Comment créer de la valeur métier grâce à l'IA

12 récits sur le terrain

IBM

Comment IBM peut vous aider

Les clients peuvent tirer parti de l'expertise sectorielle, fonctionnelle et technologique approfondie, de l'analyse et des données, des solutions technologiques d'entreprise et des innovations dans la recherche scientifique d'IBM pour exploiter tout le potentiel de l'IA, Pour plus d'informations sur les services d'IA d'IBM Consulting, visitez le site ibm.com/fr-fr/services/artificial-intelligence.

Pour plus d'informations sur les solutions d'IA d'IBM Software, accédez au site ibm.com/fr-fr/Watson.

Pour plus d'informations sur les innovations IA d'IBM Research®, accédez au site research.ibm.com/artificial-intelligence.

Pour plus d'informations sur le MIT-IBM AI Lab, accédez au site mitibmwatsonailab.mit.edu.

Synthèse

De nombreuses idées reçues sur l'intelligence artificielle (IA) sont en fait des mythes trompeurs, engendrés par le cycle de battage médiatique, endémique pour tant de technologies émergentes.

Coup d'œil derrière le rideau de l'IA

L'IBM Institute for Business Value (IBV), en collaboration avec le MIT-IBM Watson AI Lab, a interrogé des personnes impliquées dans des projets d'apprentissage en profondeur dans plus de 35 implémentations réelles d'intelligence artificielle (IA) dans le monde. Nous nous sommes entretenus avec des experts métier et technologiques de plus d'une douzaine de secteurs d'activité sur leurs objectifs et leurs défis et les enseignements qu'ils tirent de l'IA.

Petits gains ou transformation à grande échelle

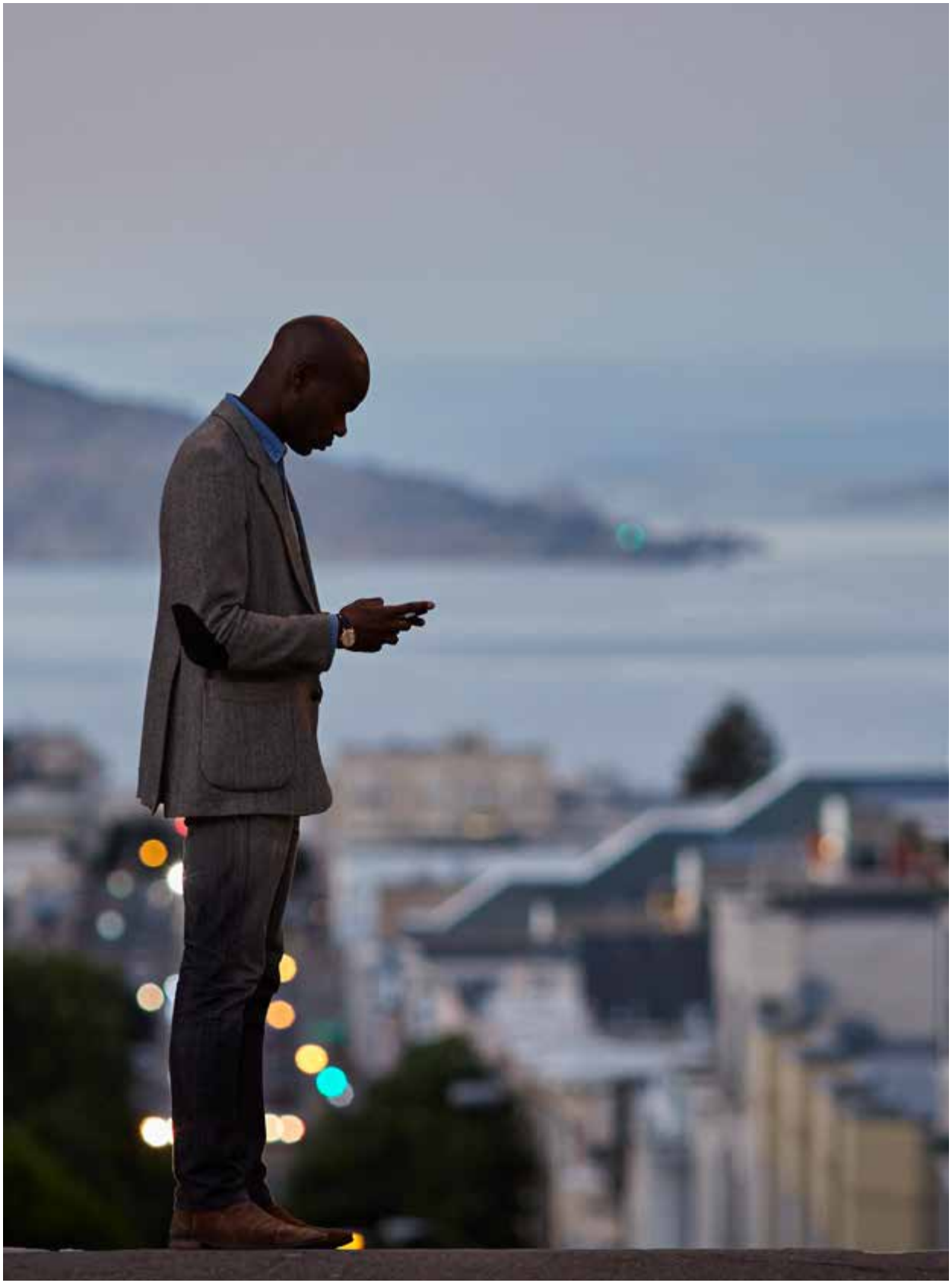
Nous avons confirmé que l'adoption de l'IA se poursuit, mais la plupart des organisations ne la mettent pas complètement en œuvre pour réaliser une transformation à grande échelle. En effet, nombreuses sont celles qui se contentent de relever des défis opérationnels distincts. D'ici à la fin de 2022, nous estimons que seule une grande entreprise sur quatre aura dépassé le stade des projets pilotes pour passer à l'IA opérationnelle.[1]

Au-delà des mythes pour comprendre la réalité de l'IA

Alors que les entreprises adoptent l'intelligence artificielle, leurs décideurs et les autres dirigeants n'adhèrent pas à certains mythes qui l'entourent, tels que « Avec l'IA, pas de raccourcis possible » ou « Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA ». En fait, ils doivent véritablement ancrer leurs décisions dans la réalité de l'IA.

Apprendre de vos pairs dans les secteurs d'activité

Dans ce document, nous levons le voile sur cinq mythes et révélons, à l'aide de données et d'exemples concrets, la vérité sur la manière dont les entreprises utilisent l'IA, afin que les dirigeants et les équipes des organisations puissent tirer des enseignements de leurs pairs.



Introduction

Une question de perception par rapport à la réalité.

Alors que les gros titres annoncent l'intelligence artificielle (IA) comme la panacée au malaise économique croissant, les dirigeants continuent de se demander ce que font réellement les entreprises avec l'IA. Quels résultats obtiennent-elles et comment y parviennent-elles ?

L'IBM Institute for Business Value s'est associé au MIT-IBM Watson AI Lab pour interroger plus de 35 organisations, afin de répondre à ces questions et à d'autres. Ces entretiens nous ont permis de connaître la manière dont les experts métier et technologiques impliqués dans des projets d'apprentissage en profondeur appliquent l'intelligence artificielle dans le monde réel de l'entreprise pour générer réellement de la valeur ajoutée.

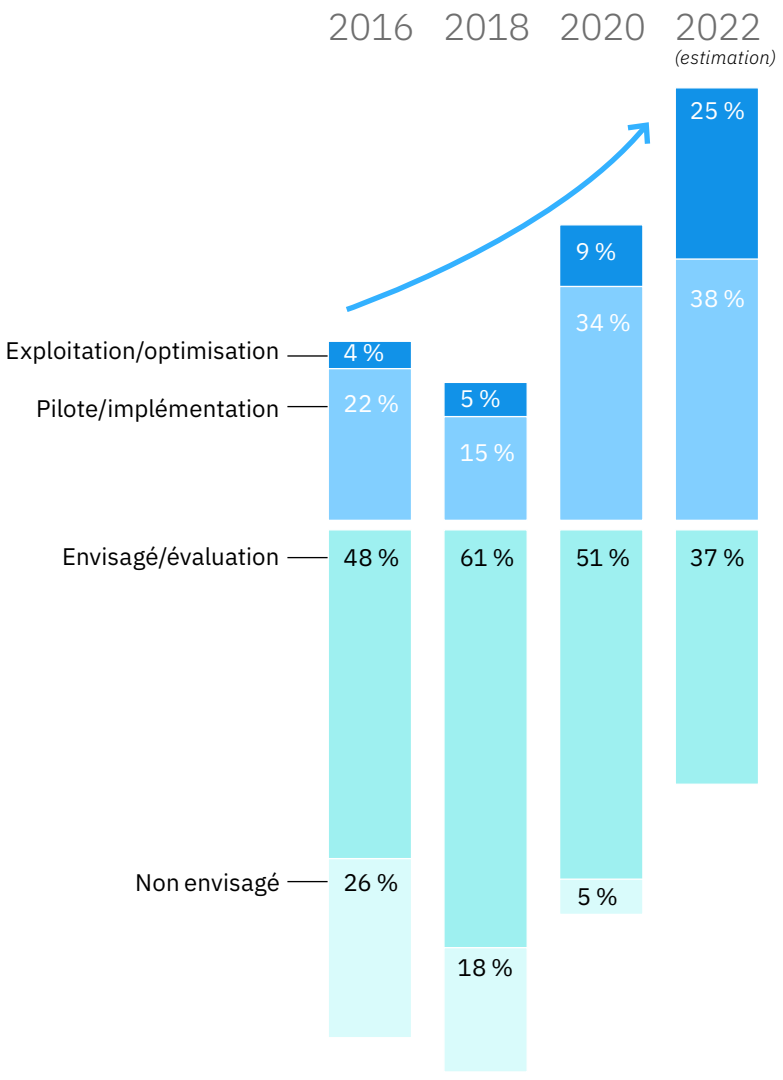
IA : Au-delà des chiffres, des expériences concrètes

L'intelligence artificielle continue de progresser régulièrement tout au long de sa courbe d'adoption technologique, ou de son cycle de battage médiatique, pour les plus cyniques (voir la figure 1).

FIGURE 1

Adoption de l'IA* 2016–2022

D'ici fin 2022, une grande entreprise sur quatre sera passée des pilotes à l'IA opérationnelle.



** Remarque : l'adoption de l'IA comprend les pilotes, l'implémentation, l'exploitation ou l'optimisation.*
Pour plus d'informations, voir la remarque 1 en fin de document.

La pandémie a freiné, puis accéléré, son adoption dans les entreprises. Le nombre d'entreprises qui pilotaient des cas d'utilisation de l'IA au milieu de la pandémie a plus que doublé par rapport à 2018, et des données récentes indiquent que ce nombre augmente.[2]

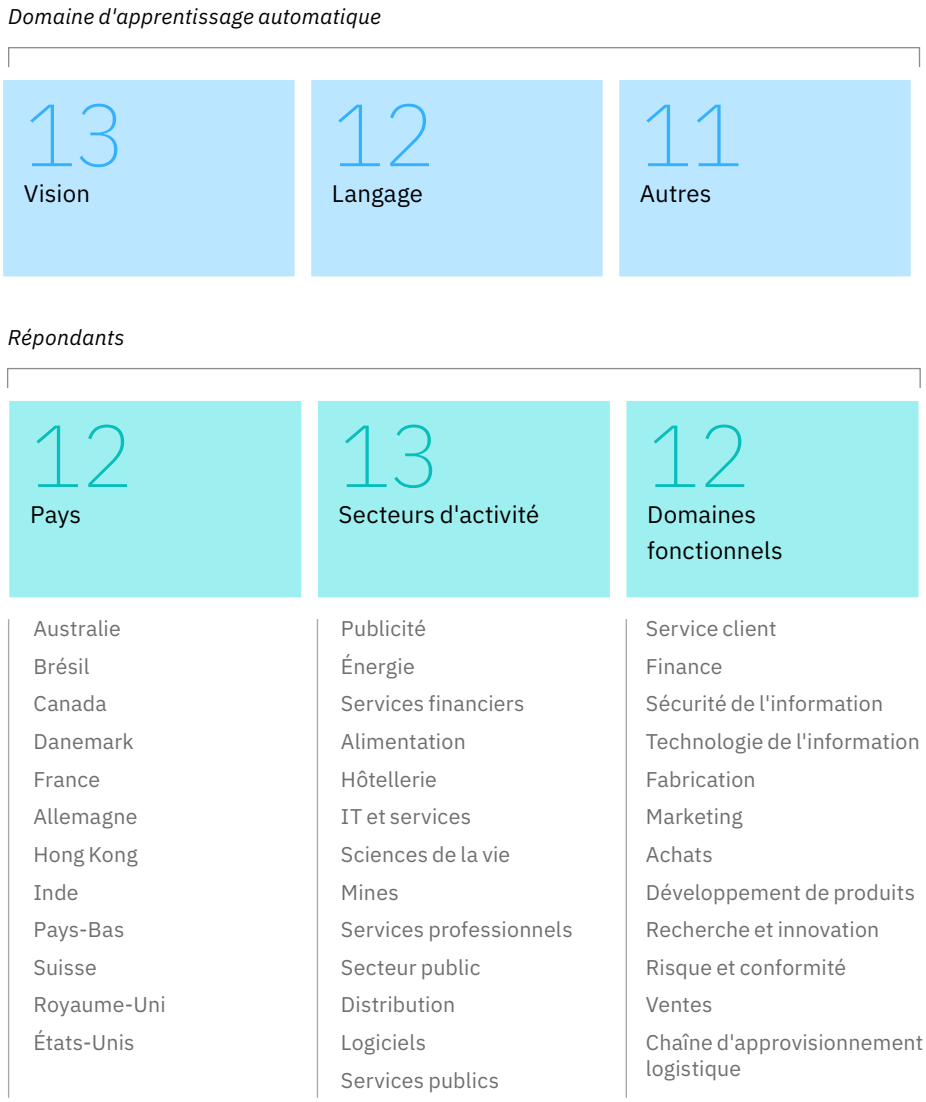
Si ces chiffres montrent une tendance à la hausse, ils ne disent pas tout ce dont de nombreux dirigeants d'entreprise et technologiques nécessitent pour évaluer l'utilisation de l'IA dans leur propre organisation.

