4 idées reçues sur le Machine Learning

4 idées reçues sur le Machine Learning

Un certain nombre d'idées reçues viennent en tête lorsque l'on parle de Machine Learning (ML), car à juste titre on l'associe au Big Data. Toutefois, bien que le ML soit une composante analytique des projets Big Data, il ne porte pas à lui seul, l'ensemble des contraintes et fantasmes liés à ce type de projets.

A cet égard, nous allons balayer 4 idées reçues sur les projets de Machine Learning. Afin de présenter simplement le Machine Learning et de répondre aux idées reçues, nous allons illustrer notre démonstration par analogie avec un physionomiste à l'entrée d'une discothèque.

1. «Le ML, c'est pour les grands groupes qui ont beaucoup de données»

Le rôle du physionomiste est de choisir les meilleurs clients (clients cibles) afin d'assurer le plus haut revenu possible, tout en évitant les risques de bagarres.Pour cela il dispose de son expérience passée (ses datas) et de son réseau de neurones (algorithmes). Ainsi, sur un certain nombre de critères (des variables) il pourra prendre sa décision de faire, ou non, rentrer les individus.

Exemples :

a/ groupe de plus de 4 + exclusivement masculin + agités + éméchés = ne pas faire rentrer.

b/ Couple + mixte + calme + sobre = faire rentrer

Dans ces exemples nous disposons de 4 critères, les neurones de notre physionomiste permettent de traiter ces informations et de prendre les décisions qui s'imposent.

Toutefois, si à ces éléments, nous décidions de rajouter un grand nombre de variables ;

- l'âge de l'individu,
- sa profession,
- son niveau d'étude,
- sa taille,
- la marque de ses chaussures,
- le motif de la sortie,
- le mois,
- le jour,
- l'heure,
- la température extérieure,
- …

La bonne décision serait bien plus compliquée à prendre.

Ainsi à partir d'un certain nombre de variables, le cerveau humain n'est plus capable d'identifier les signaux faibles contenus dans les données. Grâce au ML, le data scientist est capable de modéliser les bonnes décisions.

La problématique du ML n'est donc pas le nombre d'enregistrements (nombre de personnes se présentant devant l'établissement) mais l'analyse d'un grand nombre de variables ne pouvant pas d'être appréhendées par le cerveau humain.

Ainsi, toute entreprise disposant d'un journal de facturation dispose de la matière suffisante pour tirer parti du ML : cross selling, up selling, classification des clients selon des logiques d'achat, identification des clients mûrs, …

2. Il faut des données propres et complètes.

Contrairement à la comptabilité ou à la Business Intelligence qui nécessitent 100% des données pour être juste (chiffre d'affaires = somme de toutes les ventes), le ML n'a besoin que d'un échantillon représentatif pour élaborer un modèle.

Pour l'illustrer, notre physionomiste aura besoin d'un historique de clients suffisant pour prendre une bonne décision, sans pour autant avoir à se souvenir de toutes les personnes individuellement.

D'autre part et dans une certaine mesure, si certains enregistrements sont incomplets (valeurs manquantes), cela ne sera pas non plus problématique.

Notre physionomiste saura juger ponctuellement un client même si il ne connait pas son âge (l'indice de confiance sera alors plus faible).

3. La mise en œuvre est complexe, coûteuse et longue.

Contrairement à la mise en place d'architectures Big Data (Hadoop), où le moindre POC peu prendre plusieurs mois et nécessiter de nombreuses compétences (internes et externes), l'utilisation du ML peut être très rapide et reposer sur une seule personne. Le minimum requis étant un fichier (type csv) et un Data Scientist (DS).

A partir du ou des fichiers sources, le DS va créer en quelques jours un modèle de ML. Par l'analyse des résultats générés, il saura si les informations contenues dans les sources sont suffisantes ou non. Dans le second cas le DS devra trouver d'autres sources d'information internes ou externes (Open Data).

Par analogie, notre physionomiste tentera d'évaluer ses clients à travers l'âge, le groupe, l'attitude, si les résultats ne sont pas satisfaisants, il devra intégrer de nouveaux critères.

Ainsi, le ML peut donner des résultats très rapidement sans mobiliser les ressources internes.

Si le modèle s'avère rentable, qu'un besoin de «temps réel» existe, ou que le volume de données le nécessite, il sera alors temps de penser à mettre en œuvre l'architecture technique adéquate.

4. Le retour sur investissement est difficile à évaluer

L'idée selon laquelle il faille stocker les datas quelles qu'elles soient, coûte que coûte, sans savoir ce que l'on en fera, contribue à brouiller le calcul du ROI des projets BigData.

Les modes de consommation, les comportements évoluant très rapidement, cela reviendrait à stocker des données périmées avant même qu'on en ait besoin.

La valeur dégagée par l'usage du ML doit pouvoir être évaluée et préalablement objectivée. Pour chaque problématique métier on doit disposer d'indicateurs (KPI) permettant de comparer la situation initiale (avant l'usage du ML) à la situation finale : taux de retours de campagne de communication, nombre de transformation de devis, montant de marge, indice de satisfaction, taux de pannes…

Pour illustrer nos propos, imaginons maintenant, que notre physionomiste soit face à 100 clients souhaitant pénétrer dans l'établissement et qu'il ne dispose plus que de 10 places disponibles. Imaginons encore que parmi les 100 personnes seules 10 sont prêtes à acheter une bouteille de champagne (clients cibles).

i m a g e :

https://lh4.googleusercontent.com/JniA4T0TU73H39UkNVUftyl5JvkBR0SxshcT75cfWapva8Jp0PFanonNntqE9SPf-rroxoeD-yf_zBcKbQMkI8GBJUnb7Q_tqy_srgVk-eUQEbw12Yj4h-lk07lMogjfJGtOYkE6

Si notre physionomiste choisit au hasard les 10 personnes qu'il fera rentrer, statistiquement il aura fait rentrer 1 client cible. Si il est capable de bien modéliser ses clients cibles, il en fera rentrer, 2, 3, 5 voire 10 si son modèle est parfait (théorique). En revanche, si le choix du physionomiste est biaisé (exemple : il ne fait rentrer que ses connaissances) il peut ne faire rentrer aucun client cible. Le ML quant à lui utilise des faits objectifs sans biais humain (affect, goût, croyance…).

L'indicateur d'évaluation du ROI sera le nombre de bouteilles vendues avant l'usage de ML et après.

En conclusion, le Machine Learning est une composante analytique du Big Data, pouvant être mise en œuvre indépendamment de la composante architecturale. Il est ainsi possible de se lancer dans le Big Data par des projets de Machine Learning à haute valeur ajoutée.

L'avantage étant que contrairement à certaines idées reçues, une PME avec relativement peu de données, même partiellement incomplètes, pourra grâce au Machine Learning, rapidement et à faible coût, exploiter et mesurer de nouveaux gisements valeur.

Réagissez à cet article

Source : *4 idées reçues sur le Machine Learning | FrenchWeb.fr*