MASTER DATA SCIENCE UNIVERSITÉ PARIS SACLAY - TÉLÉCOM PARISTECH

Rapport hebdomadaire de stage

Alexis Blanchet

17 mai 2018

1 INTRODUCTION

Les données des consommateurs de l'audiovisuel français sont stockées depuis déjà plusieurs années et leur utilisation a des fins commerciales ou sociologique est en pleine explosion et ce du fait de l'apparition de nouvelles techniques de machine learning. En particulier, les opérateurs réseaux tel que SFR ont accès au données d'utilisation des décodeurs TV. Cela leur permet en autre d'établir en temps réel les courbes d'audiences pour chaque chaîne pour leur abonnés. Dans ce contexte, une compréhension fine des flux entre chaîne est importante afin d'interpréter au mieux les différentes courbes déjà fourni par le logiciel en place.

Les Programmes TV sont fournis par un plurimédia aux opérateur pour utilisation. Cependant, ces programmes sont approximatifs car ne donnent pas avec précision l'heure de début réelle et l'heure de fin des différents programmes. De plus la Publicité n'est jamais répertoriée sur ces programmes. L'objectif du Stage est de réécrire, dans un premier temps a postériori et in fine en direct, de ces programmes TV afin d'assurer aux clients une explication claire des flux d'utilisateurs.

On s'appuiera pour cela sur les Programmes TV fournis par Plurmédia et sur les courbes d'audiences des utlisateurs de SFR.

La principale difficulté est l'absence complète d'historique. Aucune chaîne ne rend publique ses fichier AsRun qui correspondent a la description exacte de ce qui a été passé a la télévision. On est donc libre de créer soit même un historique et donc d'apprendre ce que l'on veux. Afin de faciliter les choses j'ai choisi de manière arbitraire de baser mon travail sur TF1 qui dispose d'un site internet de direct/replay accessible, de deux journaux très suivi ainsi que de plusieurs programmes de taille et de sujet variés et surtout d'une audience conséquente qui évite de confondre bruit et tendances. Après visionnage de plusieurs programmes TV, il est apparu que le retard est régulier pour certains programmes spécifiques, que la publicité est régie par de nombreuses lois et que les chaînes jouent avec les dites règles pour maximiser l'efficacité et l'impact des publicité.

1.1 LÉGISLATION

Le décret du 27 mars 1992 définit la publicité comme « toute forme de message télévisé diffusé contre rémunération ou autre contrepartie en vue soit de promouvoir la fourniture de biens ou services, y compris ceux qui sont présentés sous leur appellation générique, dans le cadre d'une activité commerciale, industrielle, artisanale ou de profession libérale, soit d'assurer la promotion commerciale d'une entreprise publique ou privée ». Une période d'au moins vingt minutes doit s'écouler entre deux interruptions successives à l'intérieur d'une même émission, et ce qu'il s'agisse d'une œuvre cinématographique, d'une œuvre audiovisuelle (constitutive, ou non, d'une série, d'un feuilleton ou d'un documentaire) ou d'un programme destiné à la jeunesse.

Les œuvres cinématographiques et audiovisuelles ne peuvent pas faire l'objet de plus de deux interruptions publicitaires. Et concernant les œuvres cinématographiques, elles doivent se limiter à une durée de six minutes au total.

Toutefois, la diffusion d'une œuvre audiovisuelle ou cinématographique par France Télévisions, et celle d'une oeuvre cinématographique par les services de télévision de cinéma ne

peut faire l'objet d'aucune interruption publicitaire. Il en va différemment lorsqu'une émission est composée de parties autonomes, et pour une émission sportive ou retransmettant un évènement ou un spectacle comprenant des intervalles. Dans ces cas, les messages publicitaires peuvent être insérés entre ces parties autonomes ou dans ces intervalles sans limitation du nombre d'interruptions. Sur les chaînes privées, le temps d'antenne consacré à la publicité est encadré différemment selon leur mode diffusion :

