



► *E-Guide*

# Automatisation de l'activité analytique : les retours de Sephora et Captain Train.

## Dans cet E-guide :

Sephora et Captain Train ont beau être sur deux marchés diamétralement opposés, ces derniers ont augmenté leurs forecasts de ventes et les volumes de commandes avec l'intelligence artificielle. Découvrez les retours d'expérience de ces deux entreprises qui ont mis en place l'automatisation de leur activité analytique.

**Automatisation  
de l'activité  
analytique : les  
retours de  
Sephora et  
Captain Train.**

# Big Data : Sephora prédit les performances de sa e-boutique avec le Machine Learning

*Philippe Ducellier, Journaliste*

Sephora - la chaîne de magasins de cosmétiques d'origine française – connaît une croissance « à deux chiffres » et « une forte croissance des ventes en ligne » depuis qu'elle est passée dans le giron de LVMH. En 2016, elle a réalisé un CA de 1,5 Milliards d'euros rien qu'en France, avec un développement international et numérique soutenu (le numérique aide en plus à conquérir de nouveaux marchés), tendance confirmée en 2017.

Dans cette dynamique, le site e-commerce de Sephora devient de plus en plus central. Ce qui explique la volonté du groupe de se lancer dans un projet ambitieux à base d'AI (Machine Learning) et d'analytique prédictif.

## Contexte

Pour piloter l'activité de la e-boutique, les responsables du site doivent réaliser régulièrement des forecasts de ventes, de chiffres d'affaires et de volume de commandes. Ces estimations prévisionnelles alimentent en données différents services : le marketing, la supply chain (chaîne logistique) et le service clients.

Concrètement, le marketing utilise ces informations pour calibrer ses offres promotionnelles et ses opérations d'animation, pour définir son plan marketing, et pour le valider. La supply chain, elle, utilise le volume de commandes pour mieux anticiper son organisation (dimensionner son équipe en entrepôt, par exemple). Le but étant de respecter les délais de livraisons au client. Quant au service client, il extrapole des volumes d'appels au call-center pour s'adapter de manière pro-active.

« Jusqu'à présent, ces estimations étaient réalisées manuellement », explique Clément Marchal, Manager Data Science chez Sephora. Le travail représentait une charge importante. Sephora a donc décidé de lancer un projet pour automatiser et pour améliorer ses forcecasts avec pour but ultime de huiler sa chaîne de la donnée et de prévisions.

Mais pour y arriver, le Manager Data Science prévient qu'il faut relever trois défis majeurs : celui de la « data », celui du Machine Learning » et celui du process.

## Les données : le cœur de l'analytique

Chez Sephora, il a d'abord fallu identifier sur quelles données s'appuyer pour prédire des indicateurs clefs, et savoir comment collecter cette donnée.

« Ce projet transverse a nécessité la collaboration de nombreuses équipes », témoigne Clément Marchal, « les équipes e-store, web, web analytics, de l'IT (pour le Data Lake), et de Data Science ». Sans oublier deux prestataires externes que Sephora a choisis pour l'épauler dans ce chantier : le cabinet de conseils Avisia et Dataiku (« sa plateforme Dataiku DSS nous a permis de développer des algorithmes et de les industrialiser »).

Les prévisions de Sephora s'appuient sur deux sources principales de données : le plan marketing quotidien (qui liste les offres, événements, animations) et les informations de

navigations sur le site (visites, transactions, nombre de commandes, CA, type de livraisons utilisés). Mais pas les prix. Le but étant d'aller vite et d'arriver à un résultat pour prouver le bénéfice de la démarche, l'équipe de travail n'a pas voulu multiplier les variables.

Les deux sources alimentent un Data Lake, mais avec une différence. « Les données du plan marketing sont intégrées dans le Data Lake à la demande des métiers, pour qu'ils puissent réaliser des prévisions sur leurs dernières offres par exemple. [Mais] les données de navigation sont versées automatiquement, à J+1 ».

