE. Ceci dit, une réorganisation du flux de données au sein de l'entreprise devient nécessaire et la capacité des méthodes statistiques à traiter de gros volumes devient un élément clé.

Un dernier point important consiste au traitement des données qui sort de plus en plus des sentiers battus en traitant, on y trouve alors non seulement des fichiers plats "individus x variables", mais également des données sous forme non structurée, du texte, des images et des vidéos. On parle de fouille de données complexes. En effet cette orientation attribue une place primordiale à la préparation des données.

Déroulement du stage

Outils

Flux Cloud Platform / cloud computing

Le cloud est utilisé pour faire des requêtes SQL/nosql pour filtrer les données et créer des échantillons qui sont nécessaires pour réaliser le data mining.

R et Python

R était le langage de Data Mining utilisé à base par le département Data mais au cours du stage, le nouveau besoin d'activation nous a obligés de passer à python ce qui a simplifié la communication avec les serveurs.

Microsoft Excel

Après la modélisation, la classification et l'extraction des informations en utilisant plusieurs fonctions que nous avons codées, nous utilisons Excel pour l`analyse et pour mieux comprendre les résultats.

Microsoft PowerPoint

Microsoft Powerpoint est utilisé pour créer les diapos qui vont être présentés aux clients

Extraction des Données

L`extraction des données se fait à travers WCM et WAM, en effectuant des requête SQL dans une infra No SQL (recodage automatique pas le service cloud).

Nettoyage

La Deuxième partie du travail consiste à nettoyer les données après leurs imports dans R (Dans la Première partie du stage) et dans Python (Dans la deuxième partie du stage).

Le nettoyage en lui-même sert à gérer les valeurs manquantes, créer de nouvelles variables à partir d`un groupement d'entrées puis de les filtrer afin qu'elles deviennent plus significatives en terme d`analyse. Celles-ci peuvent être renommées par la suite ou mêmes suppriment selon leurs valeurs.

Le renommage s`effectuait au début d`une manière manuelle et prenait beaucoup de temps, mais grâce à la librairie que j'ai créée par la suite et qui assure plusieurs fonctionnalités, j'ai pu faire référence à une api qui prend en paramètre l'id des variables et renvoie leurs noms et ce pour traiter cette tâche et l'automatiser.

Modélisation

Quelques Notions

SVM

Le support Vecteur Machine (SVM) est un algorithme très puissant qui est capable de trouver des patterns fortement non linéaires. Il repose sur deux idées essentielles :

- La maximisation de la marge entre frontière de décision et les exemples les plus proches, qu'on appelle les vecteurs support.
- Le choix d'un hyperplan séparateur dans un espace de combinaison non linéaires entre les variables, dans lequel une séparation linéaire des individus sera possible.

Random Forest

Le Random Forest appartient aux méthodes ensemblistes qui au lieu d'avoir un estimateur censé tout faire, on en construit plusieurs de moindre qualité individuelle. Chaque estimateur a ainsi une vision parcellaire du problème et fait de son mieux pour le résoudre avec les données dont il dispose. Ensuite ces multiples estimateurs sont réunis pour fournir une vision globale. Dans le cas du Random Forest les estimateurs sont basés sur des arbres de décisions.

Naive bayes

Algorithme de classification qui se repose sur une hypothèse forte qui est l`indépendance des variables.

Apprentissage VS Validation

Lorsque notre objectif se tourne vers la prédiction, et en particulier à l'élaboration de modèles prédictifs, nous allons généralement utiliser nos modèles pour guider de nombreuses décisions et faire des millions de prédictions. Avec un modèle prédictif notre objectif principal n'est plus sur les données, mais sur un type de théorie. Ici, nous avons toutes les raisons d'être prudent, surtout que nous allons au-delà des limites des déclarations descriptives des données. Si nous développons des modèles prédictifs, nous devons avoir un moyen d'évaluer leur exactitude, leurs fiabilités et leurs crédibilités.

Pour cette raison je divise notre échantillon en deux. Le premier est utilisé pour l'apprentissage tandis que le deuxième est utilisé pour la validation. Pour des raisons de performance j'ai choisi de mettre le premier a 70% et le deuxième a 30%.

