



© Air France



© Pixabay



© SNCF

« Quand les “ data scientists ” nous simplifient la vie »

En ce début de XXI^e siècle, la science des données s’immisce tous les jours un peu plus dans notre quotidien. Reconnaissance vocale, moteurs de recherche, publicités ciblées, recommandations d’achats... Elle est bien sûr déjà omniprésente dans nos activités numériques. Mais cette *data science* gagne aussi progressivement tous les secteurs d’activités économiques. Décoller en A380, emprunter sereinement son train quotidien ou bien encore se remettre d’une catastrophe naturelle: voici trois cas très concrets où des algorithmes aussi discrets qu’efficaces sont déjà à la manœuvre. Autant d’applications qui nécessitent les compétences de *data scientists* chevronnés... Un secteur d’activité florissant qui pourrait créer trois millions d’emplois en Europe! Et grâce au nombre et à la qualité de ses formations, la France est bien placée pour ne pas laisser filer sa part du gâteau... Rien que sur l’Hexagone, le recrutement de 130 000 *data scientists* sera nécessaire d’ici à 2020. En ferez-vous partie?

SOMMAIRE

- Décoller à l’heure p. II
- Des sinistrés pris en charge plus vite p. III
- Améliorer l’information temps réel p. IV
- Et si vous deveniez *data scientist*? p. V

Cahier spécial réalisé en partenariat avec



Décoller à l'heure

Les Airbus A380 d'Air France qui atterrissent à Roissy sont aujourd'hui analysés par des algorithmes de maintenance prédictive.

Objectif : empêcher toute panne inopinée source d'importants retards... Reportage.

Lundi 11 décembre 2017, 10 heures du matin. Dans un froid glacial, un Airbus A380 d'Air France en provenance de Washington D.C. atterrit sur le tarmac de l'aéroport parisien Roissy-Charles-de-Gaulle. Les passagers commencent à s'ébrouer puis quittent leurs sièges, tandis que tout un ballet se met en place autour du géant des airs pour sortir les bagages des soutes, récupérer les déchets, faire des contrôles sur l'appareil, etc. Mais ce que les passagers ne savent pas, c'est qu'un autre ballet démarre en même temps, numérique celui-là...

En effet, à peine a-t-il posé les roues sur la piste que notre A380 commence à transmettre par wifi toutes les données enregistrées en continu sur son dernier aller-retour *via* les centaines de milliers de capteurs dont ses équipements sont truffés. Consommation électrique, vitesse, angle, température...

L'ensemble de ces *data* (1 gigaoctet en moyenne) est alors traité par un logiciel de maintenance prédictive baptisé Prognos. Ce dernier transmet ensuite ses résultats en temps réel aux ingénieurs du département maintenance avion d'Air France-KLM.

1 Go de données par vol

C'est là que nous rencontrons Paul-Louis Vincenti, *data analyst* au département recherche opérationnelle chez Air France Industries-KLM Engineering & Maintenance. «*Grâce à une infrastructure big data distribuée dédiée*¹, Prognos détecte en moins de trente minutes les éventuelles signatures de prémices de pannes cachées dans cette masse de données», explique-t-il. Le cas échéant, le service maintenance agit de manière préventive... De quoi éviter toute panne inopinée qui empêcherait l'avion de décoller, générant un retard important, voire une annulation de dernière minute.

Dans la pratique, les algorithmes² de Prognos mettent en œuvre des méthodes de *machine learning* par apprentissage supervisé* qui s'entraînent sur l'historique de précédents vols. «*Nous sommes en effet très attachés à l'explicitabilité des résultats*», précise P. L. Vincenti. *Voilà pourquoi nous évitons volontairement le deep learning, et les réseaux de neurones** : nous voulons toujours pouvoir justifier d'un point de vue physique les décisions prises*».

En 2017, grâce à Prognos, les équipes d'Air France ont remplacé dans les A380 huit pompes d'alimentation moteur, et quatre capteurs de rotation logés dans le nez des trains d'atterrissage. «*Jusqu'à présent, confirme P. L. Vincenti, toutes ces déposes concernaient bien des équipements déclarés défectueux après inspection*». Fort de ces résultats, l'outil de maintenance prédictive a aussi été mis en place sur les dix-huit Boeing 747 de KLM.

* Apprentissage automatique à partir d'exemples annotés
** Méthodes mimant la profondeur des couches d'un cerveau

1. Hadoop Distributed File System
2. Développés en langages Python et Spark, ils s'appuient sur la librairie Scikit learn



« Les algorithmes de maintenance prédictive des A380 doivent toujours rester explicables »



PAUL-LOUIS VICENTI
Data analyst chez Air France-KLM

AMÉLIORER LA RELATION AVEC LES PASSAGERS

Chez Air France-KLM, une seule et même plateforme *big data* centralise aussi toutes les données que la compagnie collecte sur ses clients *via* divers canaux: *call centers*, site web, réseaux sociaux, dans l'aéroport, lors de salons... ou bien encore à bord des avions. Toutes ces données sont passées à la moulinette d'algorithmes qui en tirent des recommandations adaptées au profil de chaque client: options pour le menu à bord du prochain vol, propositions de destinations et de tarifs personnalisés, temps de transport jusqu'à l'aéroport de départ, guidage du passager en correspondance dans l'aéroport... Sur ordinateur ou tablette, tous les personnels au sol ou en cabine auront bientôt accès au dossier de chaque client, et à ses éventuels échanges en cours avec d'autres services de la compagnie. Le tout se fait bien sûr dans le respect de la vie privée définie par la Cnil.