Pour en savoir plus à ce sujet et sur les défis que l'IA contribue à résoudre, nous avons interrogé des personnes qui participent à des projets d'apprentissage en profondeur dans le monde entier. D'avril à août 2021, nous avons interrogé des experts métier et technologiques dans plus d'une douzaine de secteurs d'activité sur leurs objectifs, leurs défis et les enseignements qu'ils tirent de l'IA (voir la figure 2).

FIGURE 2

Périmètre et échelle de nos entretiens

Nos entretiens révèlent comment des utilisations personnalisées de l'IA peuvent résoudre des problèmes métier distinctes.



Qu'avons-nous appris sur l'état de l'IA ?

L'IA peut-elle être un catalyseur de croissance dans l'entreprise, Absolument. Pour certains adoptants innovants de l'IA, tels que NVIDIA, NavTech et d'autres, l'IA permet de créer des offres entièrement nouvelles et même de nouveaux modèles de gestion.

Cependant, peu d'entreprises l'utilisent encore pour réaliser une transformation aussi large. En effet, elles s'efforcent principalement de résoudre des problèmes métier distincts et tangibles. Les organisations du monde entier utilisent l'IA, entre autres, pour réduire les coûts, enrichir les expériences clients et des employés, augmenter les taux de réussite et optimiser les performances de la chaîne d'approvisionnement.

Nous avons également appris que de nombreuses idées reçues sur l'IA sont en fait des mythes trompeurs, alimentés par le cycle de battage médiatique qui est devenu endémique pour tant de technologies émergentes. Malheureusement, ces idées fausses dissuadent généralement les organisations de s'engager dans les réalités plus pragmatiques de l'IA.

Dans les pages qui suivent, nous détruisons cinq des mythes les plus répandus sur l'IA en mettant en lumière des idées pertinentes et des exemples pratiques tirés de nos entretiens. Les observations et anecdotes qui suivent, issues de nos discussions avec plus de 55 professionnels de plus de 35 organisations en première ligne de l'IA, peuvent permettre de démêler le vrai du faux. Elles permettent de jeter un coup d'œil derrière le rideau, une rencontre virtuelle avec des pairs, alors que les organisations cherchent à accroître l'impact et la valeur de l'IA. (Les lecteurs qui souhaitent en savoir plus trouveront 12 études de cas détaillées en annexe).

Mythe et réalité

Mythe 1

L'IA est universelle

Mythe 2

Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas réellement de l'IA

Mythe 3

La réduction des coûts est le point fort de l'IA

Mythe 4

Avec l'IA, pas de raccourcis possibles

Mythe 5

L'IA n'apporte de la valeur ajoutée qu'au niveau du problème traité

**Mythe 1 -
L'IA est universelle**

Mythe 2 -
Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA

Mythe 3 -
La réduction des coûts est le point fort de l'IA

Mythe 4
Avec l'IA, pas de raccourcis possibles

Mythe 5
L'IA n'apporte de la valeur ajoutée qu'au niveau du problème traité

Annexe

Mythe 1

L'IA est universelle

Réalité

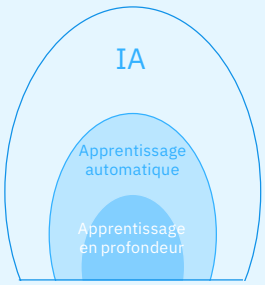
L'adéquation aux objectifs est importante. Les améliorations métier induites par l'IA découlent de nombreuses techniques.

L'une d'entre elles, l'apprentissage en profondeur, par exemple, convient généralement mieux pour résoudre les problèmes liés aux jeux de données sous-jacents (souvent volumineux) dans les domaines de la vision, du langage et d'autres modèles prédictifs. Des assistants virtuels à la détection des fraudes, l'apprentissage en profondeur modifie notre façon de travailler et de jouer. Dans ces contextes, les techniques traditionnelles d'apprentissage automatique peuvent se révéler moins efficaces.

Mais l'IA n'est pas à même de relever tous les défis métier ou de générer les résultats voulus, malgré le battage médiatique qui pourrait le laisser croire. Les organisations doivent d'abord déterminer si une initiative stratégique plus large ou un problème métier donné se prête à l'IA, un sujet que nous abordons plus en détail dans la section « Repenser votre approche de l'IA ». Les entreprises peuvent commencer par évaluer leur « richesse en données » globale et examiner des problèmes métier spécifiques.

Perspective

Définition de l'intelligence artificielle, de l'apprentissage automatique et de l'apprentissage en profondeur



Par analogie aux poupées russes, l'apprentissage en profondeur est un sous-ensemble de l'apprentissage automatique, lui-même un sous-ensemble de l'intelligence artificielle. Ces techniques sont souvent complétées par la robotique, des capteurs et des actionneurs de l'Internet des objets, des interfaces virtuelles et d'autres technologies adjacentes.

Qu'est-ce que l'intelligence artificielle (IA) et d'où vient-elle ?
L'IA permet aux ordinateurs d'effectuer des tâches qui auparavant ne pouvaient être réalisées que par des êtres humains. Mais lorsque les capacités humaines commencent à atteindre un plateau de précision, de vitesse et de puissance de traitement, l'IA commence à présenter un véritable intérêt.

Si la frénésie que suscite l'IA est très ancrée dans le 21e siècle, elle est en fait née il y a plusieurs décennies, au milieu du 20e siècle. En 1955, deux professeurs de mathématiques (de Dartmouth et d'Harvard) et deux chercheurs (de Bell Labs et d'IBM) ont suggéré qu'une « étude....de 2 mois sur l'intelligence artificielle soit menée pendant l'été [suivant] ». Selon la synthèse de cette proposition, il s'agissait de « tenter de déterminer la manière de procéder pour que des machines utilisent le langage, créent des abstractions et des concepts, résolvent des types de problèmes jusque-là dévolus aux être humains et s'améliorent elles-mêmes. » [3]

C'est ainsi qu'est née la première définition formelle de l'intelligence artificielle. Depuis, les universités et les entreprises s'efforcent de créer une IA toujours plus performante.

Qu'est-ce que l'apprentissage automatique ?
Dans leur livre « Deep Learning » publié par MIT Press, les auteurs abordent l'apprentissage automatique comme suit : « Les systèmes d'IA doivent être capables d'acquérir leurs propres connaissances, en extrayant des patterns à partir de données brutes. Cette aptitude s'appelle l'apprentissage automatique ». [4] Autrement dit, un ordinateur apprend à partir de jeux de données complexes, et devient plus intelligent à mesure qu'il apprend.

Aujourd'hui, nous utilisons des systèmes d'apprentissage automatique à diverses fins, notamment, pour sélectionner les résultats les plus pertinents en réponse à une recherche par mot-clé ou pour analyser des images visuelles, entre autres. Ces applications de l'IA font de plus en plus appel à une catégorie de techniques appelée apprentissage en profondeur.[5]

Définition de l'apprentissage en profondeur
L'apprentissage en profondeur est un sous-ensemble de l'apprentissage automatique, inspiré du fonctionnement du réseau de neurones du cerveau humain. Parmi les techniques d'apprentissage automatique utilisées actuellement, la plus importante est l'apprentissage en profondeur. Il peut :

- utiliser des données non structurées telles que des images et du texte en format libre
- modéliser des relations non linéaires, afin de modéliser des problèmes complexes
- apprendre des relations sans être préprogrammé sur la tâche cible
- améliorer sa puissance prédictive à mesure que nouvelles données deviennent disponibles

Pour les problèmes complexes où suffisamment de données sont disponibles, l'apprentissage en profondeur est généralement plus performant que d'autres méthodes d'apprentissage automatique.

Une enseigne de mode européenne utilise l'IA pour accroître l'efficacité et la durabilité

La prévision de la demande et l'efficacité des ventes ont toujours été au centre des secteurs des biens de consommation et de la distribution ; même des améliorations progressives peuvent avoir un impact considérable sur l'entreprise.

BESTSELLER, une enseigne de vêtements, voulait améliorer la précision de ses prévisions de la demande pour accroître le plus possible son volume de ventes. À l'époque, elle vendait déjà 78 % des produits qu'elle fabriquait, une performance relativement remarquable dans le monde volatile de la mode. Mais si BESTSELLER pouvait augmenter la granularité des attributs des produits utilisés dans ses algorithmes de prévision, elle pourrait encore faire mieux.

Lorsque les équipes constatèrent que les techniques d'analyse traditionnelles avaient atteint leurs limites, BESTSELLER décida d'entraîner un réseau de neurones convolutifs (CNN) à partir d'images de ses vêtements. (Un réseau CNN est une catégorie de réseaux de neurones artificiels couramment utilisés pour analyser des images visuelles). Ainsi, BESTSELLER put classer ses produits en fonction de caractéristiques supplémentaires qui ne figuraient pas dans ses jeux de données structurées.

En intégrant ces résultats dans son moteur de prévision principal, l'entreprise put augmenter l'efficacité des ventes de 82 % et réduire les échantillons de conception nécessaires de 15 %, des améliorations bienvenues pendant la période de fort ralentissement des ventes lié à la pandémie. Ce changement eut également un impact positif sur le développement durable, car l'entreprise diminua le nombre de vêtements vendus à prix réduit, donnés ou mis au rebut.

La société américaine d'arômes alimentaires McCormick recourt à l'intelligence artificielle pour compléter l'expérience des scientifiques débutant en produits alimentaires pour qu'ils soient aussi performants que leurs aînés ayant 20 ans d'expérience.

Marketing Platform augmente le taux de réponse en appliquant des techniques d'apprentissage automatique

Une agence de marketing et de publicité a utilisé des modèles d'apprentissage automatique pour prévoir la réceptivité des consommateurs aux campagnes de ses clients. Elle a intégré cette fonctionnalité dans une plateforme de données et d'analyse destinée à l'ensemble de ses clients. Une amélioration minime de la précision à grande échelle peut se convertir en millions (USD) de ventes supplémentaires. Par conséquent, les enjeux sont importants.

L'agence découvrit qu'elle pouvait augmenter les taux de réponse de 20 à 30 %, mais que cela augmentait également les coûts de calcul pour stocker, entraîner et traiter plus de données et de paramètres du modèle. Mais en migrant vers le cloud, elle bénéficie d'un plus grande visibilité sur ses coûts, et donc de meilleures informations sur la façon de les gérer. Ainsi, l'équipe a pu continuer d'accroître le taux de réponse tout en utilisant plus efficacement les ressources de calcul et en réduisant de deux tiers les coûts de traitement.

L'IA au service des centres d'appels, de la science des aliments, entre autres

Le Crédit Mutuel, un groupe bancaire coopératif français, utilise largement l'apprentissage en profondeur pour soutenir les agents humains des centres d'appels et économise, ainsi, des dizaines de milliers d'heures chaque mois.

De la même manière, mais dans un secteur d'activité complètement différent, la société américaine d'arômes alimentaires McCormick recourt à l'intelligence artificielle pour compléter l'expérience des scientifiques débutant en produits alimentaires pour qu'ils soient aussi performants que leurs aînés ayant 20 ans d'expérience.

D'autres exemples issus de nos entretiens montrent comment des utilisations personnalisées de l'IA peuvent résoudre des problèmes opérationnels distincts, dans des zones géographiques, des secteurs d'activité et même des fonctions différentes. Généralement, la bonne approche émerge après avoir choisi le jeu de données approprié pour résoudre le problème, comme le soulignent les exemples de BESTSELLER et de l'agence marketing.

Mythe 1
L'IA est universelle

**Mythe 2 -
Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA**

Mythe 3
La réduction des coûts est le point fort de l'IA

Mythe 4
Avec l'IA, pas de raccourcis possibles

Mythe 5
L'IA n'apporte de valeur ajoutée qu'au niveau du problème traité

Annexe

Mythe 2

Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas réellement de l'IA

Réalité

Les grandes entreprises résolvent des problèmes métier distincts et génèrent une valeur métier significative en combinant la science des données, l'apprentissage automatique traditionnel, l'apprentissage en profondeur et des techniques de pré-traitement.

Au cours de la dernière décennie, c'est dans l'apprentissage en profondeur que la plupart des avancées dans la recherche IA ont été réalisées. La croissance exponentielle des plateformes de médias sociaux, de recherche, de distribution, de diffusion en flux et d'autres plateformes B2C qui intègrent l'apprentissage en profondeur dans leurs modèles de gestion a fait émerger l'idée, fausse, que si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA.

En réalité, l'apprentissage en profondeur n'est qu'un outil parmi d'autres dans la boîte à outils de l'analytique de l'entreprise qui permet de mettre en œuvre de l'IA (voir la figure 3 à la page 10).