Prédiction des scores

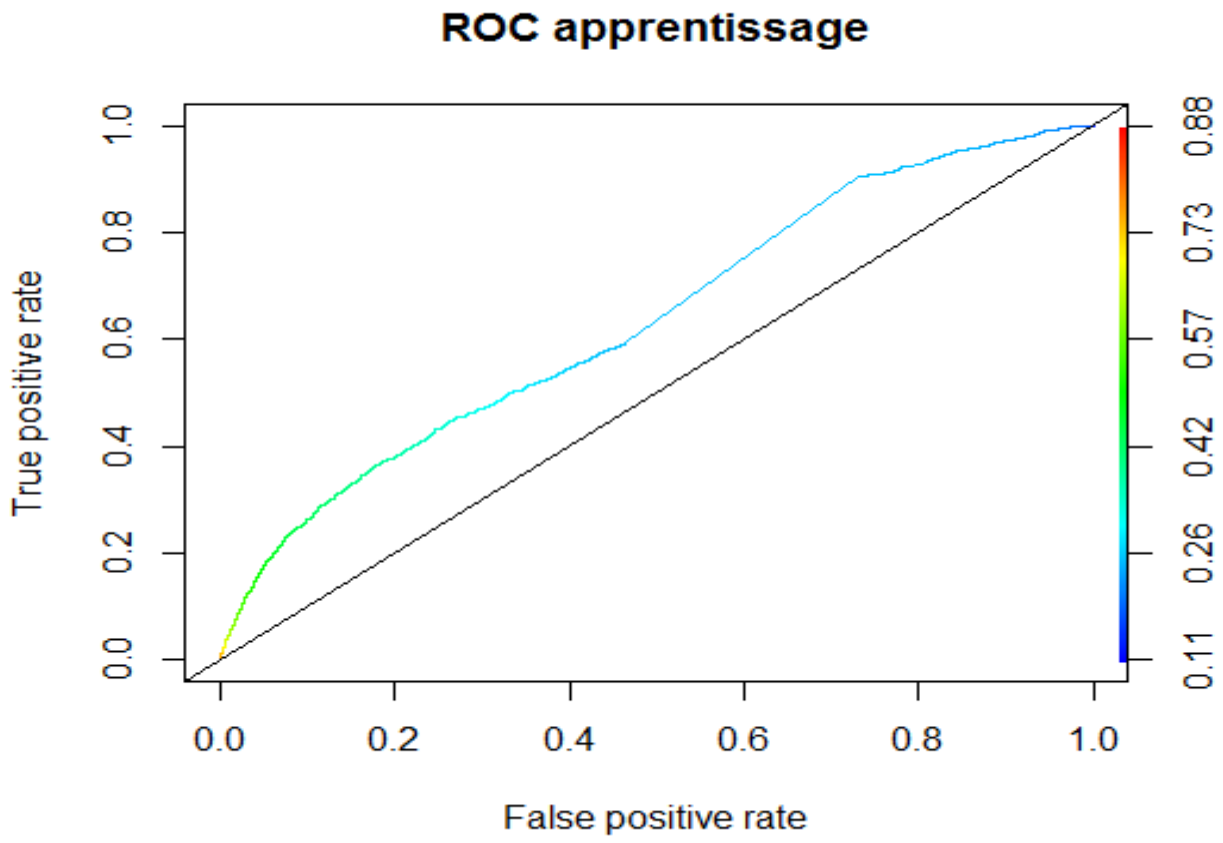
Dans cette partie je m`intéresse à la création des modelés spécifiques qui sont demandées par les clients en utilisant la méthode de Random Forest ou de SVM.

Le besoin du premier client tant de prédire les probabilités sur un ensembles d'individus qui ont atteint la partie configuration sur le site proposé par ce dernier, pour ce fait j'ai dû utiliser plusieurs modèles selon la campagne publicitaire afin de collecter les scores qui en résultent.

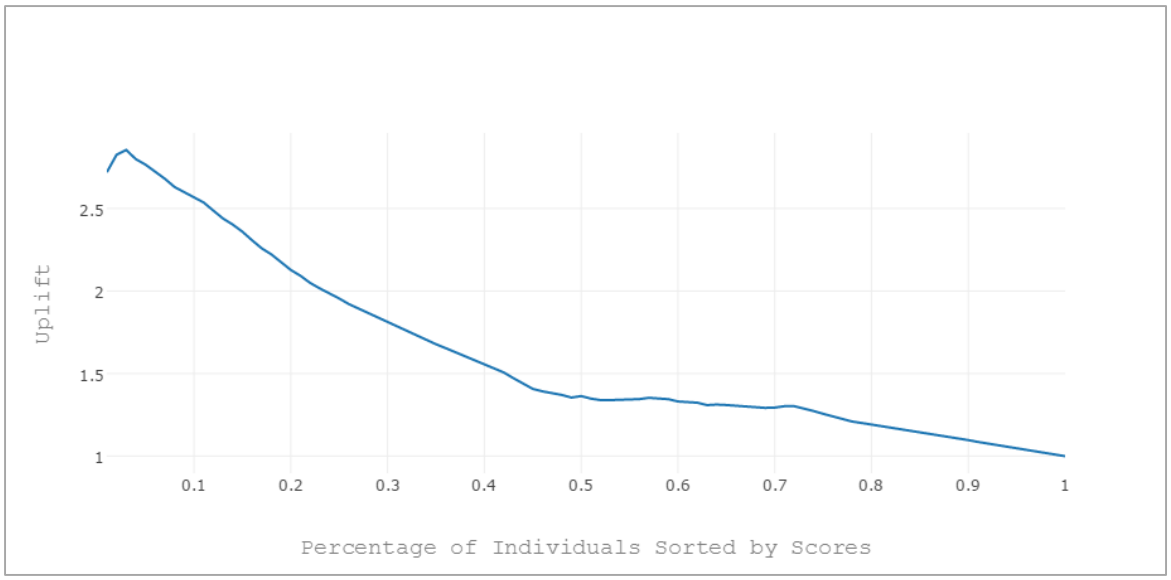
D`un autre côté, le deuxième client éprouve comme besoin de prédire plusieurs scores pour une seule compagne tels que le nombre de personnes qui ont accédés à la page d'accueil du site comparé à ceux qui ont accédé au site de l`entreprise concurrente.

Pour un score donné, je dois d'abord calculer de deux manières différentes le score puis appliquer le courbe roc et le courbe uplift afin de choisir par la suite le modèle le plus adéquat entre le Random Forest et le SVM.