Des sinistrés pris en charge plus vite

Grâce à la *data science*, on commence à prédire l'étendue des dégâts... avant même qu'une catastrophe naturelle s'abatte sur une région donnée. De quoi mieux anticiper la prise en charge des futurs sinistrés. Explications...

Inondations, tempêtes, cyclones, ouragans...
Entre 1988 et 2013, les catastrophes naturelles ont coûté 1,9 milliard d'euros en indemnisations, avec 431 000 sinistrés en moyenne chaque année. Quand des catastrophes de ce type s'abattent sur un territoire, les assureurs activent immédiatement leurs plans de gestion de crise. Principal objectif: traiter au plus vite l'arrivée en masse de dossiers de demande d'indemnisation d'assurés impactés par la catastrophe et qui risquent d'affluer en agences. Mais aujourd'hui les assureurs veulent aller plus loin...

«Data visualisation» des futurs sinistrés



ORLANE MONNET

«À la MAIF, nous développons par exemple un outil capable de prédire le nombre de déclarations de sinistres sur telle ou telle zone, avant même qu'elle ne soit touchée par une tempête hivernale ou une inondation cévenole», lance Orlane Monnet du département Études statistiques de cet assureur. Calibré sur la sinistralité des événements climatiques passés les plus significatifs (tempêtes Klaus, Xynthia, Joachim...), ce modèle de prévision utilise deux grands types de données. Tout d'abord, des données sur les communes et les assurés qui risquent d'être touchés: maisons ou appartements, domiciles ou résidences secondaires, propriétaires ou locataires, densité de population, existence ou non d'un plan de prévention des risques, etc. Ensuite, le modèle utilise aussi bien sûr des données météorologiques: force du vent prévue, durée estimée, quantité de précipitations à venir, etc.

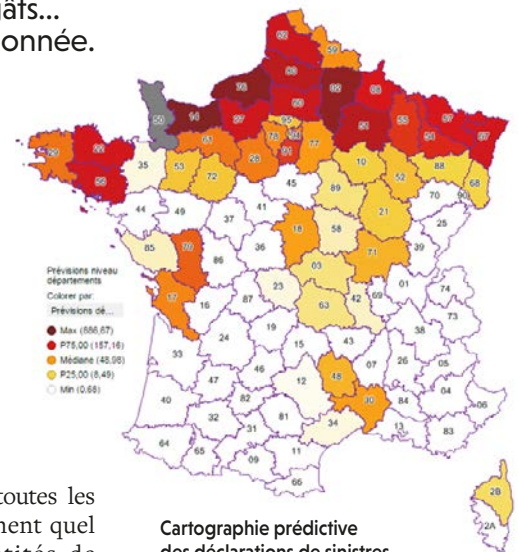
«Écrit en langage R*, notre outil se lance et s'utilise via une seule et unique interface de data visualisation** nommée *Tibco Spotfire*, indique Orlane Monnet. Elle permet de visualiser sur une carte le nombre de déclarations de sinistres prévisibles – plus les zones sont touchées plus elles apparaissent foncées.» Par simple clic sur des curseurs, elle permet aussi de mesurer la sensibilité du modèle de façon interactive; exemple, si on majore par sécurité toutes les rafales de 10 km/h, on voit directement quel pourcentage augmente les quantités de déclarations de sinistres sur la carte. La MAIF a commencé à utiliser son outil sur des inondations fin 2016 et sur de petites tempêtes début 2017.

Anticiper les réparations

À l'avenir, ce type d'outil pourrait encore être amélioré pour évaluer plus précisément les coûts attendus, et même commencer à mobiliser les entreprises réparatrices (loueurs de moto-pompes et de groupes électrogènes, réparateurs de carrosseries de voitures...). Mais pour cela, il faudrait collecter des données encore plus précises sur les assurés: équipements, habite à tel étage, arbres sur son terrain, type de toiture, parking aérien ou souterrain, terrain en forte ou faible pente... Enfin, ce type d'outil pourrait aussi servir à alerter par mail ou SMS les assurés concernés et leur prodiguer des conseils de prévention personnalisés.

* Langage informatique dédié aux statistiques et à la science des données

** En mettant en images les données, la «dataviz» est un moyen d'identifier des tendances et corrélations



Cartographie prédictive des déclarations de sinistres par département

«L'usage d'algorithmes de machine learning nous aidera à encore mieux assurer les biens de nos sociétaires»



© Darri MAIF

STÉPHANE RENOUX
responsable du programme
Data & intelligence artificielle
à la MAIF



Datavisualisation du trafic TGV un lundi matin

© 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 CHANDESRIIS
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!