Sur ce point, le Manager Data Science de Sephora a tiré un premier enseignement de ce projet, qu'il donne sous forme de conseils pour qui souhaite se lancer dans une aventure analytique similaire. Il ne faut pas sous-estimer la volumétrie des données de navigation. « Elles sont très denses, elles représentent 250 champs. On parle de 2 à 5 To de données par mois en ce qui nous concerne ».

## **Machine Learning : choisir les algorithmes n'est pas trivial**

On a tendance à le gommer, mais le Machine Learning n'est pas une technologie monolithique. La discipline regroupe au contraire une multitude d'algorithmes. Pour réaliser de bonnes prévisions, il faut donc choisir le bon algorithme.

Sephora a commencé sa sélection en se concentrant sur deux indicateurs : le CA quotidien et le volume de commandes (là encore, le choix de quelques indicateurs clefs traduisent la volonté de faire « vite et bien »). Ces deux prévisions sont réalisées pour une durée de plusieurs semaines, avec un calendrier d'estimations jour par jour.

Comment Sephora a-t-il concrètement sélectionné ses algorithmes ? « Ils devaient répondre à trois critères », répond Clément Marchal. « Ils devaient pouvoir intégrer les tendances de l'activité à court moyen et long termes (à la semaine, au mois et à l'année) ; ils devaient pouvoir ré-apprendre des dernières tendances (notre activité sur le site est volatile), et ils devaient pouvoir intégrer du contexte marketing (le plan marketing ont un impact significatif sur l'activité) ».

Plusieurs algorithmes ont au final été sélectionnés par Sephora dont « random forest ».

## **Procédure : l'analytique n'est rien s'il n'est pas utilisé par tous**

Troisième défi du projet : la question des procédures. Car avoir des outils analytiques est une bonne chose, mais que les métiers les utilisent est beaucoup mieux. « Il fallait des prévisions sans complexités, qui soient agiles et efficaces », confirme le responsable du projet.

**Automatisation  
de l'activité  
analytique : les  
retours de  
Sephora et  
Captain Train.**

La première partie du process mis en place chez Sephora est aujourd'hui le suivant. Le marketing dépose son nouveau plan sur le Data Lake. Via un simple bouton « run », le Data Lake se connecte à la plateforme de Dataiku et lance les modèles analytiques. La réception des prévisions se fait par mail sous forme de fichier Excel, avec une ligne par jour, sur toute la période de la prévision.

« Ce process simple permet en plus au marketing d'itérer et de tester plusieurs combinaison d'offres ou plusieurs combinaisons d'événements à venir », se réjouit Clément Marchal.

La deuxième partie du process débute une fois que le plan marketing est validé. Cette fois, ce sont les métiers (e-Commerce) qui vérifient que les prévisions sont cohérentes. Si tout va bien, les plans et les prévisions sont ensuite communiqués à la logistique et au service client, et confirmée au service marketing.

« Cette procédure a trois avantages majeurs : les équipes marketing sont autonomes, l'étape d'expertise métier permet d'éviter le coté boîte noire de certains algorithmes, et le métier garde la main sur les prévisions », vante Clément Marchal.

**Nous avons une chaîne de traitement de bout en bout, avec des prévisions automatiques qui restent à la main des métiers, avec la possibilité d'itérer, de les valider, de les contrôler et de les diffuser**

**Clément Marchal,  
Manager Data Science,  
Sephora**

## Résultats : un gain de 15 points sur les prévisions après trois itérations

En plus des itérations pour améliorer les actions marketing, le projet a permis d'affiner les prévisions – faites auparavant sans Machine Learning – d'environ 15 points entre les forecasts et le CA effectivement réalisé.

Pour la phase d'apprentissage de l'algorithme, le responsable ne « veut pas dévoiler de secrets ». Il avoue néanmoins que « nous avons appris sur une période de deux années » et sur des données consolidées par les métiers pour éviter les anomalies de reporting (logs ou CA incomplets).