Les figures ci-dessous illustre ces deux courbes



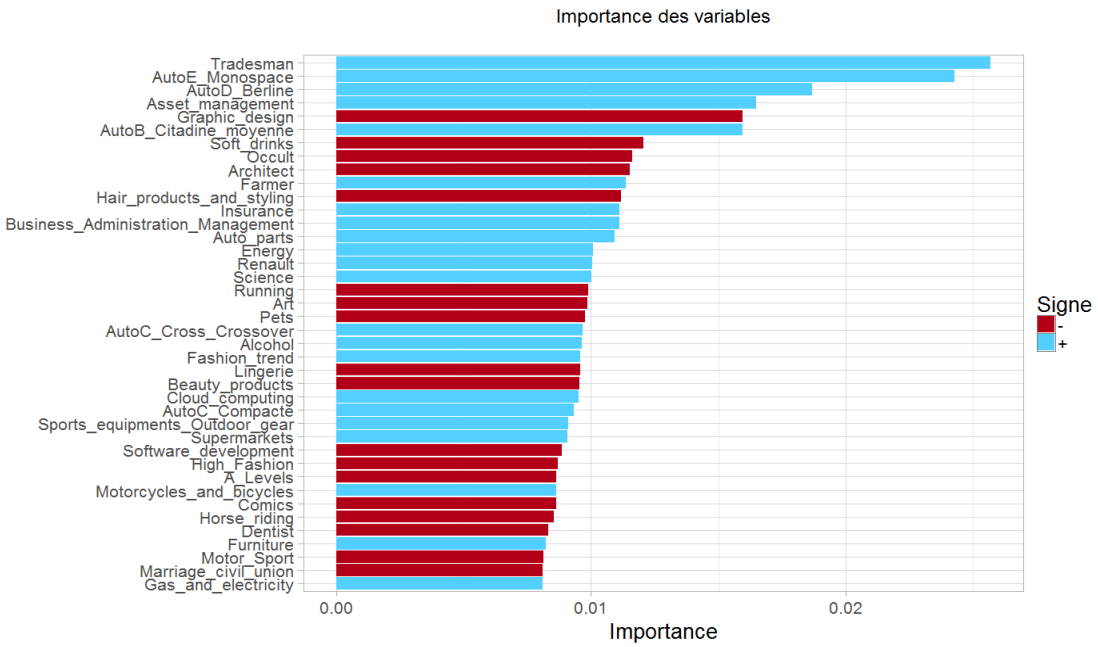
La courbe ROC (receiver operating characteristic) est une mesure de la performance d'un classificateur binaire, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs des caractéristiques de chacun de ces éléments. On la représente souvent sous la forme d'une courbe qui donne le taux de vrais positifs (fraction des positifs qui sont effectivement détectés) en fonction du taux de faux positifs (fraction des négatifs qui sont détectés incorrectement).



La courbe d'uplift mesure aussi la performance mais prend en considération la prédiction des probabilités. Elle représente les uplifts des nombres des individus cumulatifs qui ont convertis.

Importance des variables

La dernière étape de la modélisation s'intéresse au calcul de l'importance des variables en fonction de chaque score.



L'importance des variables est une méthode implémenté dans les librairies de data science de R et de python qui assure une permutation aléatoire des variables une par une et calcule son importance selon l'augmentation et la diminution de l'erreur.
L'effet positif et négatif de ces variables n'était pas donné par la méthode mais suite à des demandes clients j'ai imaginé un moyen d'y répondre grâce à des modèles linéaires en analysant les pentes des

droites obtenues. J'ai pensé à automatiser cette tâche en créant une fonction capable de reproduire toute la démarche et de donner les analyses qui en résultent.

Prédiction de quelque données 2nd party

Le client automobile a acheté des données 2nd party qui concernent les cartes grises digitalisées pour qu'ils soient exploitables dans l`étape de modélisation, mais malheureusement, le volume de ces données est négligeable par rapport aux autres données. Pour ce fait j`ai choisi de créer un modèle pour les prédire en fonction des données first party. Le modèle utilisé est naïve baise, ce choix est justifié par le besoin d`une exécution rapide.