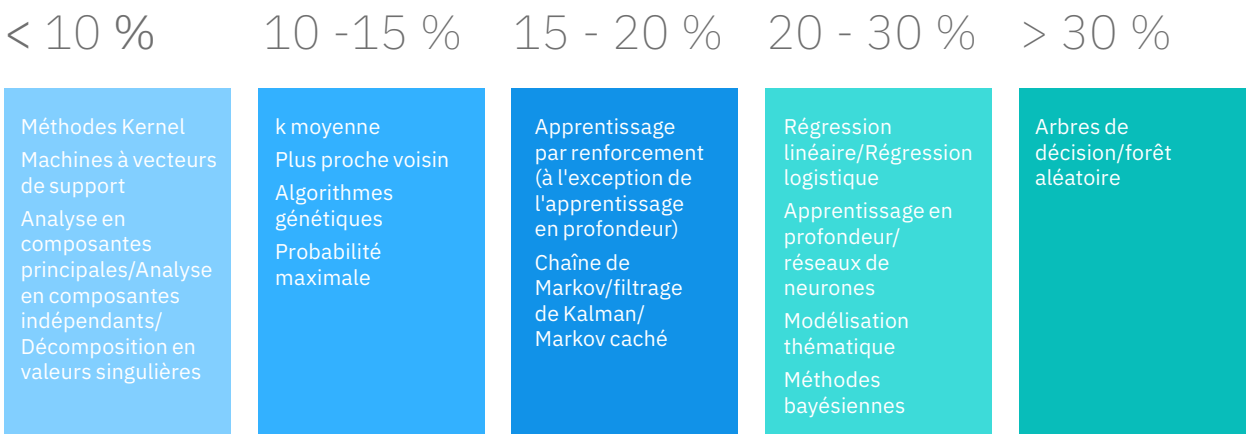
D'un point de vue conceptuel, les inquiétudes concernant les coûts de l'apprentissage en profondeur peuvent poser des problèmes importants quant à l'orientation future et à la nature de la recherche en IA (voir « Le coût entraînera-t-il la disparition de l'apprentissage en profondeur ? ») à la page 11). D'une manière pragmatique, c'est en comparant les résultats obtenus dans la pratique, généralement dans le cadre d'une validation de concept ou d'un projet pilote, qu'il est possible d'identifier les lacunes de l'apprentissage en profondeur.

FIGURE 3

L'apprentissage en profondeur n'est pas universel

Les entreprises utilisent différentes techniques d'apprentissage automatique en fonction du problème métier.

Pourcentage d'entreprises utilisant chaque technique d'apprentissage automatique



Source : 2021 IBV AI Capability Survey, données non publiées Q16A. What Machine Learning (ML) techniques does your organization employ? Choose all that apply.

KPMG utilise l'apprentissage en profondeur et d'autres outils d'analyse pour aider ses clients à économiser des millions USD.

Le cabinet mondial, KPMG, spécalisé dans le conseil, les audits et la fiscalité, a lancé un hackathon interne pour identifier la meilleure approche possible pour réduire le travail manuel lié à la documentation des projets de R&D, des investissements et des crédits d'impôt des clients. Cette approche peut générer une valeur métier tangible en réduisant les factures fiscales des clients chaque année. Le cabinet a constaté que la précision variait de 55 % en utilisant un logiciel de reconnaissance de documents prêt à l'emploi (à peu près aussi efficace qu'une recherche manuelle par mot-clé) à plus de 70 % pour l'apprentissage en profondeur.

Mais, il apparaît que les meilleures approches sont l'apprentissage automatique basé sur des règles, offrant une précision supérieure à 85 %. L'automatisation de ces processus se révèle être un moyen plus rentable d'économiser des millions USD en dépenses fiscales pour un client chaque année. Un client a bénéficié d'un crédit d'impôt supplémentaire de 40 % sur ses dépenses de R&D en adoptant cette approche.

Perspective

Le coût signera-t-il la fin de l'apprentissage en profondeur ?

Si les réseaux de neurones artificiels existent depuis les années 1950, après avoir survécu à l'arrêt des investissements et au désintérêt pendant les hivers de l'IA[6], l'apprentissage en profondeur rayonne depuis la fin des années 2000.

Une augmentation massive de la puissance informatique pour traiter les données, associée à l'explosion rapide des données structurées et non structurées, a stimulé la phase la plus récente.

Face à la croissance exponentielle et continue des données, conjuguée à la loi de Moore qui approche de sa limite (si ce n'est déjà le cas), des chercheurs en IA s'inquiètent des coûts financiers et environnementaux nécessaires pour maintenir cette tendance. Comme Neil C. Thompson l'indique dans un article du IEEE Spectrum de 2021 : « Il est clair que vous pouvez améliorer les performances de l'apprentissage en profondeur si vous utilisez davantage de puissance de calcul pour créer des modèles plus grands et les entraîner avec davantage de données. Mais quel sera le coût de cette charge de calcul ? Les coûts atteindront-ils un niveau propre à entraver le progrès ? » [7]

Par exemple, le développement et l'entraînement de GPT-3 d'OpenAI a coûté 3 millions USD, et le seul entraînement d'AlphaGo de DeepMind, filiale d'Alphabet, aurait coûté 35 millions USD.

Avec de tels coûts (qui par ailleurs, augmentent rapidement), le problème se complique : comment concilier la nécessité de disposer de modèles plus grands, de plus de données, de plus d'entraînement et de plus de puissance de calcul, avec les réalités métier inhérentes aux budgets et à l'efficacité ? Les chercheurs devront résoudre ce problème au risque de voir les avancées ralentir.[8]

Divers organismes de recherche explorent des moyens de s'adapter : différentes solutions matérielles, nouvelles méthodes d'apprentissage IA et nouvelles façons de combiner un apprentissage en profondeur puissant, riche en données et en paramètres, avec des techniques symboliques classiques basées sur le raisonnement et des règles.

Un article d'IEEE Spectrum le résume comme suit : « Si l'avènement de l'apprentissage en profondeur a pu être fulgurant, son futur peut être chaotique face à la multiplication de ces efforts de recherche. »[9]

Dans l'intervalle, les organisations doivent surveiller attentivement leur compromis entre le coût et les performances dans l'utilisation de l'apprentissage en profondeur, notamment par rapport à d'autres outils d'IA.

Zzapp Malaria : utilisation de l'IA pour le bien du monde, et pas seulement pour les affaires

Le paludisme a été à l'origine d'environ 627 000 décès en 2020, dont 96 % en Afrique.[10] Zzapp Malaria, lauréat 2021 de la XPRIZE AI, crée des approches optimisées par l'IA pour éliminer le paludisme et les applique directement sur le terrain par le biais d'une application mobile dédiée.

Dans le cadre d'un projet pilote, le réseau de neurones convolutifs de Zzapp Malaria a analysé des images visuelles pour détecter de petites étendues d'eau, des lieux de reproduction potentiels des moustiques, vecteurs du paludisme, que ne révélaient pas distinctement les images satellitaires existantes. Il a atteint une précision d'environ 75 %, mais avec une visibilité limitée sur les facteurs qui sous-tendent les prévisions. Malgré les bons résultats obtenus, ils étaient insuffisants pour être adaptés à d'autres sites.

À l'aide du réseau CNN, l'équipe extraya, entre autres, 50 caractéristiques topographiques des images, puis les utilisa dans une approche traditionnelle reposant sur la régression linéaire pour déterminer la probabilité de présence d'eau stagnante. Les résultats étaient équivalents à ceux obtenus précédemment, mais ils mettaient beaucoup plus en évidence les facteurs à l'origine des prévisions. Les résultats étant plus simples à expliquer à l'équipe, cette approche pouvait être donc être plus applicable à des sites présentant une topologie très différente. L'équipe s'est appuyée sur ce succès de l'IA pour orienter les ajustements de son approche et étendre sa couverture, afin de réduire l'incidence du paludisme dans d'autres régions.

Mythe 1
L'IA est universelle

Mythe 2
Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA

**Mythe 3 -
La réduction des coûts est le point fort de l'IA**

Mythe 4
Avec l'IA, pas de raccourcis possibles

Mythe 5
L'IA n'apporte de valeur ajoutée qu'au niveau du problème traité

Annexe

Mythe 3

La réduction des coûts est le point fort de l'IA

Réalité

Si l'utilisation de l'IA pour résoudre des problématiques métier peut effectivement réduire les coûts, elle offre bien d'autres atouts. Les organisations leaders s'efforcent activement (et stratégiquement) de l'exploiter pour se démarquer de la concurrence, améliorant l'efficacité des processus générateurs de chiffre d'affaires, favorisant la croissance et innovant dans le modèle de gestion dans le même temps.

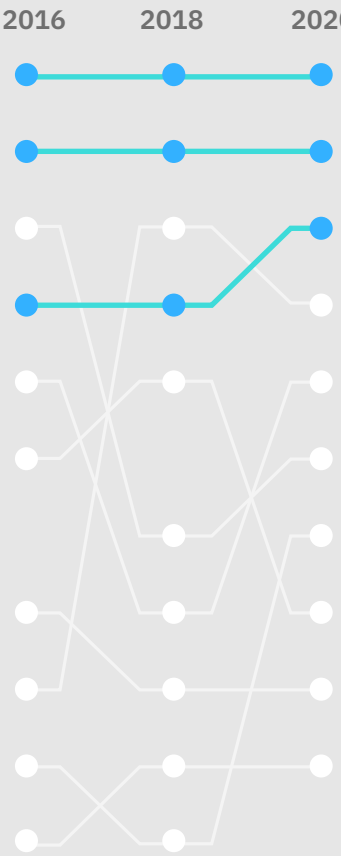
Le coût est important, mais la croissance, l'innovation et le bien-être social comptent davantage. Une enquête IBV montre de manière constante que les entreprises accordent la priorité à la croissance au centre de laquelle se trouve le client, car il s'agit du domaine dans lequel l'IA a le plus fort impact (voir la figure 4 à la page 15).

On se demandait si cette priorité ne relevait pas plus du vœu pieux que de la réalité, mais nos entretiens avec des responsables lors de cette enquête ont révélé que des entreprises ont joint le geste à la parole.

FIGURE 4

Facteurs de valeur de l'IA 2016–2020

Les entreprises accordent la priorité à la croissance du chiffre d'affaires, au centre de laquelle se trouve le client.



Facteurs de valeur de l'IA

Amélioration de la satisfaction client

Amélioration de la rétention client

Réduction des coûts d'acquisition de clients

Réduction d'autres coûts opérationnels

Croissance du revenu avec un cycle de vente plus court

Croissance du revenu avec des commandes plus importantes

Redéploiement des effectifs

Croissance du revenu par l'accélération de la mise sur le marché

Réduction des effectifs

Réduction des autres coûts d'investissement

Source : Voir la note 11 en fin de document.

L'IA favorise la croissance du chiffre d'affaires dans une co-entreprise d'assurance

IFFCO-Tokio, une co-entreprise indienne d'assurance générale, a décidé d'améliorer l'expérience des clients en leur réglant directement le coût des réparations après l'approbation de leur demande de remboursement.

La première étape consiste à capturer des images du véhicule impliqué dans un accident. Ensuite, les équipes utilisent l'apprentissage en profondeur pour classifier le modèle de voiture, les pièces endommagées et le type de dommage. Ainsi, le système d'IA peut déterminer si les pièces peuvent être réparées ou doivent être remplacées, et fournir une estimation des coûts, tout en maintenant un expert humain dans la boucle pour réduire le risque de fraude.

C'est une grande réussite, car le projet a été rentabilisé en moins d'un an. Les règlements ont baissé de 40 % et le taux d'acceptation client est passé de 30 % à 65 %. S'en est suivi une augmentation de la satisfaction, de la rétention et de l'acquisition client. L'IA ne permet pas seulement d'améliorer l'efficacité, mais elle est aussi un facteur indéniable de croissance du chiffre d'affaires.

Mythe 1 :
L'IA est universelle

Mythe 2
Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA

Mythe 3 :
La réduction des coûts est le point fort de l'IA

Mythe 4 :
Avec IA, pas de raccourcis possibles

Mythe 5
L'IA n'apporte de valeur ajoutée qu'au niveau du problème traité

Annexe

Mythe 4

Avec l'IA, pas de raccourcis possibles

Réalité

Alors que les cas utilisation des modèles d'IA varient selon le secteur et la fonction, un ensemble croissant de modèles de base et de modèles pré-entraînés « prêts à l'emploi » peut constituer un point de départ plus abordable pour les scientifiques des données d'entreprise.

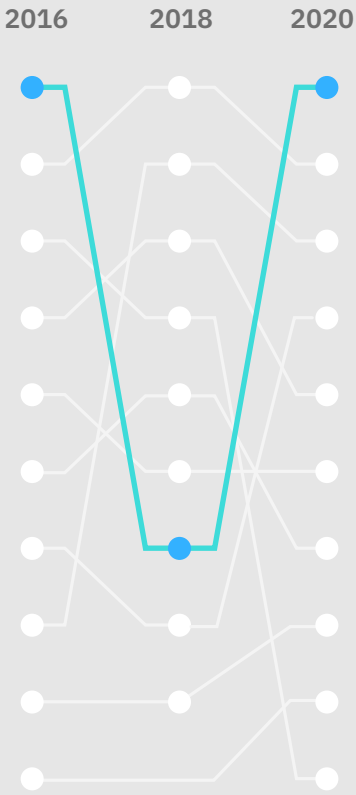
L'IBM Institute for Business Value a adopté une approche systématique pour quantifier diverses tendances dans l'IA d'entreprise depuis 2016. L'une des surprises en 2020, a été la réémergence de la « disponibilité de la technologie » comme barrière à l'adoption de l'IA, qui avait cédé la place aux « compétences et autres facteurs » en 2018 (voir la figure 5 à page 17). Nous nous sommes demandés les raisons pour lesquelles cet obstacle réapparaissait.

Nous continuons de penser que les organisations prennent enfin conscience que ce qu'elles pensaient être nécessaire pour faire fonctionner la technologie IA, des employés dotés des bonnes compétences en données, est indispensable, mais insuffisant en soi. Les nombreux scientifiques des données embauchés par les entreprises pour entraîner différents jeux de données ont consciencieusement fait ce qu'on leur demandait. Mais il semblait que pour résoudre chaque problématique métier, généralement avec un modèle d'IA différent du précédent, il fallait repartir de zéro. Il n'y avait pas de moyen simple de tirer parti des expériences précédentes.