1. sur les chaînes diffusées par voie hertzienne terrestre (c'est-à-dire la TNT), il est limité à neuf minutes par heure en moyenne quotidienne sur l'ensemble des périodes de programmation au cours desquelles cette diffusion est autorisée, et à douze minutes pour une heure d'horloge donnée. Afin de favoriser leur essor, les nouvelles chaînes de la TNT bénéficient de règles allégées pendant un délai de sept ans à compter de la date du début des émissions, le temps consacré à la publicité étant seulement limité par le plafond de douze minutes par heure d'horloge donnée. A l'issue de ce délai, elles devront également respecter la durée de neuf minutes par heure en moyenne quotidienne.
2. Sur les chaînes distribuées par câble, par ADSL ou diffusées par satellite, la durée consacrée à la publicité est fixée par voie conventionnelle avec le CSA. Elle ne peut excéder douze minutes pour une heure d'horloge donnée

Les règles encadrant la durée des messages publicitaires sont plus strictes sur les chaînes publiques : cette durée ne peut dépasser six minutes par heure en moyenne quotidienne, ni huit minutes pour une heure d'horloge donnée.

De plus, depuis le 5 janvier 2009, les chaînes de France Télévisions (France 3 Régions exceptées) ne doivent plus diffuser de publicité de marques de 20 heures à 6 heures du matin. Cette interdiction ne s'applique qu'à la publicité et ne concerne donc pas les messages d'intérêt général, les publicités génériques (pour faire la promotion de la pomme, des produits laitiers, etc.) ou les parrainages, qui peuvent continuer à être diffusés

le CSA peut de plus autoriser le dépassement du nombre d'écrans publicitaire a l'occasion d'événements particuliers. Par exemple : "Le CSA a autorisé la société France Télévisions à insérer, au sein des émissions consacrées au Téléthon diffusées en soirée le vendredi 4 décembre 2015 sur France 3 et le samedi 5 décembre 2015 sur France 2, deux écrans publicitaires dont les recettes seront intégralement reversées à l'Association française contre les myopathies, dès lors qu'il s'agit d'une interruption exceptionnelle liée à une opération caritative."

1.2 Contournement et abus

Nous avons parlé précédemment des retards pris par les chaînes sur certains programmes. Cela fait parti des abus qui visent a maximiser les rentrées d'argent des différentes chaînes. Ainsi le télécrochet phare de TF1, pourtant annoncé pour 20h55, débute en réalité vers 21h07 ! Et ce n'est pas le seul exemple. La première chaîne, M6, TMC et surtout C8 détiennent la palme des chaînes de télévision dont les programmes de soirée commencent le plus tard. Bien loin des horaires annoncés. Même le service public s'est aligné.

Alors, pourquoi les chaînes retardent-elles le début des programmes de soirée ? D'abord pour une histoire d'argent. Selon Médiamétrie, c'est entre 21 heures et 21h15 que les télé-

spectateurs sont les plus nombreux devant le petit écran. Environ 25 millions de personnes regardent la télévision à ce moment précis de la journée. C'est donc l'occasion parfaite pour les chaînes privées de diffuser des écrans publicitaires qui seront vendus au prix fort aux annonceurs.

La législation impose pourtant des contraintes : une limite de 12 minutes de publicité maximum par «heure d'horloge» (20h-21h, 21h-22h...) et pas plus de six minutes maximum par plage de publicité. C'est justement avec cette réglementation que les chaînes tentent de placer le plus le maximum de publicité.

Prenons l'exemple de TF1. De 19h59 à 20h35, la première chaîne diffuse son journal télévisé. Dès la fin du générique, elle est donc libre de placer de la publicité à raison de 12 minutes et pas plus de six minutes par plage. Pour ne pas faire fuir les téléspectateurs, devant ces longs tunnels de pub, la chaîne va donc les entrecouper de programmes courts (Nos chers voisins, la météo, C'est Canteloup...).

Arrivé à 21 heures, heure du fameux pic d'audience, le compteur repasse à 0 et TF1 en profite pour rajouter quelques minutes de pub en plus. «C'est un très bon moyen d'optimiser ses écrans», fait remarquer Laurent Fonnet, consultant et ancien responsable de chaînes.