Classification

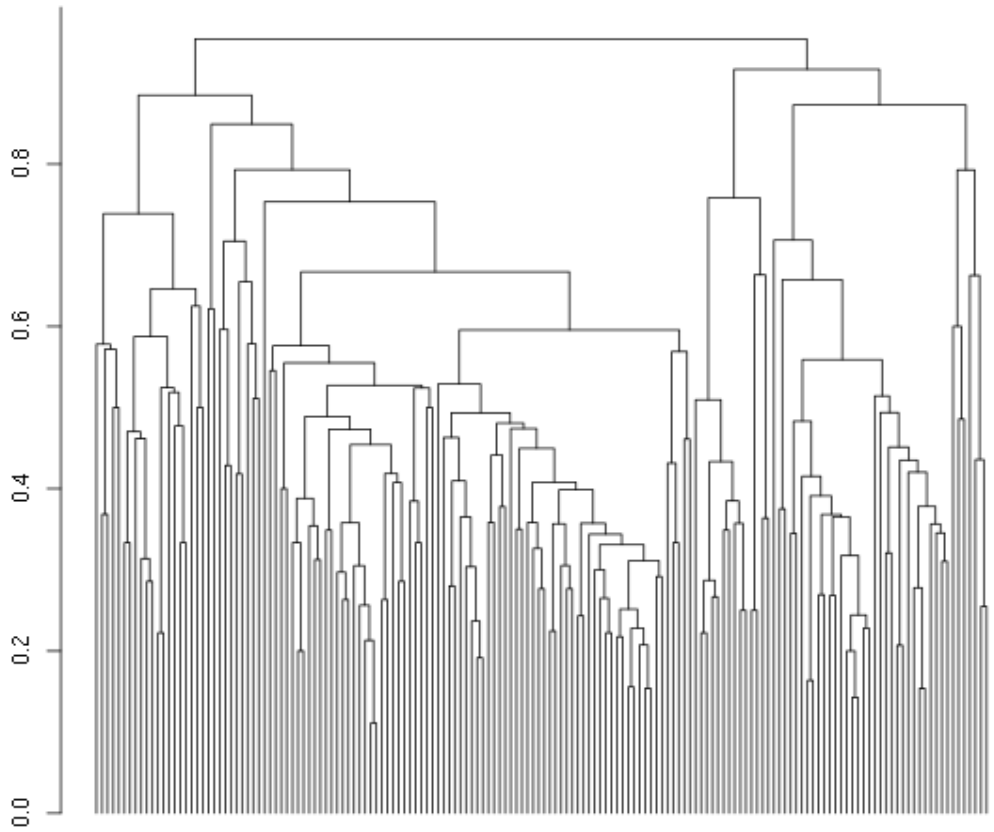
Analyse à composante principale

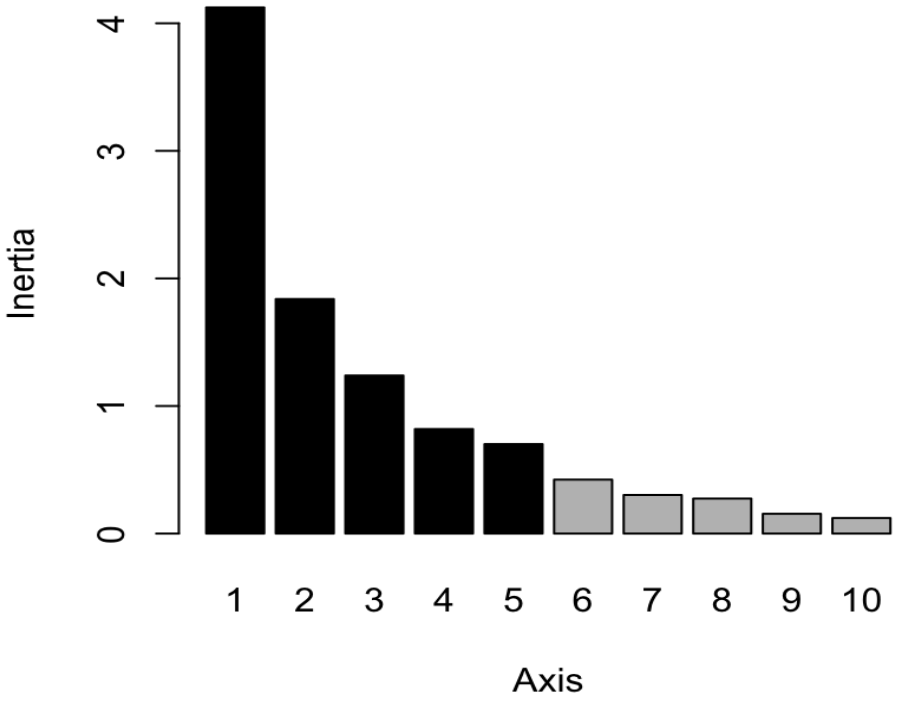
Tout d'abord, je commence par appliquer l'analyse à composante principale sur les données afin de pouvoir calculer les coordonnées des individus sur les axes principaux.

D'autre part, Elle permet d`analyser les variances cumulatives dont la valeur indique les nombres optimaux de vecteur propres qu'il faut utiliser.

Classification hiérarchique ascendante

A cette étape, s`introduit la classification hiérarchique ascendante des données qui s`intéresse à construire les classes par agglomération successives des objets deux à deux. Cet algorithme aboutie à la construction du dendrogramme qui figure ci-dessous et qui rassemble des individus de plus en plus dissemblables au fur et à mesure qu`on s`approche de la racine de l`arbre et qui va aider (en plus de l'histogramme d'inertie) par la suite à préciser le nombre de cluster optimale après l`avoir analyser.


0.0
0.2
0.4
0.6
0.8


Inertia
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
Axis

Calcule des centres de gravité et vecteurs propres

Les centres de gravité sont calculés pour chaque cluster Tandis que Les vecteurs propres proviennent de l'analyse à composante principale. Les deux sont sauvegardés, par la suite, en dur pour être utilisés ultérieurement dans la partie projection.

Description des clusters

La description des clusters se fait à travers la comparaison des Z-Value des variables, qui se calcule par la soustraction de la moyenne des individus dans un cluster, par la moyenne des individus totale.

Cette Description me permet de nommer chaque cluster. Durant le stage, j'ai obtenu plusieurs clusters tels que :

Un cluster qui correspond aux étudiants, un cluster qui correspond aux professionnels, un cluster qui correspond au retraitées et un cluster qui correspond aux gens qui aiment les loisirs…

Ci-dessous une capture d'écran qui illustre les Z-Value d'un cluster. J`en déduis que ce cluster convient aux individus ayant une expérience professionnelle.

Type		ZValue	AVGClust	AVG
Clusters	**Banking**	4.14861728	6.60617284	2.45755556
Clusters	**areers_and_occupational_trainin**	4.11174074	7.07407407	2.96233333
Clusters	**Law**	3.97644444	6.32222222	2.34577778
Clusters	**Loans**	3.68787654	4.57654321	0.88866667
KeyLife	**Real_estate**	3.62461728	6.03950617	2.41488889
Clusters	**Finance**	3.56361728	5.38395062	1.82033333
Clusters	**Asset_management**	3.36044444	4.51111111	1.15066667
Clusters	**iness_Administration_Managem**	3.34076543	5.15432099	1.81355556
Clusters	**Politics**	3.01606173	5.73950617	2.72344444
Clusters	**Public_administrations**	2.88088889	6.18888889	3.308
Clusters	**News**	2.83737037	6.37037037	3.533
KeyLife	**Insurance**	2.72135802	4.70246914	1.98111111
Clusters	**Tourism**	2.50518519	5.06296296	2.55777778
Clusters	**Entrepreneur**	2.50220988	3.29876543	0.79655556
Clusters	**Labour_law**	2.34462963	3.71851852	1.37388889
Clusters	**Lawyer**	2.33993827	3.67160494	1.33166667
Clusters	**Teaching**	2.27976543	4.43209877	2.15233333
Clusters	**Holiday_rentals**	2.0302963	3.11851852	1.08822222
Clusters	**Talent_FR**	1.96854321	3.18765432	1.21911111
Clusters	**Social_networks**	1.9382963	4.45185185	2.51355556
Clusters	**Building_and_civil_engineering**	1.93625926	3.92592593	1.98966667
Clusters	**Holidays**	1.86654321	5.88765432	4.02111111