FIGURE 5

Obstacles à l'IA 2016–2020

La disponibilité de la technologie a réémergé comme le principal obstacle à l'implémentation de l'IA en 2020.



Obstacles à l'IA

Disponibilité de la technologie

Disponibilité de ressources qualifiées ou d'expertise technique

Contraintes réglementaires

Niveau de soutien de la direction

Problèmes juridiques/de sécurité/de confidentialité liés à l'utilisation des données et de l'information

Gouvernance des données et politiques de partage au-delà des limites de l'entreprise et avec des partenaires externes

Volume/disponibilité des données à appliquer et mise en contexte pour la prise de décision

Niveau de préparation des clients

Degré de préparation des partenaires ou des parties prenantes

Degré d'adhésion/de préparation/d'adaptation culturelle dans l'entreprise

Sources: "Shifting toward Enterprise-grade AI: Confronting skills and data challenges to realize value." IBM Institute for Business Value. Septembre 2018. https://www.ibm.com/thought-leadership/institutebusiness-value/report/enterpriseai. Figure 1, Barriers in implementing AI: 2016 versus 2018, The business value of AI, unpublished data. Q9. What are the top barriers your organization faces in implementing artificial intelligence? Sélectionnez les 5 principaux obstacles.

Depuis quelques temps, des entreprises commencent à utiliser des raccourcis IA pour tirer parti de leurs solutions IA. Le « prêt à l'emploi » sont au logiciel ce que les modèles de base et pré-entraînés sont à l'IA. Ils peuvent constituer un point de départ plus efficace pour les nouveaux projets IA.

Comment ? En aidant les organisations à progresser sans générer de nouveaux ensembles de données, plutôt qu'en tirant parti des connaissances IA acquises lors de la résolution d'un problème pour résoudre des problèmes connexes. La clé de cette approche est l'apprentissage par transfert : réutiliser un modèle entraîné à l'origine pour une tâche et l'appliquer à une autre. Par exemple, un modèle de reconnaissance de véhicules peut être appliqué à la reconnaissance de camions.

La plupart des types de modèles pré-entraînés sont conçus pour résoudre un ou plusieurs problèmes métier spécifiques. Il existe un nombre croissant de modèles généralisés et gigantesques (par exemple, Alphabet's BERT, GPT-3 d'OpenAI) qui peuvent être utilisés pour relever de nombreux défis. Cela dit, ils peuvent avoir été dépassés par le Wu Dao 2.0 chinois, le premier modèle aux mille milliards de paramètres au monde.

Les modèles de base peuvent apporter de la valeur ajoutée de trois manières essentiellement :

- *Amélioration de la rentabilité :* amortissement des coûts sur plusieurs cas d'utilisation
- *Amélioration des résultats* : plus grande exactitude issue de jeux de données plus grands et robustes
- *Nouvelles fonctionnalités* : capacité à rassembler des données multimodales plus efficacement.

Mais ce n'est pas toujours vrai. L'adaptation de modèles pré-entraînés entraîne parfois une baisse trop importante des performances sur les nouvelles données. C'est précisément le problème auquel Boston Scientific, un fabricant américain d'appareils médicaux, a été confronté et qu'il a résolu.

Boston Scientific dépense 50 000 USD pour économiser 5 millions USD

Boston Scientific souhaitait automatiser son processus d'inspection d'endoprothèses pour améliorer la précision lors de la recherche de défauts tels que des liens brisés ou des imperfections de surface. Des inspections précises sont essentielles pour obtenir des résultats cliniques satisfaisants. La Food and Drug Administration des États-Unis réglemente les « taux de fuite » (la proportion de pièces défectueuses qui peuvent passer entre les mailles du filet) en fonction du risque pour les patients.

« L'inspection visuelle humaine est souvent lente, coûteuse et peut présenter des risques de qualité indésirables », souligne Eric Wespi, le responsable des sciences des données chez Boston Scientific. La société compte environ 3 000 experts qui effectuent des inspections, soit un coût de plusieurs millions USD par an.

Boston Scientific avait déjà implémenté un système automatisé basé sur des règles qui utilisait des mesures dimensionnelles et d'autres moyens pour capturer les problèmes courants. Ce système avait été réglé pour produire des résultats conservateurs, avec un taux de faux négatifs négligeable. Cependant, le taux de faux positifs de 5 % à 10 % était encore trop élevé. Trop de pièces conformes aux spécifications étaient signalées comme défectueuses.

Capables d'analyser l'imagerie visuelle, les réseaux de neurones convolutifs seraient parfaitement adaptés pour résoudre ce problème, mais de tels modèles nécessitent d'énormes volumes de données. L'équipe ne disposait pas de suffisamment de données pour entraîner intégralement ces modèles. Il lui est également apparu que la collecte ou la génération de ces données serait peu pratique et son coût prohibitif.

La solution ? Elle a commencé par segmenter le problème en tâches plus petites et plus limitées. Ensuite, elle a exploité des modèles IA open source prêts à l'emploi existants pour relever le défi redéfini. Enfin, elle a utilisé un jeu de données plus petit pour affiner ce système.

Le résultat ? L'entreprise a réalisé 5 millions USD d'économies directes sur la base d'un budget modeste d'environ 50 000 USD, et atteint une précision supérieure à celle obtenue précédemment.

Les chefs d'entreprise qui veulent gagner du temps et de l'argent avec des modèles de base et pré-entraînés doivent ne pas pas perdre de vue que s'ils permettent de faire des économies, ils peuvent ne pas être la meilleure option s'ils visent principalement la différenciation. Ces modèles étant à la disposition de tous, certains étant open source, les entreprises doivent veiller à choisir des problémes métier où la différenciation importe moins. Ou bien, elles peuvent se recentrer sur la personnalisation en incluant des données supplémentaires, généralement propriétaires ou intégrées propriétaires, pour accroître l'avantage concurrentiel.

Les organisations prennent enfin conscience que ce qu'elles pensaient être nécessaire pour faire fonctionner la technologie IA, des employés dotés des bonnes compétences en données, est indispensable, mais insuffisant en soi.

Mythe 1
L'IA est universelle

Mythe 2
Si ce n'est pas de l'apprentissage en profondeur, ce n'est pas de l'IA

Mythe 3
La réduction des coûts est le point fort de l'IA

Mythe 4
Avec l'IA, pas de raccourcis possibles

Mythe 5
L'IA n'apporte de valeur ajoutée qu'au niveau du problème traité

Annexe

Mythe 5

L'IA n'apporte de la valeur ajoutée qu'au niveau du problème traité

Réalité

Les effets des réseaux IA intra et inter-entreprise émergents apportent une vraie valeur métier au sein de l'entreprise.

La prolifération des sources de données et la possibilité accrue de les exploiter fournissent aux entreprises une masse croissante de données. Stratégiquement utilisées pour alimenter une IA réfléchie et éthique, les entreprises récoltent non seulement des bénéfices financiers, mais favorisent aussi l'essor de l'innovation ouverte. Ainsi, il en résulte des améliorations d'échelle supplémentaires, notamment dans les entreprises plus avancées ayant adopté l'IA. Comme nous le notions en 2020 dans notre rapport sur la valeur métier de l'IA :

« Les effets des réseaux, même s'ils ne sont qu'internes à l'entreprise, semblent étendre davantage encore les bénéfices des investissements en IA. Une analyse initiale suggère qu'investir dans l'IA dans un domaine d'opérations métier a tendance à amplifier l'adaptabilité et la résilience organisationnelles dans d'autres domaines, avec les gains financiers qui en découlent. Par exemple, l'amélioration de la gouvernance des données et des politiques d'accès dans une fonction s'étend aux fonctions adjacentes dans le cadre de leur intégration et de leur collaboration dans un flux de travail. Cette constatation se révèle spécialement vraie pour les investissements IA dans les fonctions centrales ou dorsales qui ont une influence ou des impacts transversaux particulièrement forts, comme dans la finance, l'informatique ou les RH. »[12]

Par exemple, la rotation des talents IA d'un service ou d'un projet à un autre permet une pollinisation croisée d'expertise, ainsi qu'un apprentissage organisationnel continu pour les collaborateurs, au sein de l'entreprise. Ainsi, il est possible de développer l'ensemble des compétences en IA au lieu de les laisser stagner.

Les effets des réseaux et autres synergies entre l'IA et d'autres technologies de transformation numérique, comme le cloud, l'Internet des objets, la sécurité et la gestion des données, viennent s'ajouter à la valeur réalisable.[13]

Comme c'est la cas pour de nombreuses technologies émergentes, nous constatons déjà que ce que l'IA catalysait déjà au sein d'institutions a également commencé à se manifester dans toutes les institutions.

NVIDIA encourage l'innovation ouverte sur le marché de l'automobile

L'approche de la société technologique NVIDIA de l'innovation concernant les modèles de gestion met en lumière la façon dont l'IA peut se propager aux clients et aux partenaires commerciaux. Pour relever l'énorme défi du calcul pour les véhicules autonomes, que certains constructeurs ne peuvent pas développer en raison de leur manque d'expérience, de matériel et de données, l'entreprise crée un ensemble partagé de fonctionnalités d'IA :

- Plateforme de données commune à plusieurs clients
- Simulation pour l'entraînement et les tests
- Traitement commun des tâches visuelles.

En fonction de leurs besoins et de leurs fonctionnalités existantes, les constructeurs automobiles participants peuvent soit louer du matériel pour véhicule autonome, afin d'entraîner leurs propres modèles sur la base d'un jeu de données plus important, soit utiliser des modèles pré-entraînés de NVIDIA. Dans les deux cas, au lieu d'investir massivement dans le matériel et les fonctionnalités de développement de l'IA, les constructeurs automobiles peuvent comptabiliser la technologie dans les dépenses d'exploitation et bénéficier d'améliorations à mesure que le matériel et les logiciels s'améliorent.

Ce que nous voyons dans les institutions dotées de l'IA indique, en fin de compte, des forces qui pourraient avoir un impact sur des économies entières.

Un grand nombre des plateformes B2C leaders, ainsi que des fabricants de matériel/logiciels, des institutions universitaires et des gouvernements, ont investi des sommes considérables pour faire progresser la recherche dans l'IA, les mettant souvent à disposition dans le domaine public. Ces plateformes ont également développé de manière constante les compétences transférables au niveau de leur direction et de leurs travailleurs du savoir (des agents libres, tous), principalement à partir de leur expérience directe à l'avant-garde de l'adoption de l'IA.

L'énergie potentielle d'une plus grande mobilité professionnelle, accélérée par les changements dynamiques de l'espace de travail pendant la pandémie, est libérée par la « Grande Démission » en énergie cinétique qui peut transformer les économies mondiales.

Qu'est-ce qui fait obstacle ?

La diffusion des connaissances, c'est-à-dire, la distribution des connaissances et des talents, n'est pas toujours en parfaite adéquation avec la « capacité d'absorption », à savoir la capacité des organisations à s'adapter et à intégrer ces informations et ces compétences.[14] Des barrières institutionnelles peuvent se dresser sur le chemin, de même qu'une gestion rigide qui résiste à l'inconfort que peut engendrer le changement.

La valeur transformatrice de l'IA, à travers ses impacts financiers, économiques et sociétaux, ne peut se concrétiser que si les dirigeants des entreprises plus traditionnelles mettent de côté leurs idées nostalgiques sur le fonctionnement passé. Ils doivent pleinement saisir les opportunités d'innovation de manière stratégique, réfléchie et concrète.

Un point de départ essentiel consiste à distinguer la perception de l'IA de sa réalité émergente.

Autres documents à consulter

Depuis la fin 2020, l'IBV développe des documents sur la création de fonctionnalités d'IA de premier rang. Il adopte une vision globale de l'IA à l'échelle de l'entreprise et lie de nombreux thèmes pertinents pour tirer profit de l'adoption de l'IA sur le plan financier et la rentabilité.

Chacun de ces documents offre un ensemble de recommandations concrètes relatives à un thème spécifique, qui sont également adaptées aux entreprises qui ont atteint un niveau de maturité plus ou moins élevé dans leur adoption de pratiques métier IA.

Nous vous suggérons de consulter ces guides d'actions tangibles qui synthétisent des projets IA et l'expertise de centaines de professionnels IA et d'autres experts, que vous trouverez dans chacun des rapports suivants :

- *Strategy and vision:* Rethinking your approach to AI
- *Donnnées et technologies :* Faire face au dilemme des données de l'IA
- *Engénierie et opérations:* Concepts éprouvés pour la mise à l'échelle de l'IA

À propos des auteurs

Nicholas Borge

Chercheur, FutureTech, MIT Computer Science and AI Lab
njborge@mit.edu
linkedin.com/in/nicholasborge

Nicholas Borge est membre de l'équipe du projet FutureTech du MIT, où il participe à des recherches sur la rentabilité de l'IA et le futur du travail. Nick a été directeur de l'automatisation intelligente chez Sony Music, a fondé une start-up d'IA et a plus de 11 ans d'expérience dans le conseil en stratégie et en technologie pour des entreprises du Fortune 500. Nick est titulaire d'un master du MIT en ingénierie et gestion ; il est Fellow du programme System Design and Management du MIT.