**Deux années entières de données pour l'apprentissage de l'algorithme**

Plus précisément, le modèle a scanné les résultats et les plans marketings des deux dernières années. Puis la validation « a été réalisée avec une base de test sur laquelle on a lancé des prévisions puis nous avons regardé le gap entre le résultat prédit et le CA réel », explique Clément Marchal.

Son équipe a fonctionné par Stream de trois à 3 à 4 semaines sur les étapes d'exploration de la donnée, de data préparation, et de modélisation. « Cela nous a permis de faire les premières prévisions et les premiers modèles rapidement ».

Ensuite, Sephora a itéré sur ces étapes pour affiner l'outil. « Il nous a fallu environ trois itérations pour arriver à des prévisions de qualité », évalue le responsable.

**Automatisation de l'activité analytique : les retours de Sephora et Captain Train.**



Face à la pertinence des prévisions, la logistique et le service client de Sephora – conquis d'après Clément Marchal - s'appuieraient désormais sur ces évaluations pour gérer leurs effectifs respectifs. Quant à l'automatisation, elle aurait permis au service marketing de se concentrer sur des actions à fort potentiel stratégique.

***Témoignage recueilli lors du salon Big Data Paris 2018***

**Automatisation  
de l'activité  
analytique : les  
retours de  
Sephora et  
Captain Train.**

# TrainLine Europe reste fidèle à Dataiku

*Alain Clapaud, Journaliste*

Captain Train est née en 2009, au moment où la SNCF a donné accès à son moteur d'achat de billets à des tiers. La startup est entrée en concurrence frontale avec Voyages-SNCF et a peu à peu étendu ses ventes aux services de transports par train ou par bus de 22 pays européens. Sa réussite a poussé son rival britannique TrainLine à prendre le contrôle du français en 2016 sur une valorisation de 200 millions d'euros. Captain Train représente désormais la marque britannique sur le continent européen et c'est bien la plateforme technique créée par Captain Train qui motorise aujourd'hui le service TrainLine Europe.

## Captain Train, un service 100% Ruby

Techniquement, la réussite du français s'appuie sur un service ultra-efficace à l'interface très simplifiée. L'ensemble des développements ont été réalisés en interne sur la plateforme Ruby on Rails. Le fonctionnement du site TrainLine Europe met en œuvre des modules développés dans ce langage, sur le moteur de messagerie interapplicative RabbitMQ pour les échanges et sur la base de données PostgreSQL.

La particularité du service, c'est que TrainLine ne dispose pas de la base de l'ensemble des lignes de ses partenaires transporteurs ferroviaires. Lorsqu'un utilisateur du service effectue une requête pour trouver un billet de train sur le réseau de la SNCF, de la Deutsche Bahn ou de Renfe en Espagne, sa demande est transmise aux systèmes de billetterie

destransporteurs qui renvoient une réponse à TrainLine. Ce dernier n'a donc pas à stocker sur ses serveurs l'intégralité des informations des transporteurs européens.

« Techniquement nous n'avons pas véritablement besoin d'avoir recours aux technologies que l'on a baptisé Big Data pour notre cœur d'activité » explique ainsi Cédric Raud, Développeur Full Stack chez TrainLine. En fait , l'entreprise a « lancé une initiative sur la Data voici deux ans maintenant pour [son] département marketing afin d'avoir une vision sur le comportement des utilisateurs et pour pouvoir analyser l'offre ferroviaire ».

## Un besoin Big Data issu du marketing

Si TrainLine Europe arrive à supporter la charge avec des solutions "classiques", le département marketing, lui, a poussé la startup à s'intéresser aux technologies Big Data pour analyser les sources de trafic du site et l'efficacité des opérations promotionnelles.