Description des dimensions

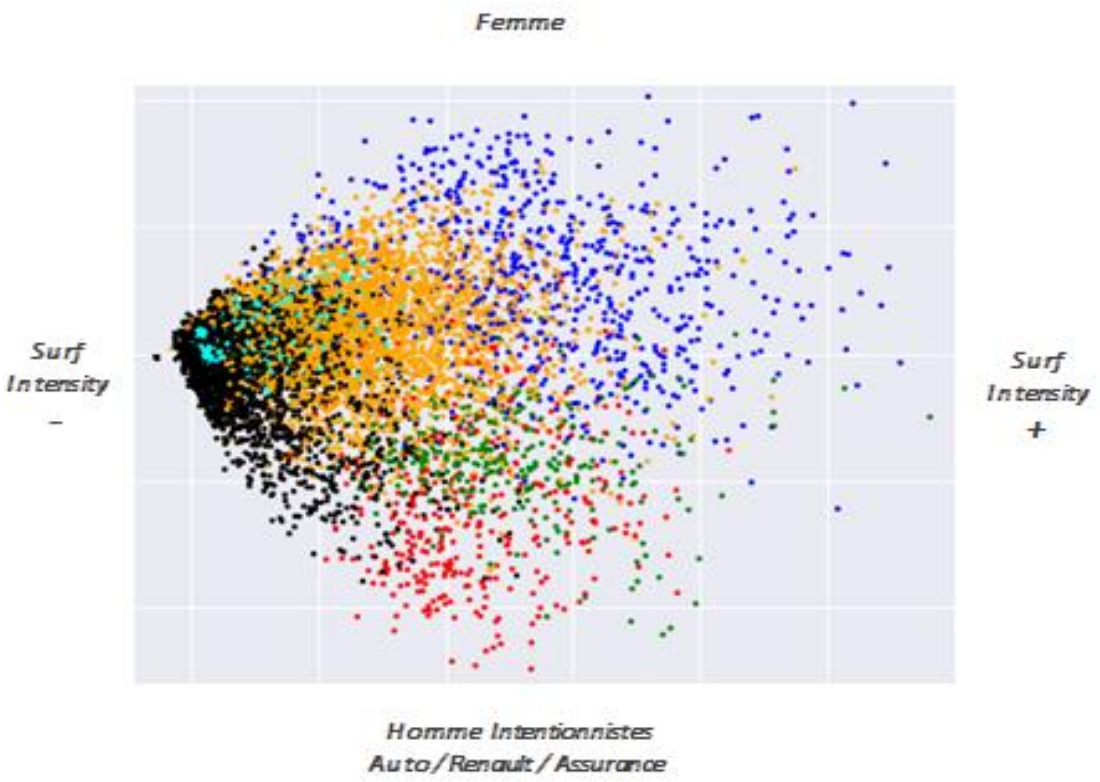
Pour réaliser la description, j`utilise la commande transform de l'ACP qui retourne les poids des variables correspondants à chaque axes, pour pouvoir ensuite nommer les axes du nuage d'individus.

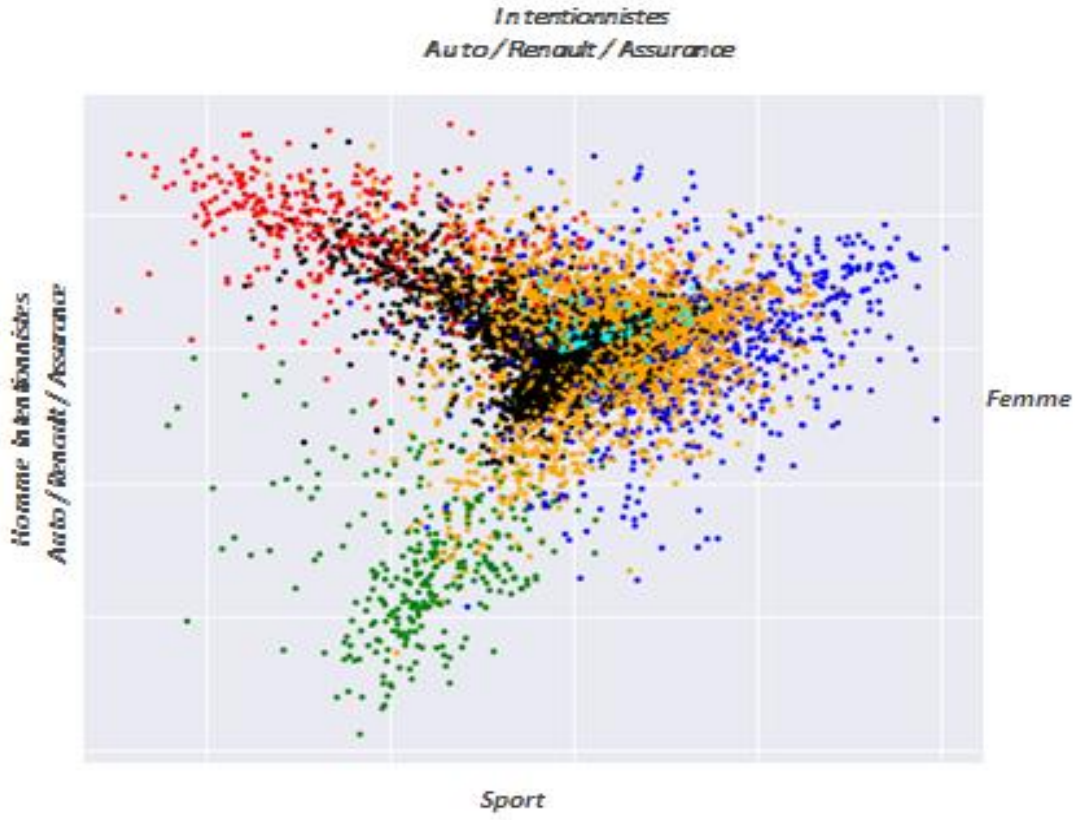
Le tableau suivant met en evidence une fraction de la description du deuxième et du troisième axe qui montre que l'extrémité du premier désigne les femmes et que l'extrémité du deuxième désigne les intentionnistes auto.