Subhro Das, titulaire d'un doctorat

Membre du personnel de recherche, MIT-IBM Watson AI Lab
subhro.das@ibm.com
linkedin.com/in/subhrodas/

Subhro Das est membre de l'équipe de recherche du MIT-IBM Watson AI Lab au sein d'IBM Research. En tant que chercheur principal du laboratoire, il travaille au développement de nouveaux algorithmes d'IA en collaboration avec le MIT. Ses recherches portent sur les méthodes d'optimisation pour l'apprentissage automatique, l'apprentissage par renforcement, l'apprentissage automatique digne de confiance et les algorithmes d'IA centrés sur l'humain. Il est titulaire d'un master et d'un doctorat en Electrical and Computer Engineering de la Carnegie Mellon University.

Martin Fleming, titulaire d'un doctorat

Chief Revenue Scientist, Varicent
martin@fleming41.com
linkedin.com/in/flemingmartin

Martin Fleming est Chief Revenue Scientist chez Varicent, un fournisseur de logiciels de gestion de la performance commerciale à Toronto. Martin est également chargé de recherche au Productivity Institute, un consortium de huit universités britanniques. Ses recherches se situent à l'intersection de la technologie, de la productivité et de l'économie. Il est l'auteur de la publication à paraître, « Breakthrough, A Growth Revolution ». Auparavant, Martin a été économiste en chef et directeur principal de l'analyse chez IBM.

À propos des auteurs

Brian Goehring

Responsable mondial de la recherche, IA,
IBM Institute for Business Value
goehring@us.ibm.com
linkedin.com/in/brian-c-goehring-9b5a453/

Brian Goehring est Partenaire associé à l'IBM Institute for Business Value, où il dirige le programme de recherche métier IA, en collaboration avec des universitaires, des clients et d'autres experts pour développer un leadership d'opinion basée sur les données. Il apporte plus de 20 ans d'expérience dans le conseil stratégique auprès de dirigeants dans la plupart des secteurs et fonctions métier. Il a obtenu un AB en philosophie de l'Université de Princeton et possède des certificats en études cognitives et en allemand.

Neil Thompson, titulaire d'un doctorat

Directeur, FutureTech, MIT Computer Science and AI Lab
neil_t@mit.edu
linkedin.com/in/neil-thompson-5724a614

Neil Thompson est directeur du projet de recherche FutureTech au Computer Science and Artificial Intelligence Lab du MIT et chercheur principal dans le cadre de l'Initiative on the Digital Economy du MIT. Il a été professeur adjoint en innovation et stratégie à la MIT Sloan School of Management et professeur invité au Laboratory for Innovation Science d'Harvard. Il a travaillé dans des organisations telles que le Lawrence Livermore National Laboratory, Bain & Company, les Nations Unies, la Banque mondiale et le Parlement canadien. Il est titulaire d'un doctorat en affaires et politiques publiques et d'une maîtrise en informatique et statistiques de l'Université de Californie de Berkeley, ainsi que d'une maîtrise en économie de la London School of Economics.

Contributeurs

Adam Bogue
Responsable du développement commercial,
IBM Research

Alex Gorman
Directeur de programme, Client Advocacy,
IBM Software

Cathy Reese
Associée principale, Practice Leader,
IBM Consulting

Shannon Todd-Olson
Associée principale, IBM Consulting

Remerciements

Les auteurs et contributeurs remercient le MIT-IBM Watson AI Lab, et ses codirecteurs Aude Oliva et David Cox, pour le financement de ce projet, ainsi que Seth Dobrin, Glenn Finch et Sriram Raghavan pour leur soutien.

Annexe

Périmètre et échelle des études de cas détaillées

Études de cas dans différents secteurs, sur différentes fonctions et sur différentes techniques d'apprentissage automatique.

Entreprises interrogées

Nom	Secteur d'activité	Fonction	Solution métier	Technique d'apprentissage automatique
BESTSELLER	Consommation	Création de mode	Amélioration de la conception et de l'efficacité des ventes par l'extraction des attributs des produits à partir des images du catalogue	Vision
Boston Scientific	Industrie	Conception d'appareils médicaux	Réduction du coût de la main-d'oeuvre par l'automatisation de l'inspection visuelle d'endoprothèses en utilisant l'apprentissage par transfert	Vision
Crédit Mutuel	Secteur bancaire	Service client	Appels plus courts en suggérant de meilleures réponses aux conseillers client grâce au traitement automatique du langage naturel hiérarchique	Langage
Global Bank	Secteur bancaire	Audit interne	Capacité d'audit accrue par l'amélioration de la qualité de la documentation à l'aide d'un « relecteur instantané »	Langage
IFFCO-Tokio	Assurance	Automatisation des demandes	Réduction des versements des indemnisations d'assurance par paiement direct aux demandeurs en utilisant des évaluations automatisées	Vision
KPMG	Services professionnels	Crédits d'impôt	Augmentation des crédits d'impôt en offrant une meilleure documentation par la recherche de documents	Langage
Marketing Platform	Services professionnels	Ciblage publicitaire	Contrôle des coûts d'entraînement de modèles de ciblage en exposant les incitations marginales des expérimentations	Autres
McCormick	Consommation	R&D/conception de produit	Efficacité R&D accrue en suggérant des profils d'arômes initiaux pour l'expérimentation	Autres
Navtech	Technologie de l'information	Ventes	Accès à des catalogues de produits numériques en créant une plateforme de vision par ordinateur	Vision
NVIDIA	Technologie de l'information	Conduite autonome	Mise en commun de données et offre d'une technologie VA en tant que service pour disposer d'un nouveau modèle de gestion	Vision
Suncor	Énergie	Exploitation des sites	Développement d'un système d'avertissement en amont pour les problèmes de production de diesel par la prévision de conditions de traitement défavorables	Autres
Zzapp	Technologie de l'information	Santé publique	Utilisation de l'imagerie satellite pour identifier l'eau stagnante pour le traitement par insecticide antipaludéen	Vision

Source: Neil Thompson, PhD: http://www.neil-t.com/teaching-cases/

Annexe

BESTSELLER

Boston Scientific

Crédit Mutuel

Global Bank

IFFCO-Tokio

KPMG

Marketing Platform

McCormick

Navtech

NVIDIA

Suncor

Zzapp

BESTSELLER

Utilisation de l'IA pour libérer toute la valeur des données de votre entreprise

Résumé

La prévision de la demande consiste à fournir des caractéristiques de produits à des algorithmes. Plus la quantité d'informations est importante, plus il est aisé de déterminer la variabilité des modèles de demande historiques et meilleures sont les prévisions futures.

Cependant, l'évaluation d'attributs granulaires supplémentaires de produits peut être difficile et chronophage. L'apprentissage en profondeur apporte une solution en classifiant rapidement et précisément les produits avec une intervention manuelle minime, augmentant, ainsi le nombre de caractéristiques disponibles pour les algorithmes de prévision. L'entreprise du secteur de la mode BESTSELLER illustre ce ce fonctionnement.

Opportunité : Réduction du gaspillage et amélioration des délais d'exécution

Dans le secteur de la mode, environ 80 % des marchandises se vendent sur deux saisons chaque année, tout le reste étant vendu à prix très réduit, donné, voire jeté. Cette surproduction se traduit par des pertes de profits et pose également un énorme problème de durabilité.

BESTSELLER conçoit, fabrique et vend des vêtements pour le marché indien. Pour chacune de ses quatre marques, son équipe conçoit et modélise 3 500 échantillons, mais n'en sélectionne que 1 100 pour la production. Les vêtements retenus sont classés dans 5 000 à 6 000 références (SKU) par couleur et taille, notamment. L'entreprise produit 1,5 million de pièces. BESTSELLER peut vendre environ 78 % de sa production, une performance relativement bonne dans le secteur de la mode. Mais il est possible d'augmenter encore ce pourcentage en adaptant plus efficacement la production aux préférences client. Alors qu'elle prévoit de plus que doubler son portefeuille de marques en le faisant passer de quatre à neuf d'ici la fin de l'année, l'amélioration du pourcentage de vente peut avoir un impact considérable sur la rentabilité.

Défi : Insuffisance d'éléments de conception disponibles pour l'analyse

BESTSELLER a voulu identifier les facteurs qui déterminent les ventes d'un produit particulier. Ainsi, elle pourrait tenir compte des informations dans le processus de conception, afin d'améliorer les ventes, en alignant plus clairement le nombre de produits vendus sur celui des produits fabriqués. Cette initiative aurait également pour avantage d'améliorer potentiellement la conception. Cependant, une première analyse utilisant des données sur les attributs des produits, tels que la couleur et la taille, ainsi que le stockage et l'emplacement, révéla que la quantité de données sur les produits était tout simplement insuffisante pour créer des inférences significatives. Un ensemble de données plus riche était nécessaire.

Les vêtements peuvent être décrits selon leur forme, leur coupe, le tissus, les styles et divers éléments de conception. En fait, BESTSELLER utilisait une taxonomie de plus de 7 000 patterns de conception et 4 000 couleurs. Si un grand nombre de ces caractéristiques sont discernables par la simple observation des images des produits, très peu de ces informations étaient balisées dans les données maître des produits. BESTSELLER devait pouvoir extraire ces informations rapidement et efficacement.

Solution : Analyse des images pour enrichir les caractéristiques disponibles

Elle a donc extrait des caractéristiques supplémentaires des images en utilisant la vision par ordinateur. BESTSELLER a pris 10 000 images (un catalogue d'une saison) et développé un modèle pour chacune de ses quatre marques. En seulement trois semaines, elle a pu développer et entraîner un réseau de neurones convolutifs pour classifier une image selon diverses caractéristiques. Ces fonctions dérivées de l'apprentissage en profondeur pourraient être introduites dans des techniques d'analyse traditionnelles, telles que la régression ou l'analyse en composantes principales, afin de mieux comprendre les facteurs qui développent les ventes.

Résultats : Amélioration de l'échantillonnage de conception et des ventes

Même avec la baisse globale des ventes due à la pandémie, BESTSELLER a enregistré des améliorations remarquables tant au niveau des ventes et de l'efficacité de conception au cours des 18 mois suivants. Les ventes ont atteint 82 % (plus quatre points de pourcentage par rapport à 78 %), et l'entreprise a réduit de 15 % le nombre d'échantillons créés pour chaque marque, sans réduire le nombre final de conceptions sélectionnées. L'efficacité de l'échantillonnage s'est accrue en permettant aux concepteurs de se limiter à un plus petit nombre de modèles ayant une plus grande probabilité de vente.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Boston Scientific

Éviter les pièges de l'apprentissage par transfert

Résumé

L'apprentissage par transfert consiste à réutiliser un modèle, entraîné à l'origine pour une tâche, pour une autre tâche. Ainsi, les connaissances acquises lors de la résolution d'un problème métier peuvent être appliquées à un problème connexe ; par exemple, un modèle de reconnaissance de véhicules peut être appliqué à la reconnaissance de camions.

L'apprentissage par transfert permet de réduire la quantité de travail et les coûts d'entraînement, mais il peut également s'accompagner d'une énorme chute des performances (jusqu'à 45 %). Par conséquent, il n'est donc pertinent que dans des cas limités. L'expérience de Boston Scientific, cependant, montre qu'une organisation peut tout de même obtenir des performances élevées avec l'apprentissage par transfert, en « réduisant » le problème, permettant ainsi à son modèle d'atteindre des niveaux de performance de plus de 99 % et de diminuer les coûts de main-d'oeuvre de plus de 5 millions USD.

Opportunité : Inspection coûteuse d'endoprothèses essentielle à la sécurité des patients

Boston Scientific fabrique des endoprothèses pour une gamme d'applications chirurgicales. Les équipes doivent les inspecter pour identifier les défauts éventuels tels que des liens brisés ou des imperfections de surface. Des inspections précises sont essentielles pour obtenir des résultats cliniques satisfaisants. C'est la raison pour laquelle, le « taux de fuite » (la proportion de pièces défectueuses susceptibles de passer entre les mailles du filet) est réglementée par la Food and Drug Administration des États-Unis en fonction du risque pour les patients.

Traditionnellement, les inspections sont réalisées par des experts humains, mais les résultats ne sont pas optimaux. « L'inspection visuelle humaine est souvent lente, coûteuse et peut présenter des risques de qualité indésirables », indique Eric Wespi, le responsable en science des données chez Boston Scientific : Intuitivement, cela se comprend : les gens ne sont généralement pas performants dans les tâches qui nécessitent une attention soutenue pendant de longues périodes, lorsque la probabilité d'un événement est peu fréquente. De plus, l'appréciation peut varier d'une personne à l'autre. En outre, les experts coûtent cher. Chez Boston Scientific, 3 000 experts environ effectuent des inspections pour un coût qui s'élève à plusieurs millions USD chaque année.

Défi : La classification des images nécessite une grande quantité de données pour l'entraînement

Boston Scientific avait déjà implémenté un système automatisé basé sur des règles qui utilisait des mesures dimensionnelles et d'autres moyens pour capturer les problèmes courants. L'équipe avait réglé le système pour produire des résultats conservateurs, avec un taux de faux négatifs négligeable. Cependant, le taux de faux positifs de 5 % à 10 % était encore trop élevé. Trop de pièces conformes aux spécifications étaient signalées comme défectueuses par les inspecteurs humains.

Les réseaux de neurones convolutifs (CNN) sont particulièrement bien adaptés à la classification d'images, mais l'entraînement de ces modèles nécessitent une énorme quantité de données pour les entraîner. Dans de nombreux cas (notamment pour des défauts nouveaux et plus rares), l'équipe ne disposait pas de suffisamment de données pour entraîner intégralement ces modèles. La collecte ou la production de ces données n'auraient pas été pratiques et le coût aurait été prohibitif.