Mais ce n'est pas la seule explication. Touche pas à mon poste ! présenté par Cyril Hanouna sur C8 et Quotidien porté par Yann Barthès sur TMC finissent, de leur côté, de plus en plus tard, vers 21h10. Avec une incidence logique : la diffusion retardée du prime de ces deux chaînes vers 21h20.

Depuis l'année dernière, déjà, les téléspectateurs de C8 ont pris l'habitude de voir la chaîne de la TNT de Canal+ ne pas respecter le début de son programme du soir pourtant annoncé à 21 heures. Pour les autres chaînes, c'est donc un véritable casse-tête. «Ces deux talk-shows obligent les chaînes privées historiques à s'aligner», observe Laurent Fonnet.

Plus de quatre millions de téléspectateurs sont alors devant les deux émissions phares de la TNT. Quatre millions de personnes ainsi susceptibles de rater le début du prime time de TF1 et M6...

Les téléspectateurs peuvent bien s'agacer de la diffusion tardive de leur soirée télé. Et ils n'ont pas le choix. Le conseil supérieur de l'audiovisuel (CSA) ne peut agir puisque de nombreuses exceptions existent (émissions en direct, circonstances exceptionnelles liées à l'actualité...) et permettent aux chaînes de s'affranchir de leur obligation de respect des horaires.

Les magazines de télé, eux, se tiennent aux grilles communiquées par les chaînes. Mais alors pourquoi s'obstinent-elles à annoncer leurs programmes à 21 heures, ou 20h55 comme pour TF1, si elles choisissent au final de les faire débuter au-delà de 21h05 ? Pour ne pas décourager ses téléspectateurs et profiter des EPG, ces petits synthés qui affichent le programme en cours lorsque vous changez de chaîne. Une petite astuce qui permet de capter des téléspectateurs qui étaient alors en train de zapper en quête d'un programme susceptible de les intéresser.

On voit donc bien la nécessité d'obtenir un programme TV claire et véridique afin de pouvoir étudier divers paramètres et de tirer la maximum de plus-value des données des décodeurs TV.

2 MÉTHODOLOGIE

On se contente pour le moment de détecter les points (minutes) des fins de programmes/débuts de publicité. Ces points sont souvent liés a des chutes d'audience notables et on pourra ensuite d'affiner et de raffiner les données afin de permettre une réécriture encore plus précise des programmes TV.

Afin de détecter ses Points importants que l'on nommera Change Points (CP), nous allons tout d'abord constituer un historique de trois journées ou l'on notera exactement les débuts et les fins des programmes TV ainsi que des publicité. Ces trois journées sont le lundi 30 avril, le lundi 7 Mai et le mercredi 9 Mai pour TF1. Ces trois journées seront dans l'ordre la journée servant d'entraînement, la journée pour les cross-validation, et la journée Test pour noter le modèle.