Variable	PC-2
female	24%
Family	14%
aged_18_to_24	14%
lower_middle_class	13%
Marriage_civil_union	13%
Pregnancy	13%
student	12%
Personal_care	12%
Healthcare_and_medicine	12%
Diet_and_nutrition	11%
Cooking	11%
Infants_children	11%

Variable	PC-3
aged_18_to_24	29%
student	28%
aged_25_to_34	26%
Car_buyers	10%
Car_brands	10%
AutoC_Compacte	10%
AutoC_Cross_Crossover	10%
AutoD_Berline	10%
[redacted]	10%
Local_car_brands	10%
AutoB_Citadine_moyenne	9%
AutoE_Monospace	9%

Les schémas suivant présente un nuage d'individu sur les axes (1,2) et les axes (2,3).

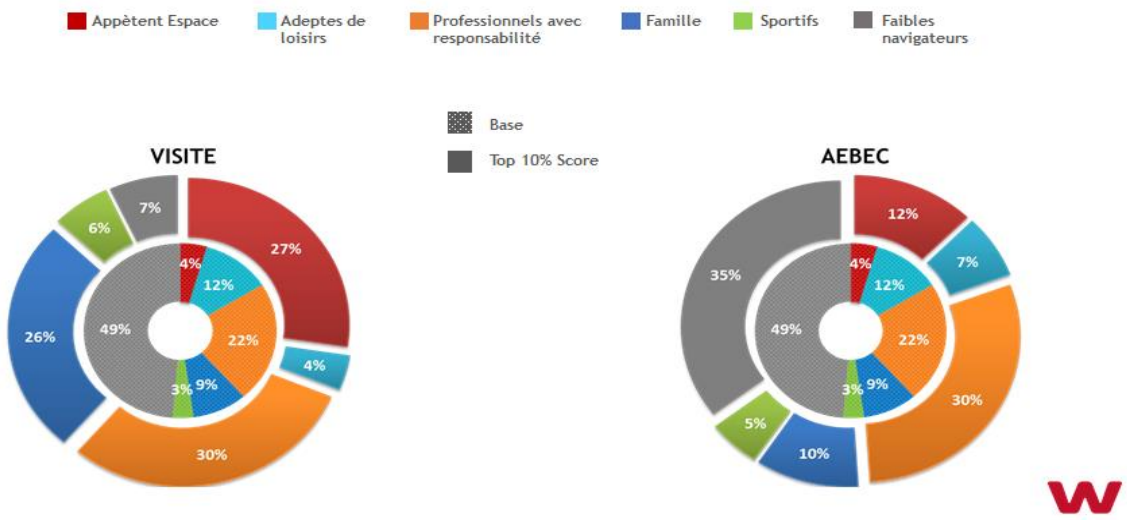




Ventilation des audiences

Le diagramme de ventilation a pour but de préciser le pourcentage de chaque audience pour les 10 pourcent qui ont les scores les plus élevés contre le pourcentage de chaque audience dans l'échantillon totale.

La figure ci-dessous illustre le diagramme de ventilation de deux scores différentes.



Analyse et mise en forme des données

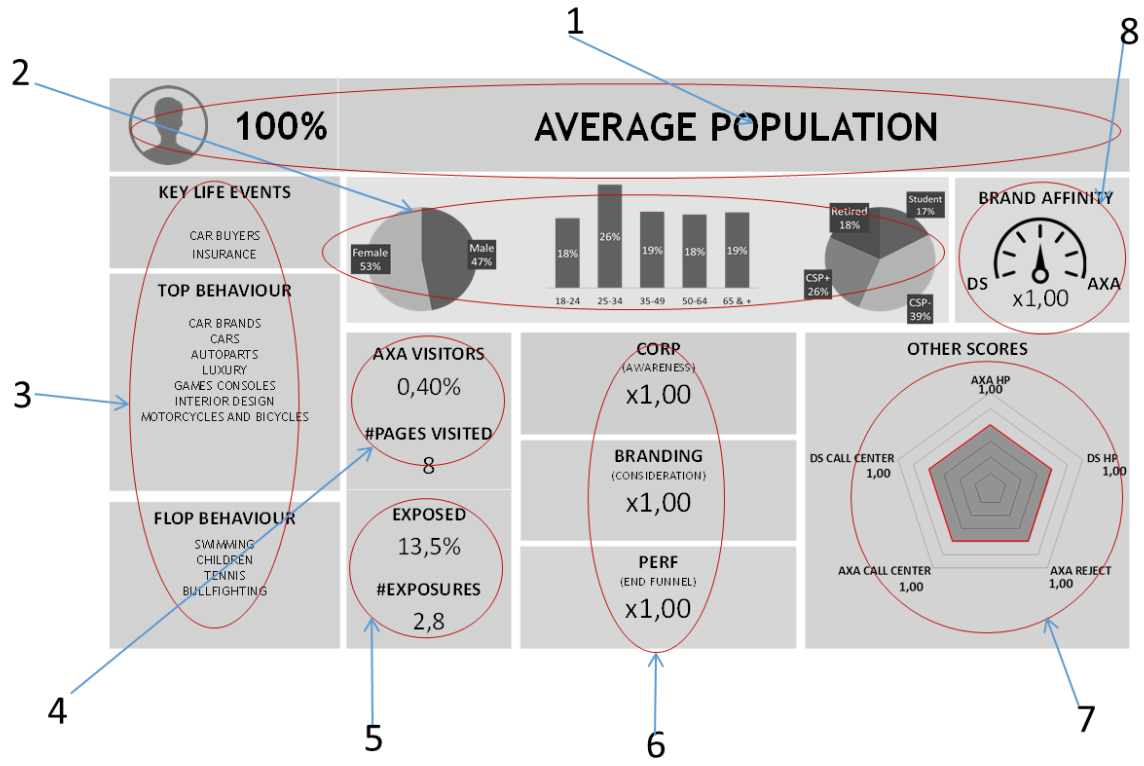
Après avoir effectué la démarche de modélisation, de classification, on passe à la partie analyse et mise en forme des données pour les mettre plus tard dans la diapo. Pour ce faire on utilise Microsoft Excel qui va aider à accomplir les tâches suivantes :

- Calcul du nombre de visiteurs, demandeurs, les nombres des exposés et les nombres des clics sur le site du client.
- Création des diagrammes correspondant aux données socio démographiques calculées précédemment sur R ou python.
- Calcul des taux de conversion et pression.
- Utilisation des données provenant de R ou python concernant les z-value pour mieux les visualiser et les analyser pour nommer les axes
- Utilisation des données provenant de R ou de python concernant la ventilation pour crée les diagrammes.