Solution : L'apprentissage par transfert appliqué à un problème réduit

L'équipe s'est demandé s'il ne serait pas plus efficace de partir d'un modèle pré-entraîné. Elle a adopté l'approche suivante :

1. Réduction de l'échelle du problème : pour chaque défaut, l'inspection pouvait être segmentée en tâches plus petites telles que « Cette partie de l'image contient-elle un lien ? » et « Ce lien est-il rompu ou non ? »
2. *Personnalisation des modèles existants* : plusieurs réseaux CNN open source ont été utilisés (par exemple, VGG16, EfficientNet [B0 à B7], Mask R-CNN, YOLOv3, ResNet-50 et Inception-v3). Dans chaque cas, l'équipe a démarré avec les poids pré-entraînés du modèle open source, a personnalisé les deux dernière couche du réseau, puis a ré-entraîné les modèles en utilisant ses propres données.
3. *Test des exigences en données* : l'équipe a constaté qu'elle avait besoin de moins de données que prévu (par exemple, 100 à 1 000 exemples de chaque défaut et 50 000 à 60 000 exemples d'endoprothèses non défectueuses) pour être plus performante que les inspecteurs humains.

Pour améliorer la robustesse des modèles, elle a également augmenté les données d'entraînement en générant des exemples supplémentaires par le biais de la perturbation (il pouvait s'agir de simples ajustements qui ne devaient pas impacter la classification, comme des ajustements de luminosité ou l'ajout de bruit).

L'intégralité du processus a été effectué avec un budget relativement modeste de 50 000 USD. L'entraînement de modèle a été rapide et peu coûteux, prenant 1 à 2 secondes par image sur neuf modèles, 2 à 10 heures pour entraîner chaque modèle sur une seule unité de traitement graphique et avec une petite équipe d'environ trois personnes.

Résultats : Excellente performance du modèle et réduction des coûts de main-d'oeuvre

La précision obtenue a été supérieure à 90 % pour tous les modèles, les réseaux plus petits comme le VGG16 donnant même de bons résultats pour les problèmes simples. La précision a augmenté pour les modèles plus sophistiqués et avec plus de données. Par exemple, EfficientNet peut atteindre jusqu'à 97 % pour un réseau B0 avec 100 exemples et plus de 99 % pour un réseau B7 avec 1 000 exemples.

L'apprentissage par transfert n'atteint généralement pas ce niveau de performance. Les performances baissent généralement de manière significative, ce qui nécessite plus de données pour compenser le déficit. Dans ce cas, l'application des modèles existants à un problème plus simple semble avoir éliminé ce besoin.

Le déploiement des neuf modèles a permis de réduire de 5 millions USD les coûts de main-d'œuvre directs en diminuant le nombre de pièces signalées pour une inspection humaine et en réaffectant plusieurs experts à d'autres projets à forte valeur ajoutée.

L'expérience de Boston Scientific suggère que l'apprentissage par transfert fonctionne bien quand les conditions s'y prêtent :

- Il existe un modèle générique à exploiter. Dans le cas des tâches de traitement des images, les premières couches de ces réseaux semblent être hautement transférables, même lorsque la tâche est notablement différente.
- L'habituelle chute de performance liée à l'apprentissage par transfert peut être éliminée en utilisant le système sur un problème métier plus simple et en faisant des ajustements sur le réseau.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Crédit Mutuel

Utilisation de l'IA pour obtenir les bonnes informations pour les conseillers client

Résumé

La recherche de l'efficacité du service client est souvent en contradiction avec la volonté d'approfondir la relation avec le client. Les spécialistes sont mieux à même de résoudre des questions relatives à des produits ou services spécifiques, mais peuvent ne pas disposer du contexte nécessaire pour servir des clients dont la relation avec l'entreprise ne se limite pas à un seul domaine. Le Crédit Mutuel utilise l'IA pour mettre à l'échelle les points de contact dédiés aux clients pour de nombreux produits, en fournissant des informations structurées à ses représentants client.

Opportunité : Améliorer le service fourni par les conseillers humains

Au Crédit Mutuel, chaque client a un conseiller attitré. Le conseiller est le premier point de contact. Il oriente le client dans ses relations avec la banque sur les divers produits comme les comptes courants, l'épargne, les emprunts immobiliers, les investissements. Plus les conseillers accèdent rapidement et facilement aux informations dont ils on besoin, plus ils peuvent répondre rapidement aux demandes des clients (et plus ils disposent de temps pour servir d'autres clients). Avec environ 3 millions d'appels entrants et 7 millions courriers électroniques par mois, l'amélioration du délai de résolution peut avoir un impact significatif.

Défi : Documentation incohérente entre les produits et les groupes

Le grand nombre de produits dont est chargé un seul conseiller présente un défi dans la mesure où il doit disposer des informations nécessaires à portée de main. Pour répondre à une demande d'un client, les conseillers (qui sont souvent des généralistes) utilisent des moteurs de recherche ou appellent des collègues pour fournir les informations sur des produits spécifiques. Mais les banques du réseau Crédit Mutuel organisent leurs informations de façon différente, ce qui complique la recherche. De plus, la langue et la terminologie peuvent également ne pas être les mêmes. Par conséquent, les modèles linguistiques typiques prêts à l'emploi, ne suffisent pas pour hiérarchiser les informations présentées aux conseillers.

Solution : Intégration d'une terminologie personnalisée et classification hiérarchique

Pour créer une recherche linguistique personnalisée pour ses produits, le Crédit Mutuel a collecté toutes les questions posées à ses conseillers sur une période de trois à quatre mois, puis a structuré les réponses à ces questions (ce qui a pris quatre autres mois). Ce processus a été répété pour chacun des 11 domaines métier actuellement en production. Ensuite, les équipes ont entraîné un modèle d'apprentissage en profondeur pour générer des intégrations terminologiques personnalisées et l'ont utilisé pour entraîner un modèle de machine à vecteurs de support pour chaque domaine, afin de sélectionner les réponses les plus pertinentes pour chaque question. La banque a également créé des dizaines de milliers d'étapes de dialogue pour collecter les informations manquantes par rapport à la question initiale. La catégorisation de domaines initiale (qui dans ce contexte ne pouvait porter que sur des questions préliminaires courtes et simples) a été élaborée à l'aide d'un modèle FastText[15] qui s'est avéré aussi performant que le système de meilleure tentative suivante, BERT, mais beaucoup plus rapide, produisant un score F1 de 90 % avec seulement un entraînement hebdomadaire de 10 à 15 secondes et un délai de catégorisation de 20 à 30 millisecondes. Ce fractionnement a permis de réduire le nombre de classes dans chaque modèle SVM propre à un domaine.

Résultats : Meilleure qualité de réponse et résolution plus rapide des appels

Les modèles linguistiques améliorés ont permis d'améliorer la qualité et la rapidité des réponses. Désormais, l'assistant virtuel peut fournir la bonne réponse au client dans 85 % des cas (et 2 millions de réponses supplémentaires chaque année), tout en réduisant le délai de résolution de 3 minutes à 1 minute en moyenne. Les gains de temps globaux (pour les clients et les conseillers) s'élève à des dizaines de milliers d'heures chaque mois.

Ce cas montre l'utilisation de l'IA non pas pour fournir une réponse spécifique, mais comme partie intégrante d'un flux de travail humain, générant un ensemble plus restreint et plus ciblé de réponses proposées, où l'être humain peut appliquer son jugement subjectif.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Global Bank

L'IA peut accroître les capacités en complétant les processus existants avec des perturbations minimes

Résumé

Dans les secteurs hautement réglementés tels que la banque, il est extrêmement important de gérer une documentation sur les processus. Ainsi, les banques peuvent fonctionner de manière plus cohérente et reproductible et prouver leur conformité lors d'audits externes. Pour garantir la précision, l'exhaustivité et la compréhension, les banques procèdent à des audits internes sur la documentation au cours desquels leurs équipes examinent les contrôles de processus et tentent de les répliquer. Cependant, les comptes-rendus sont souvent complexes et peu structurés (texte en format libre), et les audits manuels prennent du temps. Une banque mondiale montre comment l'apprentissage en profondeur peut mettre à l'échelle ces opérations manuelles, tout en complétant, et non pas en remplaçant, les personnes et les processus existants.

Opportunité : L'assurance de l'audit améliore la qualité des contrôles

Les banques utilisent une multitude de processus, de l'ouverture d'un nouveau compte d'épargne à la réalisation d'un virement, chacun étant exposé à des risques de fraude ou de blanchiment d'argent, par exemple. La gestion de ces risques est cruciale, car le secteur financier est fortement réglementé, et s'expose à des sanctions sévères en cas de transgression. De plus, face à la concurrence croissante, notamment des fournisseurs de services en ligne, changer de banque n'a jamais été aussi facile. La confiance est donc un facteur important de rétention client.

Les banques atténuent ces risques par le biais d'un système de contrôles rigoureux. Certains sont automatisés, mais la plupart sont manuels. Pour vérifier la conception de ces contrôles et leur application efficace, le service d'audit interne (AI) d'une banque procède à des tests par échantillonnage pour s'assurer qu'ils fonctionnent. En cas de détection d'un problème (par exemple, une ouverture de compte qui ne se justifie pas), un plan d'action corrective est implémenté, tel que l'augmentation de la fréquence d'actualisation de la liste des entités bloquées. Plus le nombre de contrôles vérifiés est élevé, et le plus fréquemment possible, plus l'assurance fournie à l'entreprise est grande.

Défi : Améliorer la qualité de la documentation pour réaliser des audits efficaces

Pour répliquer des contrôles et évaluer leur efficacité, le service AI s'appuie sur une documentation de qualité. Elle doit au minimum contenir suffisamment d'informations pour savoir ce qui doit être fait, de quelle manière et les résultats attendus. Si certaines de ces informations sont manquantes, l'auditeur peut devoir s'entretenir avec les propriétaires des contrôles ou les responsables de la documentation, afin de procéder à des révisions, ce qui augmente le travail d'audit. En outre, la documentation est essentielle pour permettre aux régulateurs de comprendre les contrôles appliqués, et à ce titre, elle doit également souligner les responsabilités, les délais et d'autres informations concernant le processus.

En soi, le processus d'audit représente un nombre significatif de tâches manuelles. Global Bank réalise environ 1 000 audits par an sur environ 10 contrôles chacun qui durent environ trois heures en moyenne. Global Bank continue d'augmenter sa capacité (le nombre d'auditeurs va augmenter de 30 %) et dispose déjà de l'un des plus importants services d'audit internes au monde. Il est donc important d'optimiser la productivité de ces ressources.

Solution : Utilisation du traitement automatique du langage naturel pour signaler proactivement les lacunes potentielles dans les informations

Global Bank entreprit d'accroître l'efficacité du processus d'audit en améliorant la qualité de la documentation à l'aide d'un « correcteur instantané » pour les rédacteurs de documents. L'objectif était de développer un modèle de traitement automatique du langage naturel pour signaler automatiquement toute information importante qui pourrait manquer dans la documentation des contrôles sur la base d'un test des 5 W : What (quoi), Why (pourquoi), Who (qui), When (quand), Where (où). Global Bank pouvait utiliser le système lors de la rédaction initiale du document, ou parcourir les documents existants pour identifier ceux qui présentaient des problèmes.

Global Bank créa une preuve de concept basée sur un modèle BERT (Bidirectional Encoder Representations from Transformers) pré-entraîné (une technique de traitement automatique du langage naturel). Elle utilisa le modèle de reconnaissance des entités nommées, en essayant d'identifier les termes qui représentaient chacun des 5 W). Pour reconnaître la terminologie de Global Bank, le modèle devait être affiné et, du fait de l'arrivée de nouveaux modèles et des déploiements prévus des fonctionnalités dans d'autres services, le modèle devait être entrainé de nouveau plusieurs fois. Or BERT était un modèle important et complexe dont le ré-entrainement allait nécessiter des ressources de calcul importantes. De plus, Global Bank ne pouvait utiliser que du matériel sur site pour des raisons de sécurité.

La solution de Global Bank comportait deux volets. Tout d'abord, création d'un nouveau modèle plus facile à ré-entraîner. La banque s'associa à IBM pour créer ce modèle, amorcé lors d'un engagement antérieur, et procéda à une implémentation sur site à l'aide d'IBM® Watson Studio. Ensuite, augmentation des données disponibles pour ce nouveau modèle. Auparavant, la banque avait créé un système de marquage en Python directement connecté à sa plateforme d'audit interne. Ce système permit aux auditeurs de faire de nouvelles annotations lors des audits. La banque compléta ce système avec le modèle BERT d'origine, permettant ainsi aux auditeurs de recevoir des commentaires instantanément et de rendre ces données disponibles pour le nouveau modèle IBM.[18]

Résultats : Audits plus efficaces et capacité d'audit accrue

Le système offre trois avantages clés. Tout d'abord, en fournissant un retour d'information immédiat sur ce qui manque dans la description des contrôles, il améliore l'exhaustivité et l'exactitude de la documentation au moment de la rédaction. Ainsi, les nouveaux arrivants peuvent se familiariser beaucoup plus rapidement avec la documentation. Ensuite, il améliore la cohérence des comptes-rendus par rapport aux normes de l'entreprise. Enfin, l'optimisation de la qualité réduit les allers-retours entre les auditeurs et les propriétaires des contrôles, chacun étant ainsi plus productif.

En seulement quatre mois, l'adoption rapide du système donnait déjà des résultats. Cinquante utilisateurs actifs entrèrent collectivement 12 000 entrées (annotations personnalisées) dans plus de 5 000 contrôles et ajoutèrent des centaines d'entrées supplémentaires chaque semaine. Grâce à l'efficacité accrue du processus de vérification, Global Bank a économisé environ 30 000 heures de travail et l'a déployé pour renforcer l'assistance, ce qui n'aurait pas été possible autrement.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

IFFCO-Tokio

Les améliorations du processus AI génèrent de meilleurs incitations client

Résumé

Comme de nombreux secteurs, les compagnies assurance font face au défi de fournir des incitations adéquates. En général, les assurés ne sont pas motivés pour rechercher le meilleur prix pour des services dans la mesure où leur compagnie d'assurance paie les factures. Mais l'IA peut parfois rendre possibles les changements de processus qui leur permettent de libérer cette valeur.