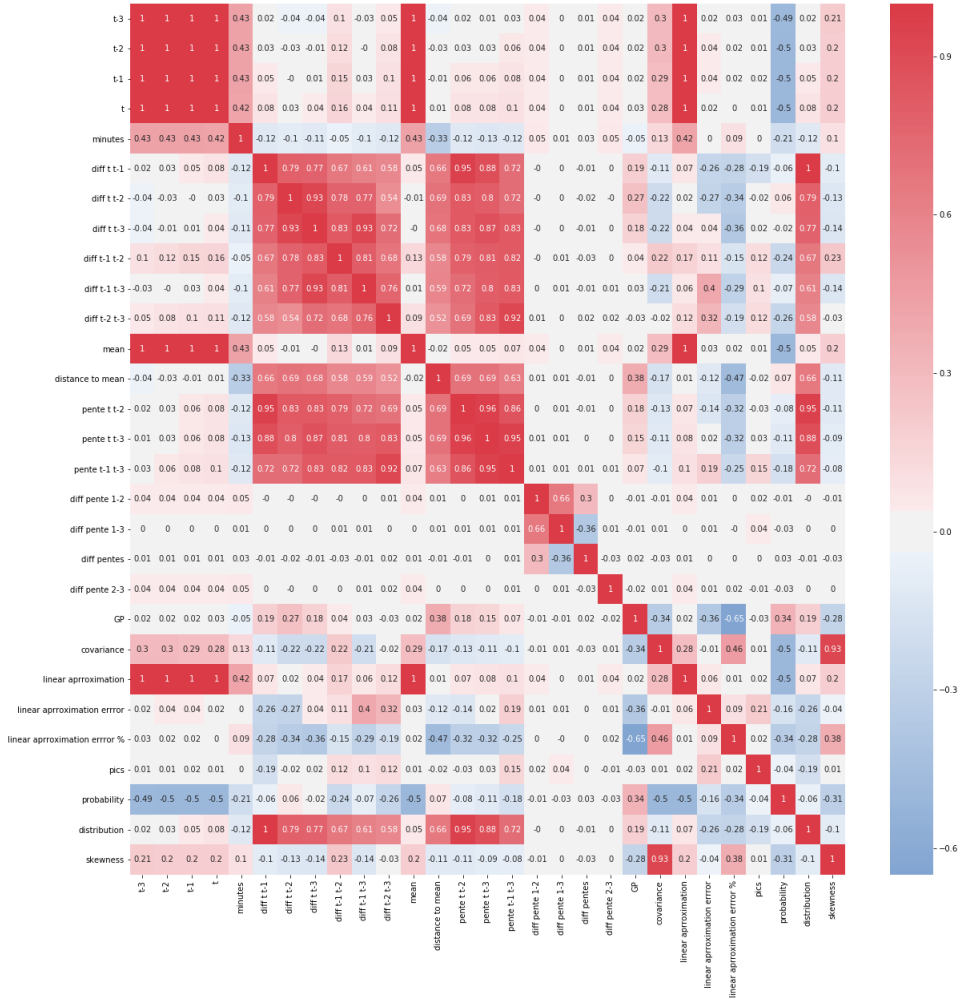
2.1 FEATURES

Il s'agira ensuite d'extraire différents features des Courbes d'audiences (Real Time Series). Afin de détecter les événements notables dans la journée, on utilise une fenêtre mouvante de 4 minutes (de t a t-3) afin d'utiliser le passé proche mais pas le futur(le modèle doit être conçu pour de l'utilisation en direct).Le choix de 4 minutes est arbitraire et pourra être sujet a révision. Les features utilisés sont dissociés en deux catégories distinctes :

1. Les features a l'intérieur de cette fenêtre : on prend par exemple les pentes, la distance, la distance a la moyenne, la pente relative, la distance relative, la probabilité des distances(*), la présence de pics...
2. Les features entre fenêtres : la covariance de chaque fenêtre,la moyenne de chaque fenêtre,la skewness et le kurtosis, la KFDR(Kernel For Density Ratio) (**)

(*) l'ensemble des distances entre points consécutifs d'une journée pour une chaîne permet de trouver statistiquement la distribution gaussienne de ces distances et donc de trouver la probabilité (on enlève les outliers pour trouver cette distribution)

(**)ces features sont sujets a de constantes évolutions en fonction des résultats trouvés et de l'avancée de mes recherches. ci dessous la matrice de corrélation des features ad hoc.



2.2 ÉVALUATION DES PERFORMANCES

Donnons pour commencer quelques définitions :

Soit (E) la classe des événements, (non E) et alors un point régulier auquel rien de notable n'est survenu. On a alors le tableau suivant :

	classé E	classé non E
vrai E	True Positive	False Negative
vrai non E	False Positive	True Negative

De ce tableau nous allons extraire plusieurs mesures qui seront importantes dans la suite des travaux :

Tout d'abord l'Accuracy qui le calcule le ratio de points bien classé sur le nombre de point total. Ici elle sera peu utile du fait de l'importance du déséquilibre entre les deux classes considéré dans notre cas.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