A l'époque du choix de Dataiku, Captain Train n'avait qu'une petite structure avec 2 développeurs et un Data Scientist. « Les personnes qui travaillaient sur le SEO utilisaient Google Analytics, celle qui mettaient en œuvre les bannières utilisaient Google Adwords et DoubleClick, celles qui s'occupaient de l'affiliation utilisaient la plateforme Tradedoubler. Or nous avons un besoin d'avoir une vision sur nos coûts d'acquisition, sur nos paniers moyens. Toutes ces analyses étaient réalisées manuellement », des extractions à la mise en forme, en passant par le partage des résultats.

La tâche « occupait totalement notre Data Scientist ». L'expert de l'analyse statistique consacrait donc le plus claire de son temps à d'autres choses que celles « les plus génératrices de valeur ».

Les fondateurs de Captain Train ont alors cherché une solution. « L'objectif était que le Data Scientist puisse traiter ses données de manière totalement autonome, sans avoir recours aux développeurs. L'atout de Dataiku est de gérer le volet extraction, traitement et analyse de données mais aussi tout le volet préparation et analyse, directement dans l'outil. Tout est intégré, facile d'accès, là où les solutions Big Data plus classique nécessite des connaissances sur les outils et des compétences que nous n'avions pas en interne à l'époque ».

La solution Dataiku a été préférée à d'autres, comme celles éditées par Talend ou encore Omniscope de Visokio (dont les tarifs étaient peu adaptés aux moyens d'une startup).

« En plus du prix, l'avantage de Dataiku était aussi dans la modernité de sa solution avec la capacité d'héberger en interne une solution bâtie sur des technologies Web, ou la présence d'une brique prédictive intégrées à l'outil. C'est un choix que nous ne regrettons pas aujourd'hui car Dataiku continue à répondre à nos besoins ».

## **Les premiers usages de Dataiku chez Captain Train**

Le premier projet à mettre en œuvre Dataiku a porté sur la performance des différents canaux mis en œuvre par l'équipe marketing pour conquérir de nouveaux clients.

Les pipelines d'extraction de données ont été modélisés sous Dataiku et, lorsque ceux-ci ont été jugés satisfaisants, ils ont été automatisés. C'est à dire qu'ils sont lancés chaque nuit afin d'alimenter une base de données dédiée aux analyses. « Nous avons développé en interne une brique logicielle qui nous permet de visualiser ces reporting. Cette brique permet à chacun de créer son propre rapport et donc de permettre aux équipes d'explorer cette donnée et d'identifier celles qui ont le plus de sens dans leur métier ».

Le second usage de Dataiku a été d'analyser le comportement des utilisateurs du service. Un moyen pour la startup de valoriser cette compréhension auprès des transporteurs et de faciliter ses relations commerciales avec ses partenaires.

La plateforme est alimentée par les données de production via un connecteur développé par Captain Train en Python. « Autre point très intéressant pour nous, Dataiku fonctionne aussi en mode collaboratif. Cela signifie que plusieurs personnes peuvent travailler sur l'outil, avec cette notion de comptes utilisateurs. On sait que tel Data Scientist est intervenu sur une partie d'un pipeline, qu'un développeur est intervenu sur une autre partie. C'est une bonne pratique, intéressante et très utile, pour assurer la maintenance des chaînes de traitements ».

## Cohabitation Franco-britannique dans le SI

Maintenant intégré à TrainLine, le système d'information de Captain Train cohabite avec celui créé par son alter-ego britannique - dont les choix technologiques sont assez différents.

Alors que le français privilégie la maîtrise interne de son infrastructure, l'anglais a clairement misé sur le Cloud (Amazon Web Services) pour développer ses activités. TrainLine UK utilise en effet les instances Amazon EC2, une base Oracle sur Amazon RDS. Pour le volet Big Data, ses Data Scientists mettent œuvre Amazon Kinesis pour l'intégration des flux de données et des clusters Hadoop avec Amazon EMR.

La convergence des deux systèmes d'information n'est pas à l'ordre du jour. Mais Cédric Raud précise que Dataiku pourra – s'il le souhaite - se connecter à cette infrastructure AWS sans difficulté.