Opportunité : Payer les clients directement

IFFCO-Tokio règle environ 500 000 sinistres automobiles en Inde chaque année. En règle générale, les clients faisaient réparer leurs véhicules dans des ateliers privés qui fournissaient des devis qu'IFFCO-Tokio devait approuver, mais ce processus posait plusieurs problèmes. Ces ateliers sont incités à gonfler l'estimation des coûts, et les clients ne sont pas incités à rechercher les meilleurs prix. Cette situation crée des litiges sur les frais de réparation et engendre des retards. Les clients, dont beaucoup dépendent de leur véhicule pour vivre, étaient gravement pénalisés par les retards qui allaient jusqu'à 20 à 30 jours par règlement. Certains ne pouvaient pas se permettre d'attendre, renonçant aux réparations et conduisant des véhicules peu sûrs.

IFFCO-Tokio décida de régler les coûts de réparation directement aux clients, afin de vraiment les inciter et de leur donner la possibilité de prendre en charge le processus en main. Mais deux obstacles se présentaient : fournir rapidement des devis et obtenir des estimations correctes sans devis de la part des ateliers de réparation.

Défi : Opérations manuelles et données de mauvaise qualité

Initialement, IFFCO-Tokio procédait manuellement. L'entreprise développa une application pour smartphone que les clients pouvaient utiliser pour télécharger des photos des dommages, recevoir un devis, décider eux-mêmes si le devis est acceptable et recevoir le paiement indépendamment du délai de réparation. Désormais, le client était aux commandes.

La nouvelle approche remporta un franc succès auprès des clients, mais elle était longue et imprécise. Ce processus prenait jusqu'à 5 heures par sinistre, dont une grande partie était consacrée à l'intervention d'experts pour inspecter les pièces et indiquer dans des formulaires les décisions de réparation ou de remplacement, ainsi que des estimations de coût pour chaque pièce. Et la qualité des images n'arrangeait rien. Auparavant, les photos des dommages étaient prises dans l'environnement contrôlé d'un garage professionnel, mais les clients soumettaient fréquemment des images prises sous des angles incorrects, avec un mauvais éclairage ou des reflets.

Solution : Capture d'images assistée et taille variable des données d'entraînement

IFFCO-Tokio avait parié sur l'apprentissage automatique pour accélérer le processus en fournissant automatiquement une première estimation pour chaque pièce, mais son équipe savait que la qualité et la cohérence des photos étaient essentielles. Pour améliorer la qualité de la capture d'images, l'équipe a étendu l'application en ajoutant des modèles pour appareil photo pour assister les clients dans la composition et a ajouté des instructions supplémentaires. Elle a également augmenté la quantité de données d'entraînement pour les types de pièces où l'éblouissement ou la réflexion rendent les dommages particulièrement difficiles à discerner (par exemple, 3 fois plus d'images pour les pièces métalliques et 5 fois plus d'images pour les pièces en verre). Grâce à cette combinaison d'images de meilleure qualité pour l'inférence et l'entraînement, il est possible d'utiliser des techniques d'apprentissage en profondeur pour classifier le modèle de voiture, les pièces endommagées et le type dommage. Sur cette base, le système peut déterminer si les pièces peuvent être réparées ou si elles doivent être remplacées et en estimer le coût.

Lors de l'automatisation d'un processus de prise de décision, le contrôle humain est généralement réduit. Il peut en résulter des abus potentiels, un client pouvant, par exemple, demander une indemnisation pour des dommages déjà indemnisés. IFFCO-Tokio ayant anticipé une augmentation de la fraude, l'assureur a créé un moteur de détection des fraudes pour identifier les images précédemment utilisées. Cependant, peu de fraudes sont détectées dans l'ensemble du système, car il est surveillé par un expert humain expérimenté.

Résultats : Gains de temps et réduction des coûts des règlements, augmentation de la rétention/acquisition client

Le système connaît un succès retentissant, tant de manière attendue qu'inattendue. Il réduit considérablement le travail des experts. Le temps de traitement de bout en bout est passé à 30 minutes par demande en moyenne (y compris les négociations avec les clients), et le nouveau système a été amorti en moins d'un an. Plus surprenant, IFFCO-Tokio constate également une réduction des versements de 40 %, et le taux d'acceptation est passé de 30 % à 65 %. De plus, le système est plus résilient, car les clients pouvaient toujours recevoir leurs règlements lorsque les garages étaient fermés pendant la pandémie. Enfin, et c'est peut-être le plus important, le système a permis d'améliorer directement la satisfaction et la rétention client et même l'acquisition de clients. L'IA est devenue non seulement un moteur d'efficacité accrue, mais aussi un moteur d'augmentation du chiffre d'affaires.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

KPMG

C'est la complexité des tâches et non pas la disponibilité des données qui détermine le choix de la méthode d'apprentissage automatique

Résumé

Les approches de traitement automatique du langage naturel et d'exploration de texte reposent généralement sur l'apprentissage en profondeur pour la reconnaissance d'entités nommées (NER). Ces approches permettent d'extraire du sens (par exemple, à partir de phrases et de paragraphes) et sont souvent performantes sur des tâches complexes. Cependant, l'expérience de KPMG dans le domaine de la recherche et de la classification de documents montre que lorsqu'une sous-tâche est suffisamment simple, l'apprentissage automatique traditionnel peut être la meilleure approche.

Opportunité : Meilleure justification des demandes de crédit d'impôts pour le R&D

Les incitations fiscales pour la recherche et le développement (R&D) aux États-Unis peuvent être importantes. Elles peuvent constituer jusqu'à 11 % - 15,8 % des dépenses R&D supplémentaires. Ces chiffres peuvent encore augmenter, car de nombreux états offrent également leur propre crédit R&D. Pour les plus petites organisations de recherche, en particulier, ces crédits d'impôt peuvent être décisifs pour rendre un projet de R&D commercialement viable ou obtenir des fonds d'investisseurs pour lancer des projets. Des entreprises de toutes tailles font appel à KPMG pour documenter le R&D effectué et pour leur permettre d'obtenir le crédit d'impôt maximal.

Défi : Nombreuses opérations manuelles

Aux États-Unis, l'IRS (Internal Revenue Service) évalue le bien-fondé des demandes de crédit d'impôt de R&D à l'aide d'un test en quatre parties. Ce test permet de vérifier que les activités faisant l'objet de la demande :

- Impliquent la création d'un composant métier ou l'amélioration d'un composant existant
- Sont de nature technologique
- Découvrent de nouvelles informations qui éliminent l'incertitude relative à la méthodologie, la fonctionnalité ou la conception du composant métier
- Impliquent un processus d'expérimentation via la simulation, la modélisation ou le test

Comme cette évaluation est clairement subjective, il convient donc de fournir des preuves indéniables pour recevoir un réponse positive.

Ces preuves sont généralement recueillies dans la documentation d'une entreprise et peuvent prendre différentes formes. Il peut s'agir de présentations, de courriers électroniques, de compte-rendus de réunion, de rapports de laboratoires, de dossiers de test et de dessins techniques. Le contenu est généralement non structuré, le volume ingérable, et/ou il est stocké dans divers référentiels. Dans certains cas, il peut être très limité, comme dans les environnements agiles ou en évolution constante. Dans tous les cas, la réglementation n'indique pas ce qui constitue une preuve « suffisante ». Par conséquent, il est important de vérifier toutes les informations disponibles et d'en présenter le plus possible qui soient pertinentes et de haute qualité.

KPMG accompagne ses clients dans leur préparation aux audits et possède l'expérience nécessaire dans la gestion du processus de reconnaissance. Traditionnellement, il s'agit d'une approche descendante, en commençant par une liste de projets, en parcourant des documents sur ces projets, en effectuant des recherches manuelles dans les documents en utilisant des mots-clés et en lisant et en étiquetant des sections spécifiques des documents qui satisfont à chacun des quatre tests. Ces opérations manuelles importantes nécessitaient d'établir des priorités qui pouvaient exclure des preuves précieuses. Elles étaient également chronophages pour les scientifiques et les ingénieurs des clients qui s'y consacraient. KPMG a alors envisagé de recourir à l'apprentissage automatique pour être plus efficace.

Solution : Des approches basées sur des règles plus performantes que l'apprentissage automatique

KPMG lança un hackathon interne avec quatre équipes en compétition pour résoudre un sous-ensemble du problème (granularisation de document) via d'autres méthodes. Les équipes reçurent 1 000 documents avec des sections étiquetées, et il leur fut demandé d'indiquer un taux de confiance quant à la pertinence de chaque document pour chacun des quatre tests.

Les documents furent découpés en sections par la tokenisation des mots et des phrases. Les équipes essayèrent différentes approches, notamment, l'apprentissage statistique, comme des expressions régulières, des machines à support de vecteur, des arbres de décisions et une forêt aléatoire, l'apprentissage en profondeur pour la reconnaissance d'entités nommées (NER) et des approches basées sur des règles. Elles relevèrent une précision de 55 %, en utilisant un logiciel prêt-à-l'emploi de reconnaissance de documents (à peu près aussi efficace qu'une recherche manuelle par mots clés) à plus de 70 % pour l'apprentissage en profondeur. Mais, les approches basées sur des règles se sont avérées plus performantes avec une exactitude supérieure à 85 %. Cela s'explique probablement par un degré relativement élevé de normalisation entre les formats de documents, ce qui rend la segmentation des documents relativement simple.

Résultats : Des preuves plus probantes donnent plus de poids aux demandes de crédits

Le système est maintenant utilisé avec succès pour plusieurs clients de KPMG. Chaque mois, il traite plus de 5 000 documents, mais avec un changement essentiel : la recherche passe d'une approche sélective et descendante à une approche ascendante et exhaustive. La législation fiscale évoluant peu dans le temps, le système ne nécessite qu'une maintenance et des améliorations minimes et permet de tirer le meilleur parti de l'expertise humaine. Des preuves anecdotiques montrent l'ampleur de son impact. Par exemple, un client de KPMG est parvenu à obtenir un crédit d'impôt supplémentaire de 40 % pour son étude sur le crédit d'impôt recherche, en utilisant l'apprentissage automatique pour examiner la documentation des projets R&D, afin de déterminer l'éligibilité au crédit d'impôt.

Il convient de réfléchir sur la performance relative de l'apprentissage en profondeur par rapport à d'autres méthodes. Ces résultats confirment que même si le jeu de données est suffisamment grand, l'apprentissage en profondeur tend à être plus performant uniquement lorsque les données ou les problèmes sont extrêmement complexes. Dans ce cas, de simples règles et mots-clés ont suffi à identifier les informations pertinentes pour chaque test, tout en offrant une explicabilité accrue.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Marketing Platform

Maîtriser le coût de l'IA grâce à une visibilité accrue

Résumé

Une entreprise dotée d'une infrastructure sur site n'a généralement aucune visibilité sur le coût de l'entraînement et de l'exécution des modèles d'apprentissage automatique. Par exemple, il peut exister une déconnexion entre la valeur générée par le modèle et les ressources en ingénierie et en informatique requises pour atteindre l'objectif.

Cette étude de cas montre que la transparence du lien entre la demande de calcul et le coût peut contribuer à créer des incitations et à réduire les coûts de manière significative.

Opportunité : Migrer vers le cloud pour tirer parti plus efficacement des données existantes

Marketing Platform aide les distributeurs et les associations à but non lucratif à améliorer les retombées de leurs opérations marketing en prédisant les personnes chez lesquelles chaque campagne aura le plus d'écho. Une petite amélioration de la précision, à grande échelle, peut générer des millions USD de ventes ou de dons supplémentaires : les enjeux sont donc de taille.

Les jeux de données disponibles pour ces modèles sont énormes. Marketing Platform gère une coopérative de données avec des milliers de membres, dont 25 % à 40 % fournissent régulièrement des données concernant notamment des transactions, des dons ou des abonnements. L'organisation combine ces données avec des données de tiers compilées sur tous les sujets, de la démographie aux données de recensement en passant par les revenus des ménages. À l'issue de l'ingénierie des caractéristiques, les données sont constituées de 12 000 variables et couvrent la quasi-totalité de la population américaine.

Avec un ensemble de données aussi riche, Marketing Platform était à la limite de ce que son infrastructure existante sur site pouvait gérer. L' équipe n'est parvenu à entraîner des modèles que sur des données internes, et encore, seulement une petite partie d'entre elles (par exemple, un échantillon de 50 000 à 100 000) à la fois. Marketing Plaform savait que si elle pouvait utiliser plus de données, il existait une énorme possibilité de générer de la valeur supplémentaire.

Défi : Coût initial du cloud significativement plus élevé

La migration vers le cloud (IBM® Cloud Pak for Data) a amélioré la capacité de l'entreprise à gérer ses données et à tirer parti de tous ses actifs de données, à la fois hors ligne et en ligne. Le gain d'évolutivité des ressources de calcul a également permis d'exécuter une entraînement sur 600 000 enregistrements (au lieu d'un maximum de 100 000 auparavant) et 800 caractéristiques (au lieu de 150-200). Ainsi, en combinaison avec des outils d'apprentissage automatique comme XGBoost,[16] le taux de réponse a augmenté de 20 à 30 %, une hausse spectaculaire des retombées pour les clients.

Dans un premier temps, la mise à l'échelle des ressources de calcul (ainsi que le coût initial du changement et la courbe d'apprentissage) a entraîné une augmentation des coûts totaux. Comme dans un environnement sur site, le coût de l'infrastructure était indépendant de l'utilisation, les scientifiques des données pouvaient exécuter ce qu'ils voulaient, la seule contrainte étant la disponibilité des ressources de calcul. Avec une évolutivité effectivement illimitée, les expériences devaient être conçues de manière plus réfléchie.