Présentation

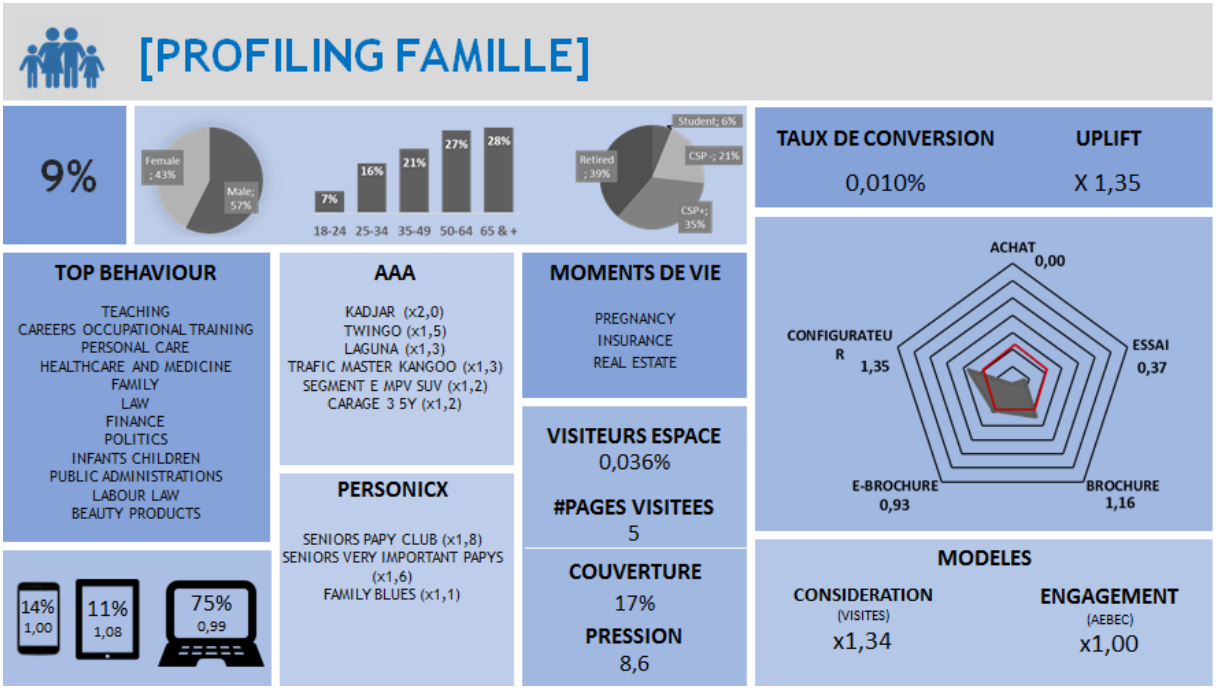
A cette étape on s'intéresse à vérifier l'avis des annonceurs par rapport à la présentation des clusters et cela est établit sur Power point.

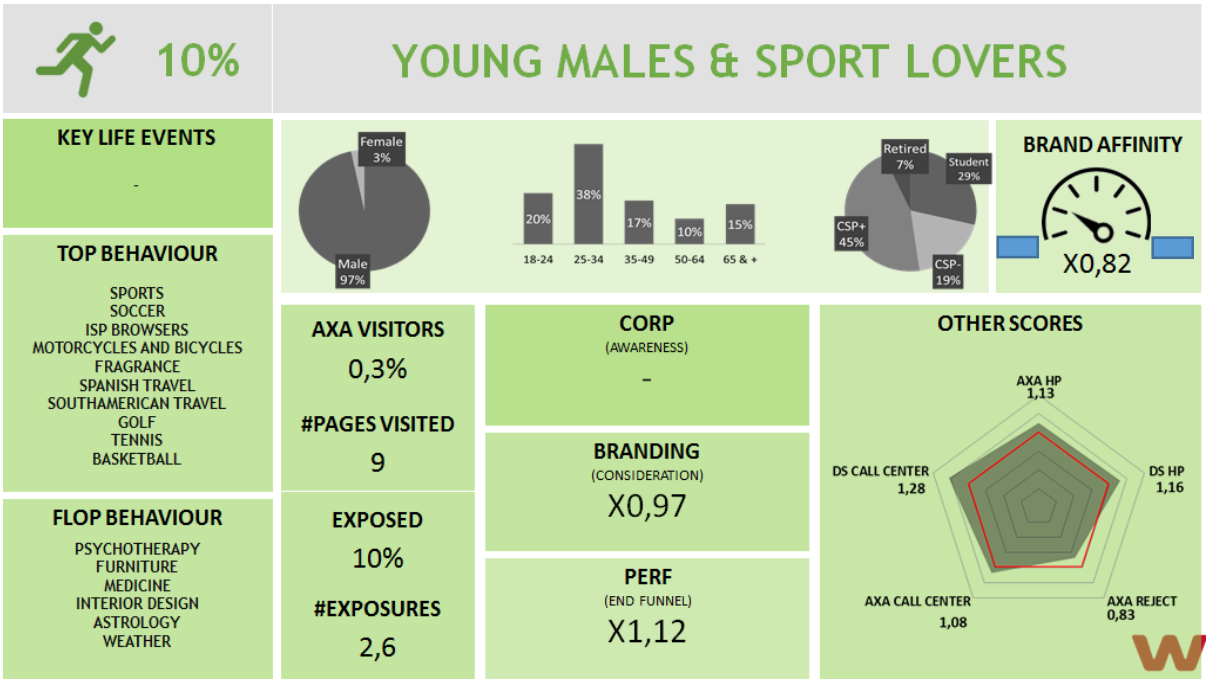
Tous les diapos consacrées aux audiences suivent plus ou moins le prototype suivant ;



