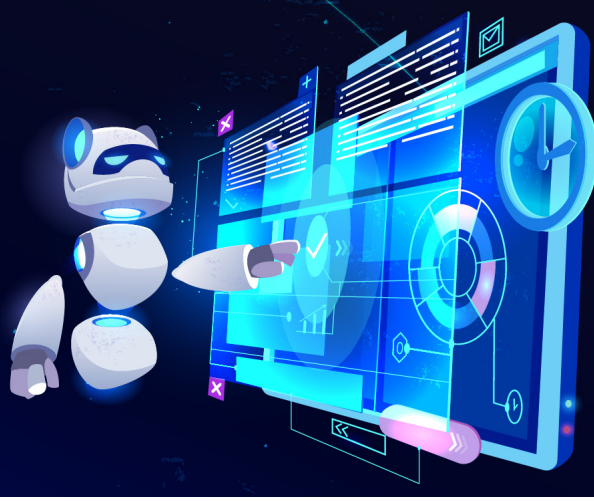
AI & DATA

6 ÉTAPES CLEFS POUR RÉUSSIR VOTRE MODÈLE DE PRÉDICTION



Pouvoir prédire l'évolution future de votre entreprise ou des comportements de vos clients est un atout puissant dans l'orientation que prendra votre stratégie. C'est pourquoi les modèles prédictifs ont un réel attrait et impact stratégique pour toute entreprise. Mais comment réussir ces derniers ? Beaucoup de paramètres sont à prendre en compte pour le succès de ces modèles.

Dans cet article, nous allons vous décrire 6 étapes clés, pour selon nous, réussir un modèle prédictif. Une partie sera consacrée à l'implémentation de ces étapes sous Dataiku.

A travers notre expérience dans l'établissement de nos nombreux modèles prédictifs, nous avons pu constater de nombreux freins auxquels peuvent se heurter certaines équipes :

- Une tendance à partir directement dans la modélisation dès la définition du problème sans préparer suffisamment les variables au préalable. L'expérience prouve que c'est une erreur.
- Une tendance à abuser de la complexité en utilisant systématiquement les derniers modèles populaires par exemple.
- Un manque de recul par rapport à la problématique traitée est aussi une source d'échec car elle entraîne un manque de discernement dans le choix des variables, du périmètre d'études, etc.

Un beau modèle prédictif mérite une belle méthodologie que nous allons vous proposer en 6 étapes.

1 QUALITÉ DES DONNÉES ET TRAITEMENT DES DONNÉES MANQUANTES

Dans un véritable contexte professionnel dans le domaine de la data, **la première étape primordiale est de récupérer les données dispersées dans des sources différentes.** Il faut ensuite étudier la qualité de celles-ci et éventuellement mettre en œuvre des modèles de traitement / imputation de données manquantes. Les données aberrantes (outliers) doivent être éliminées et le périmètre restreint à des individus suffisamment bien renseignés pour construire un modèle de qualité.

Dataiku peut être utile pour ce faire compte tenu de ses grandes capacités de requêtage et détection de données manquantes. Il est également réputé pour avoir un certain nombre de connecteurs qui permettent cette récupération et ainsi permettre facilement d'importer des données et de changer les types et les paramètres de ces dernières. Dataiku a une ingestion intelligente de la donnée qui permet facilement et très rapidement de voir la distribution des données, les outliers et les données manquantes, offrant ainsi des possibilités de nettoyage rapide.

2 DE LA DONNÉE BRUTE À L'INFORMATION PERTINENTE

Il n'est pas possible de créer un modèle sans avoir transformé la donnée brute en information pertinente : c'est ce que l'on appelle le **feature engineering.**

Dans un premier temps, il est important d'**avoir une réflexion à la fois en termes de pertinence mais aussi d'éliminations de variables inutiles** dans le cadre de votre étude. Vient ensuite le travail le plus laborieux et chronophage d'un projet data.

Dataiku offre de nombreuses opérations prédéfinies : plusieurs possibilités en mode formule, SQL, script avec du python, etc.

