



© PEXELS

Améliorer l'information temps réel

SNCF met au point des algorithmes de *big data* capables de prédire en temps réel l'affluence à bord d'un train, et l'heure prévisionnelle de ses futurs passages en gare en cas de situation perturbée. Explications avec Maguelonne Chandesris qui dirige l'équipe Innovation & Recherche «Statistique, Économétrie et Datamining» de SNCF.



MAGUELONNE CHANDESRIS

*Des futurs
du passé...
au futur
du présent*

Quel est l'objectif principal de ce projet ?

M. C. Baptisé RELUANCE (qui contient «retards et affluence en avance»), ce projet vise deux grands objectifs : prédire en temps réel le niveau de confort à bord (par exemple si on trouvera une place assise) et à quelle heure est prévue l'arrivée d'un train dans telle ou telle gare en cas de situation perturbée... Si le second point intéresse l'ensemble des 5 millions de voyageurs qui empruntent nos trains chaque jour, le premier concerne ceux qui prennent des trains régionaux, sans réservation. Pour y parvenir, nous avons développé des algorithmes de *machine learning* déjà testés sur plus de 3000 trains en circulation. Chaque jour, ils s'alimentent de centaines de milliers de données, à partir desquelles ils fournissent des millions de prédictions, au fur et à mesure de la remontée en temps réel des informations !

Sur quels types de données se basent-ils ?

M. C. Dans la pratique, nos algorithmes utilisent deux grands types de données. Tout d'abord celles qui concernent la localisation précise des trains en temps réel, collectée par des capteurs placés au sol et à bord des trains. Mais aussi des données sur le nombre de passagers qui montent et descendent des trains à chaque gare, obtenues par des capteurs installés à bord, au niveau des portes. Toutes ces données sont acheminées en temps réel, *via* des interfaces de programmation (API), vers nos algorithmes de calcul, hébergés sur nos serveurs situés près de la gare de Lyon.

Comment fonctionnent ces algorithmes ?

M. C. Leur principe général est facilement compréhensible. Très concrètement, ils vont chercher dans l'historique de ce type de données les situations les plus semblables à la situation actuelle. Une fois celles-ci identifiées, ils s'appuient sur ce qui est advenu après ces situations semblables passées (les «futurs du passé»)... pour prédire ce qui va se passer maintenant (le «futur du présent»). Point important : ces algorithmes dits «d'apprentissage non-paramétrique» réalisent leur phase d'apprentissage en continu, sans aucun reparamétrage à effectuer, facilitant la maintenance industrielle. Au final, ils améliorent sensiblement les prédictions de l'heure d'arrivée des trains à destination, et une marge d'erreur sur l'affluence inférieure à vingt voyageurs... pour des trains pouvant transporter plus de neuf cents personnes.

Quelles sont les prochaines étapes ?

M. C. Forts de ces résultats, nous avons déjà démarré les travaux d'industrialisation. Il nous faut notamment implémenter ces algorithmes de manière robuste dans «l'immeuble numérique» du groupe SNCF, afin que ces prédictions puissent alimenter tous nos systèmes d'analyse et de diffusion : affichages en gare, applications clients, centres opérationnels de supervision... Ces informations en temps réel seront en effet aussi très utiles aux agents qui régulent la circulation des trains et les flux de passagers en gare. Au final, tous nos trains devraient bénéficier de ces prédictions algorithmiques à l'horizon 2020.

© SNCF

« Et si vous deveniez “data scientist” ? »

Comme l'illustrent les trois précédents articles de ce cahier, les entreprises ont de plus en plus besoin de *data scientists*. Mais comment se former à cette nouvelle profession ? Le point sur les différentes voies pour y parvenir...

« *Métier le plus sexy du XXI^e siècle* ». C'est en ces termes que la prestigieuse Harvard Business Review qualifie le job de *data scientist* ! Un métier en plein boom consistant à extraire de nouvelles connaissances à partir de données (*data*), grâce à des techniques issues des mathématiques appliquées, de la statistique et de l'informatique. En quelques années à peine, les formations se sont multipliées. Alors, comment s'y retrouver ?

Les formations de qualité remplissent généralement trois critères : elles sont souvent d'un niveau master, adossées à des laboratoires ou des chaires... et à un écosystème industriel. L'objectif étant de pouvoir ainsi adapter les enseignements au rythme des évolutions de la recherche et de l'industrie. On trouve de tels masters dans les universités et les grandes écoles, le plus réputé restant le MVA (Mathématiques, Vision, Apprentissage) de l'ENS Paris-Saclay... qui a fêté ses vingt ans en 2017 ! Certains sont plutôt à dominante « maths », d'autres plus à dominante informatique.

Trois casquettes

Car dans la pratique, le métier de *data scientist* en regroupe plusieurs. Il y a bien sûr le concepteur d'algorithmes, nécessitant une solide formation générale en maths, en informatique scientifique et en modélisation. Mais il y a aussi l'« intégrateur » ou « développeur informatique » qui lui met en œuvre des algorithmes et des méthodes d'apprentissage, et doit gérer problèmes opérationnels et informatiques. Sans oublier le « spécialiste métier » ou « utilisateur » : véritable expert dans la manipulation de ces outils, et capable de comprendre le contexte d'utilisation (médical, marketing, transports...).

CEPE, École Polytechnique, Telecom ParisTech, université Paris-Descartes... de nombreux organismes proposent aussi des formations continues qui se sont fortement développées ces dernières années, notamment dans le domaine du *big data*. Enfin, même s'ils ne remplacent pas les formations classiques, les cours en ligne ouverts à tous (Mooc) peuvent aussi se révéler très utiles.



« Les data scientists devront de plus en plus se spécialiser. »

NICOLAS VAYATIS

Responsable du master MVA de l'ENS Paris-Saclay



« L'enjeu est à la fois de faire évoluer nos statisticiens en interne vers ces nouveaux métiers et de recruter en externe. »

OLIVIER BAES

Responsable du Datalab de la MAIF



« Nous avons créé avec l'École des ponts et chaussées la première Chaire en France regroupant optimisation et apprentissage. »

ISABEL GOMEZ GARCIA DE SORIA

Directrice de la Recherche Opérationnelle chez Air France



« Nous initions des concours en sciences des données depuis 2014. »

MAGUELONNE CHANDESIS

Data scientist chez SNCF Innovation & Recherche

En chiffres

- La France aura besoin de **130 000 data scientists** d'ici à 2020
- La **data science** pourrait créer **3 millions d'emplois** en Europe
- Un **data scientist** gagne en moyenne de **45 000 à 55 000 euros** par an

DATA CHALLENGES: DE VRAIS DÉFIS!

Aujourd'hui, de nombreuses formations intègrent des « data challenges », compétitions dans lesquelles des équipes d'étudiants data scientists s'affrontent pour résoudre des problèmes très concrets : prévoir la consommation électrique d'une zone, la fréquentation d'une gare, l'intérêt de consommateurs pour tel ou tel produit, améliorer des algorithmes de recommandations d'achat en ligne, etc. Souvent proposés par des entreprises, y compris les grandes (Air France, SNCF...), ces *data challenges* permettent aussi à certains étudiants de se faire repérer pour décrocher un stage... voire un job !