- 1 : logo, nom et proportion de l'audience dans la base de données
- 2 : profil sociodémographique des utilisateurs
- 3 : comportements des utilisateurs, moments clé de leurs vies, centre d`intérêt et autres
- 4 : % des visiteurs dans une audience et nombre moyen des visites
- 5 : % des utilisateur exposes dans une audience et nombre moyen des exposes
- 6 : les uplift des scores
- 7 : diagramme d`araignée présente l`aptitude des utilisateurs à choisir entre les différents scores
- 8 : aptitude des utilisateurs à choisir entre les deux marques ou avoir un choix neutre

Et voici ci-dessous un exemple d'un cluster pour chacun des deux clients.





Projection et Activation

La dernière étape du projet consiste à projeter les modèles et la classification sur l'ensemble des données en utilisant le cloud puis les Appliquer pour prédire la probabilité de configuration de chaque modèle pour chaque individu de la base.

J'ai choisi de faire la projection de la classification en utilisant une méthode mathématique qui consiste à calculer les coordonner des individus sur les différentes dimensions. C'est-à-dire en multipliant les nouvelles données par les vecteurs propres de l'ACP.

Après je calcule les distances euclidiennes entre chaque point et les centre de gravité et je compare les distances obtenues de tels sorte que le point ayant la distance minimale le séparant avec les différents centres de gravite appartient au cluster de ce dernier.

Une fois que toutes les démarche sont mises en place, j'active le projet pour les clients en créant un code python sur le serveur qui contient les méthodes de projection. Ce code s'exécute une fois par jour et renvoie les résultats sur WCM ou le client peut directement consulter et analyser sa compagne.

Création d'une Library en Python

Pour optimiser le travail, j'ai créé une librarie sur python.

La librairie contient les classes et les fonctions suivante :

- **idToLabel**
  Fonction qui transforme les id des variables par leurs noms en utilisant une api crée par Weborama.

- **MultiColumnLabelEncoder**
  Class qui permet à travers sa méthode transform() de remplacer des modalités par des nombres (categorical variables) et de les retransformer en modalités à travers sa méthode inverse_transform()

- **Uplift_courbe**
  Class qui permet d'observer la courbe uplift sur la console ou sur le browser

- **Roc**
  Class qui permet de visualiser la courbe Roc avec sa méthode courbe() et de calculer l'auc avec sa méthode auc()

- **Var_imp**
  Class qui permet le calculer de l'importance des variables

- **Classification**
  Class qui permet de faire l'acp, visualiser le dendograme et de classifier les individus.

- **Cloud**
  Class qui permet la visualisation des nuages des points en prenant en considération les groupes (en 2d ou 3d)

- **Traintest**
  Création d'un échantillon d'apprentissages et d'un échantillon de validation.

- **dimDesc**
  Fonction qui permet la description des dimensions en fonction des variables.

- **clustDesc**
  Fonction qui permet la description d'un cluster en fonction des variables.

- **centre_grav**
  Fonction qui calcule les centres de gravite de chaque cluster.

- **projCluster_ParDistance**
  Fonction qui fait la projection des cluster en utilisant les vecteurs propres et les centres de gravite.

- **score_uplift**

Fonction qui calcule le score du model dans chaque cluster.

- **decription_socio**
  Fonction qui décrit les données socio démographiques.

- **technoDevices**
  Fonction qui calcule les proportions des devices utiliser dans chaque groupe.

- **conversion_taux_et_uplift**
  Fonction qui calcule les taux et les uplift des firsts partis.

- **visite_taux_et_moyenne**
  Fonction qui calcule taux des visites et les moyenne des pages visitée.

- **exp_taux_et_moyenne**
  Fonction qui calcule taux et les moyenne des expositions.

- **ventilation_topScore**
  Cette fonction fournie les informations nécessaires pour crée un diagramme de ventilation.

- **proportion_groupes**
  Fonction qui donne les proportions de chaque cluster.

Conclusion

Dans le cadre de mes études en deuxième année à l'Ecole Supérieure d'Informatique pour L'industrie et l'Entreprise, j'ai eu l'opportunité d'effectuer mon stage au sein de weborama où j'ai pu m'approfondir dans le monde de la data mining et acquérir de nouvelles connaissances. J'ai pu exploiter l`exactitude de la data science au profit du marketing et établir plusieurs modèles prédicteurs moyennant les différentes technologies et techniques choisies.

Ce travail m`a été très instructif pour plusieurs raisons. Techniquement, il m'a permis de développer mes compétences dans l`analyse des données et surtout de toucher à plusieurs aspects de cette discipline qui est en évolution continue. De plus, Il a été très enrichissant puisqu'il m'a donné l'occasion d'étudier et d'utiliser un ensemble de technologies tels que R et PHYTON. Hormis le côté technique et fonctionnel, ce projet a été une occasion pour découvrir le travail dans une hiérarchie professionnelle au sein d'une grande société et les difficultés inhérentes comme la répartition du temps et des efforts.

En participant à la réalisation de plusieurs projets, j`ai pu découvrir la rigueur nécessaire à la satisfaction du client ainsi que les pratiques requises à l'exerce du marketing digital.

Dans ce qui suit, j`envisage comme perspective d'améliorer les performances des modèles qui prédissent les donnes third party ainsi que la méthode de l`importance des variables afin de les rendre plus efficaces. Je souhaite enfin que ce modeste travail apporte progrès, évolution et satisfaction aux responsables de weborama, aux membres du Jury et aux personnes intéressées par le monde de data mining.

Annexes

http://www.weborama.com/fr/presentation/dn/

http://www.atinternet.com/solution/partenaires/weborama/

Data Science : Livre Fondamentaux et étude de cas écrit par Eric Biernat et Michel Lutz