La sensibilité est plus intéressante car elle donne le nombre de points importants bien détectés par l'algorithme, mais ne considère pas le nombre de False Negative

$$Recall/sensitivity = \frac{TP}{TP + FN} \tag{2}$$

la precision permet de combler la défaillance de la mesure précédente

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

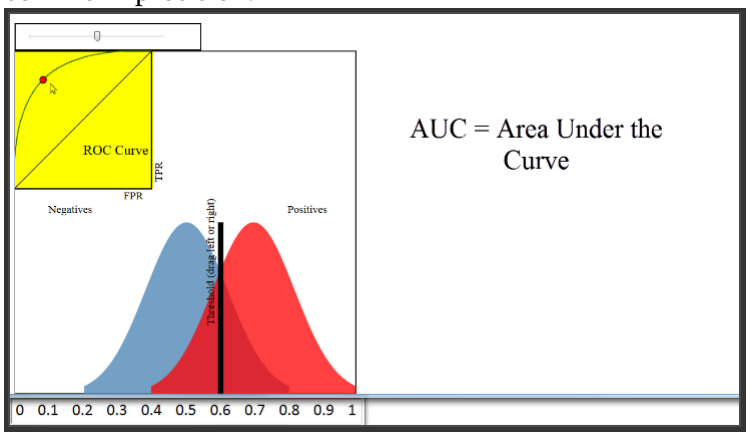
la moyenne permet d'avoir une idée générale de la bonne classification pour les deux classes et est utile car pénalise aussi bien la non détection que la détection abusive.

$$G - mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{FP + TN}} \tag{4}$$

Enfin la plus classique des mesures est la F-beta-measure qui est aussi appelée f-score est un moyen de combiner precision et sensibilité de manière a pénaliser tout comportement abusif, et ce en se centrant sur la classe que l'on veut détecter, au contraire de la G-mean. C'est cette mesure que l'on utilisera dans la suite.

$$F - beta - measure = \frac{(1 + \beta^2) * Recall * Precision}{\beta^2 * Recall + Precision} \tag{5}$$

Pour la suite, on définira une mesure approximative qui permet mieux de rendre compte de la temporalité de la série. En effet si l'algorithme se trompe d'un indice, et donc d'une minute en réalité, on considérera qu'il n'y a pas d'erreur. Cela permet d'avoir une meilleur compréhension des erreurs faites par l'algorithme et autorise une certaine imprécision temporelle qui ne gêne pas. On notera pour la suite mesure-2 la mesure autorisant les deux voisins directs du points comme imprécision.



2.3 MODÈLES

Afin de résoudre ce problème on utilisera trois groupes d'algorithmes réparti ainsi :

1. Tree Boosting : On utilisera 2 Catboost moyennés, 2 XGBoost moyennés et deux LightGBM moyennés. On prend a chaque fois deux fois le même algorithme pour réduire la variance de chacun de ces modèles et non pas a des fin ensembling.
2. Deep Learning : On utlisera ici un seul modèle de Réseau de Neurones fait avec Keras et TensorFlow. Il s'agit d'un réseau relativement simple constitué de deux couches cachées.
3. Autres : On utilise ici 4 Support Vector Classifier et 4 KNeighborsClassifier avec différents paramètres a des fins de test, et d'ensembling. Cette partie variera sans doute afin de permettre une réduction de la variance.

On entraîne tous ces modèles et on agrège ensuite les probabilité de prédictions sur le jour de cross validation pour entraîner, sur ce nouveau dataset, une Régression Logistique. On présente ici les résultats :