D'autre part, il semble important d'évoquer le « **leakage** » qui diminue fortement les performances de la modélisation. Cela peut provenir du choix des variables qui se déduisent de l'information à prédire, d'informations du futur non connues à la date de prédiction ou de doublons existants entre le jeu de données d'apprentissage et de test.

3 COMPÉTITION ENTRE MODÈLES

Il est souvent intéressant de mettre en compétition plusieurs modèles car il n'est pas impossible d'avoir des à priori sur la famille de modèles que l'on pourrait utiliser : des plus simples comme la régression logistique aux plus sophistiqués comme les réseaux de neurones.

Dataiku permet facilement cette mise en compétition des modèles implémentés. Il est possible de coder soi-même son propre modèle, mais attention aux autorisations de l'administrateur. Il faut prévoir un délai pour obtenir de la DSI les droits de téléchargement des librairies Python. Dataiku facilite le choix du meilleur modèle d'un point de vue statistique. Grâce à des KPIs standardisés quelle que soit la famille du modèle, permettant ainsi de comprendre les variables influentes et de comparer les modèles.

4 PRÉCISION ET/OU EXACTITUDE

Privilégier la précision ou l'exactitude est une décision pleine de conséquence.

Les KPIs peuvent être antagonistes entre le nombre de personnes prédites dans une classe qui ont réellement cette caractéristique et le nombre de personnes ayant une caractéristique et étant bien prédits. L'arbitrage doit prendre en considération des aspects économiques : si on offre un avantage à l'une des classes d'individus, on risque de le fournir à des individus qui n'y aurait pas droit (liée à l'impureté de la prédiction). A l'opposé, si on est trop strict, peu de personnes ayant droit à cet avantage n'y aurait accès, d'où le risque d'insatisfaction de bons clients.

Dataiku permet d'arbitrer de manière interactive cette antagonisme.

5 UN OU PLUSIEURS MODÈLES ?

Il ne faut pas négliger de **se poser les bonnes questions.** Par exemple : préfère-t-on un modèle universel ou essaye-t-on de juxtaposer des modèles plus spécialisés ? Cela est notamment pertinent si on constate que le taux de mal prédit peut-être lié à plusieurs informations. Pour une sous population, certains facteurs sont importants tandis que pour le reste de la population, d'autres facteurs seront retenus et le fait de les séparer permet ainsi la création de 2 modèles.

Dataiku permet de le faire sans avoir à tout réécrire, à réintroduire d'autres variables qui pesaient peu dans le modèle général.

6 FRUGALITÉ DU MODÈLE

La dernière étape, non négligeable, de cette méthodologie invite à **réfléchir sur la frugalité du modèle** (empreinte carbone et explicabilité des monnaies).

Nous proposons d'axer la réflexion autour de l'idée suivante. Arriver à obtenir une forme de frugalité :

- **En diminuant le nombre de variables utiles** (moins de mémoire, de stockage),
- **En choisissant à performance quasi égale le modèle le plus facile à calculer, le plus rapide et le moins itératif**
- **En mettant le modèle au point sur un échantillonnage de donnée.** L'expérience montre qu'à partir d'une certaine volumétrie, on ne gagne pas en précision dans le résultat du modèle obtenu.
- **En espaçant les mises à jour du modèle :** si les résultats sont stables, il n'y a pas d'intérêt de refaire tourner quotidiennement une segmentation de clients. 1 fois tous les 6 mois peut être suffisant.

Ces points gagneront en importance dans les années à venir.

En conclusion, nous pensons que les modèles vont devoir être de plus en plus explicables, notamment dans le cas de modèles influant sur une décision prise à l'encontre d'un individu.

Il est primordial d'avoir cette vision frugalité/explicabilité à chacune des 6 étapes de la méthodologie.

Vous souhaitez en apprendre davantage ou collaborer sur la mise au point de modèles prédictifs de comportement sous Dataiku ou ailleurs ?

Contactez-nous : contact@aid.fr