	Précision E	Recall	F2	Précision-2	Recall-2	F2-2
LightGBM	0.32	0.78	0.61	0.44	0.86	0.72
CatBoost	0.9	0.47	0.52	0.93	0.50	0.55
.SVC(kernel='linear')	0.77	0.47	0.51	0.88	0.51	0.56
SVC(kernel='rbf', gamma=0.1, C=1)	0.65	0.56	0.57	0.73	0.60	0.62
SVC(kernel='rbf', gamma=0.7, C=0.5)	0.77	0.47	0.51	0.80	0.54	0.54
SVC(kernel='poly', degree=3)	0.77	0.36	0.41	0.85	0.39	0.44
Neural Network	0.67	0.61	0.625	0.75	0.66	0.67
XGB	0.31	0.82	0.625	0.39	0.88	0.7
KNeighborsClassifier(5, weights='uniform')	0.68	0.59	0.61	0.78	0.66	0.68
KNeighborsClassifier(5, weights='distance')	0.48	0.64	0.60	0.63	0.77	0.74
KNeighborsClassifier(10, weights='uniform')	0.62	0.61	0.61	0.75	0.7	0.7
KNeighborsClassifier(10, weights='distance')	0.60	0.61	0.61	0.72	0.7	0.7
Logistic Regression (Stacking)(C=0.001)	0.52	0.71	0.66	0.65	0.82	0.78

On peut se demander ou l'erreur est localisée en terme d'« erreur temporelle » et on va donc découper la série sur différentes sections. En effet il est plus grave de ne pas détecter la publicité durant le prime time que durant les programmes de la nuit. On présente ici les résultats pour les tranches horaires déjà défini par l'outil de visualisation de SFR :

	Précision-2	Recall-2	F2-2
3h-27h	0.65	0.82	0.78
6h-13h	0.54	0.78	0.72
13h-20h	0.86	0.82	0.83
20h-27h	0.60	0.87	0.80
6h-24h	0.66	0.83	0.79
10h-13h	0.68	0.86	0.82
12h-15h	0.72	0.61	0.63
6h-11h	0.47	0.63	0.61
13h-16h	0.66	0.6	0.61
14h-18h	0.9	1	0.98
16h-19h	1	1	1
19h-22h	0.72	1	0.93
20h-23h	0.69	1	0.91
23h-27h	0.5	0.71	0.65

On trouvera le code pour l'obtention de tous ces résultats sur le dépôt git du Stage ainsi que divers graphes permettant la visualisation des données et des résultats importants.

L'ensemble du modèle et de ses composants est sauvegardé pour pouvoir prédire divers journées sans avoir a le ré-entraîner a chaque fois. Pour la suite on prédis les CP de chacune des trois journées et l'on va se servir de ces prédictions et de la réalité (historique annoté) pour mettre en place la nouvelle couche de décision de notre modèle.

!!! La sauvegarde et la chargement du modèle LGBM ne fonctionne pas pour la prédiction et ces valeurs sont donc soumises a une baisse de 0.02 pour chacune des mesures quand le LGBM est retiré (mesure temporaire en attente de résolution du problème) !!!

3 Plan pour le futur

Une fois que l'on a détecté des Change Points on va essayer de voir de quel type il s'agit : on a alors deux options qui se présentent.

1. Changement de Programme : Il s'agit donc de la fin d'un programme, avec ou sans publicité.
2. Plage publicitaire au milieux d'un Programme : Nous nous trouvons dans un programme et après une page publicitaire nous recommencerons le même programme

Cette distinction est importante, en effet nous ne traiterons pas les deux catégories de la même manière. L'une nécessite d'autre traitement pour parvenir a définitivement étiqueté le Change Point, pour l'autre il suffit de prédire la durée de la publicité pour finir le travail sur une période donnée.

Nous allons donc nous attaquer maintenant a la classification des événements détectés en trois catégories distinctes (et normalement équilibrées) qui permettrons de classer les événements détectés. Nous ne devons en effet pas oublier que le score de la classification n'est pas parfaite et qu'ainsi nous ne détectons pas tous les CP ni ne détectons uniquement les CP.

Labels Possibles :

- Dans un programme
- A la fin d'un Programme
- Erreur (Faux Change Point)

La dernière catégorie est importante et permettra de détecter a postériori des erreurs de classification. Cette étape et son degré de Précision/Recall entraînera potentiellement un changement dans les paramètres du modèles précédent afin de booster la précision ou le recall.

on a ici les premiers résultats présentés sous forme de ROC curve (la classe 0 est un CP erroné, 1 dans un programme et 2 a la fin d'un programme)