Solution : La connexion à des incitations marginales permet de compenser les surcoûts

Heureusement, le passage au cloud a également permis aux responsables de Marketing Pateform de mieux comprendre les dépenses et finalement de les optimiser. L'équipe pouvait désormais générer un coût de calcul par modèle et intégrer ces incitations marginales lors de la structuration de l'exploration et de l'analyse des données.

Elle a également pu optimiser l'utilisation des ressources de calcul en améliorant l'allocation de grappe, le flux de données et le pipeline de modélisation global. Pour le modèle d'apprentissage automatique lui-même, elle a effectué des tests pour optimiser les quelque 100 paramètres de modèle requis et en a corrigé plusieurs en fonction de ce qui fonctionnait bien, afin de réduire le nombre de ceux à ajuster à chaque fois.

Résultats : Réduction des coûts d'entraînement et déploiement plus étendu

Il en résulte une réduction spectaculaire des coûts d'entraînement qui sont passés de 1 500 dollars à quelques centaines de dollars par cycle d'entraînement, même avec l'augmentation de 30 % des performances du modèle.

Le succès de la plateforme dans ce domaine alimente la transformation de ses fonctionnalités de science des données, le nombre de praticiens aux États-Unis passant de 40 à 4000 dans l'année.

Les principaux points à retenir suggèrent que :

- La visibilité des coûts d'entraînement à un niveau de granularité par exécution de modèle peut contribuer à rendre le coût du passage à de nouvelles techniques d'apprentissage automatique moins élevé que prévu.
- Le volume compte : même de petits gains d'exactitude peuvent avoir un impact énorme sur une organisation lorsqu'ils peuvent être appliqués à grande échelle.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

McCormick

Augmentation de la creativité en R&D grâce à l'exploration dirigée par l'IA

Résumé

L'IA est généralement utilisée pour faire des recommandations basées sur ce qui a bien fonctionné dans le passé, mais les solutions qui en découlent sont alors semblables aux précédentes. L'expérience de McCormick montre que l'IA peut être également utilisée pour explorer un espace de solution et aboutir à de nouvelles combinaisons créatives qui n'auraient pas été tentées autrement. Ainsi, l'IA peut permettre d'améliorer et d'accélérer un processus d'expérimentation créative.

Opportunité : Utiliser l'IA pour accélérer le développement de nouveaux profils d'arômes

McCormick crée toute une gamme de produits, notamment des assaisonnements, des sauces et des arômes (dont certains peuvent être également vendus à d'autres entreprises (B2B) pour les incorporés dans des produits de tiers). Une formule est une combinaison d'ingrédients spécifiques dans des proportions précises, normalisées à des fins de cohérence, qui produit un profil d'arôme qui décrit l'expérience gustative. La création d'arômes pour les produits B2B peut être un processus concurrentiel, car plusieurs entreprises produisent des formules pour répondre aux demandes des clients. Pour augmenter le succès des nouveaux arômes, McCormick a examiné deux facteurs importants :

- Mise à l'échelle de l'expérience des scientifiques en produits alimentaires Un aromaticien débutant fait généralement son apprentissage pendant sept ans, au cours desquels il acquiert de l'expérience et accumule des connaissances importantes. C'est cette expérience cumulative qui permet la creativité, par la compréhension de ce qui fonctionne, de ce qui ne fonctionne pas et des degrés de liberté entre les deux.
- Amélioration de l'efficacité du processus d'expérimentation. La production d'un arôme candidat pour un client passe par plusieurs étapes. Les aromaticiens formulent une gamme de profils d'arômes différents, produisent des échantillons des arômes, les testent (à la fois isolément et après cuisson dans une cuisine test), et réitèrent jusqu'à ce qu'ils obtiennent un arôme pouvant être présenté au client.[17] L'amélioration de l'efficacité de ce processus accélère la mise sur le marché et augmente la productivité des scientifiques en produits alimentaires.

McCormick s'est demandé ce que pouvait apporter l'IA dans ce domaine. Si les aromaticiens pouvaient tirer parti d'informations de l'expérience empirique, il serait alors sans doute possible d'en extraire davantage, plus vite, grâce à l'analyse. Si ces informations pouvaient être capturées et systématisées, alors les aromaticiens pourraient mieux explorer l'espace des arômes de deux manières. La première consiste à trouver des arômes optimaux dans une zone proche de ce qui est bien connu. La seconde vise à trouver de nouvelles zones prometteuses dans l'espace des arômes qui n'ont pas été explorées. Ainsi, il serait possible d'accélérer le développement et d'améliorer la qualité

Défi : Affectation de crédit et grand espace de recherche

L'entreprise a exploité des données sur environ 350 000 formules créées sur plus d'une décennie, couvrant les attributs de produit tels que la catégorie (produits de boulangerie, produits apéritifs salés), le format (assaisonnement, condiment, sauce liquide ou en poudre), les quantités, le type et les mesures de succès telles que les notes de dégustation des produits. Elle a également capturé des attributs fonctionnels comme la durée de conservation, le flux et des attributs non fonctionnels comme le niveau de granularité, la teneur en sodium et des valeurs FEMA[18] (attributs sur 40 000 matières premières). Avec une dimensionnalité si élevée dans le jeu de données, l'équipe devait trouver un moyen de condenser le problème pour qu'il reste gérable.

Solution : Représentations graphiques et dimensionnalité réduite

Un nouveau système d'apprentissage en profondeur, SAGE, a été développé pour générer de nouveaux profils d'arômes. Il accepte deux entrées principales définies par l'utilisateur : (1) une formule de départ (par exemple, un profil d'arôme pour un barbecue coréen) et (2) toutes les contraintes caractéristiques souhaitées dans les formules en sortie (telles que « doit comporter de la mangue »). Le système génère ensuite des formules avec différents niveaux d'écart par rapport à la formule de départ : quatre avec seulement des ajustements mineurs qui optimisent la performance anticipée, quatre avec une plus grande liberté, mais toujours avec des contraintes et quatre qui diffèrent de manière significative. Les aromaticiens disposaient ainsi d'un éventail d'options à partir desquelles ils pouvaient itérer, en fonction du niveau de nouveauté souhaité.

Pour soutenir ces opérations l'équipe a dû recourir à quelques astuces. Elle commença par réduire la dimensionnalité des données en agrégeant 40 000 matières premières distinctes dans 3 000 groupes. Ensuite, elle échantillonna 3 000 formules à des fins d'entraînement sur un total de 350 000, chacune portant une étiquette de taux de « succès ». Enfin, elle formula le modèle sous la forme d'un problème de réalisation graphique, en définissant des mesures de distance entre les matières où chaque formule était représentée sous la forme d'un vecteur.

Résultats : Performance équivalente à 20 ans d'expérience

McCormick constata que les performances des scientifiques en produits alimentaires débutants étaient similaires à celles d'un collègue ayant 20 ans d'expérience, réduisant ainsi de manière significative le nombre d'essais nécessaires. Elle remarqua également que le système permettait d'utiliser plus efficacement les connaissances à l'échelle mondiale. Dans un cas, le système recommanda un profil d'arôme du Canada à un aromaticien aux États-Unis qui n'avait aucune expérience préalable du marché canadien, ce qui permit d'accroître la créativité, tout en ciblant également mieux l'expérimentation.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Navtech

Les nouvelles plateformes créées par l'apprentissage en profondeur

Résumé

Certaines innovations ne sont économiquement réalisables que lorsqu'un fournisseur centralisé peut faire les investissements initiaux requis et répartir les coûts sur une base de clientèle suffisamment large. C'est de plus en plus vrai pour l'apprentissage automatique, en particulier, lorsque les tâches automatisées reposent sur la perception (par exemple, la reconnaissance d'image), car leur développement et leur maintenance peuvent nécessiter d'importantes quantités de données et un grand nombre de ressources de calcul. Pour de nombreux particuliers et entreprises, cette technologie est tout simplement hors de portée.

Navtech a identifié une possibilité d'apporter la vision par ordinateur avancée à des négociants en diamants du monde entier en créant un modèle et le fournissant en tant que service. Cette approche peut être gagnant-gagnant et constitue un excellent exemple de la façon dont l'IA peut non seulement améliorer l'efficacité et les performances, mais aussi libérer de nouvelles fonctionnalités et de nouveaux modèles de gestion.

Opportunité : La numérisation des catalogues permet d'accroître les ventes

L'Inde compte à elle seule environ 300 000 négociants en diamants. Beaucoup d'entre eux sont de petites entreprises, dont la capacité de stockage est limitée, qui élargissent généralement leur offre en proposant des bijoux sur mesure. L'ajout d'options de bijoux sur mesure peut doubler le taux de conversion (deux fois plus de clients qui trouvent quelque chose qui leur plaît et effectuent un achat) par rapport à la seule offre de leurs bijoux en stock.

Les catalogues visuels sont une composante clé du processus de ventes. Chaque détaillant en gère un pour son propre stock et le complète avec des images d'autres bijoux, afin d'inspirer les clients qui recherchent des pièces sur mesure. Traditionnellement, il s'agit de brochures ou de magazines physiques, mais ces formats ne peuvent présenter qu'un nombre limité d'articles et ne peuvent pas être mis à jour fréquemment.

Si les catalogues numériques sont plus flexibles, ils présentent d'autres défis. Le personnel compile des images provenant de diverses sources, comme des photographies de stock, le Web et des catalogues de fabricants, et les classe manuellement dans des dossiers. Le traitement est lent (30 à 60 secondes par image, jusqu'à un million d'images), sujet aux erreurs (beaucoup d'images sont en double et il est difficile de se souvenir de ce qui a déjà été vu) et ne permet qu'une catégorisation très générale (par exemple, bagues versus colliers). L'idéal serait de pouvoir créer automatiquement des catalogues et de permettre aux clients d'effectuer des recherches en fonction de critères supplémentaires.

Défi : La vision par ordinateur coûte trop cher pour les négociants en diamants

Les systèmes de vision par ordinateur qui utilisent l'apprentissage en profondeur pour classifier des images pourraient contribuer à améliorer la vitesse et la précision, mais ils sont inabordables pour la plupart des négociants. L'apprentissage en profondeur consomme beaucoup de ressources et nécessite d'énormes quantités de données et de ressources de calcul tant pour l'entraînement que pour l'implémentation. Le système peut ne pas être utilisé assez fréquemment pour justifier cet investissement, en particulier lorsque le coût comparatif de la main-d'œuvre pour la classification manuelle est faible. Par exemple, en Inde, les salaires des employés de la distribution débutent à environ 100 USD par mois. L'étude de rentabilité de la vision par ordinateur pour ces négociants a donc peu de chances d'être attractive.

Solution : Créer une fois et fournir en tant que service

Dr M.I.M. Loya, directeur général des technologies émergentes chez Navtech, a eu une idée : créer un système de vision par ordinateur et le proposer en tant que service. Cela pouvait être une solution gagnant-gagnant. Navtech disposait des ressources nécessaires pour réaliser l'investissement initial et offrir l'accès au système pour une somme modique, et les négociants pouvaient bénéficier d'un accès continu à un faible coût au système.

Navtech sélectionna trois attributs pour son pilote et créa un modèle d'apprentissage en profondeur pour chacun d'eux :

- Pour la catégorie de produit (par exemple, bagues, bracelets) et le style, l'entreprise utilisa un réseau VGG16[19] pour classifier les images. Le réseau dorsal open source entraîné par ImageNet fut optimisé en entraînant de façon personnalisée la partie supérieure et les première et seconde couches du réseau.
- Pour la coupe des diamants (par exemple, en rond, en carré), Mask RCNN fut utilisé pour la détection et la classification des objets (car la précision atteinte fut seulement de 55-56 % en utilisant VGG16). Les données d'entraînement de ce modèle furent étiquetées par un stagiaire qui a dessiné manuellement un masque polygonal autour de la forme de chaque diamant.

Résultats : Un accès plus large à un apprentissage automatique de pointe à un coût gérable

Le système permet aux négociants de créer des catalogues numériques plus volumineux beaucoup plus rapidement. Il peut classifier jusqu'à 100 images par minute (contre une à deux par minute avec la classification manuelle), avec une exactitude de 90 % à 93 % pour la catégorie et le style de produit et 85 % à 86 % pour la coupe de diamant (contre 80 % pour l'étiquetage manuel). En outre, l'équipe a pu le faire en utilisant un jeu de données relativement petit (seulement 3 000 images étiquetées pour chaque modèle), ce qui est quelque peu surprenant. Navtech réalisa un post-traitement, où les résultats d'un modèle (tel que le style) furent utilisés pour augmenter le niveau de fiabilité des prévisions pour un autre modèle (tel que la coupe de diamant). Néanmoins, il est possible que l'effet de levier élevé de ce petit ensemble de données soit également dû à une diminution de la complexité des problèmes par rapport à ImageNet (des images de bijoux varient moins que des images de chats, par exemple).

Cette expérience illustre les compromis coûts-bénéfices de l'apprentissage en profondeur, où certains cas d'utilisation ne peuvent être effectués que par des acteurs plus importants et centralisés pouvant servir des publics plus larges. Ces systèmes sont fournis à grande échelle et doivent donc l'être dans le cadre d'une architecture de produits et de services plus large, soutenue par un développement logiciel traditionnel, et ils sont beaucoup plus rentables.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

NVIDIA

Création d'une plateforme de calcul et de données pour les véhicules autonomes

Résumé

Pour créer des modèles d'IA sûrs et fiables pour les véhicules autonomes (VA), une énorme puissance de calcul et de très grandes quantités de données d'entraînement sont nécessaires, sans parler des compétences, des ressources et l'expertise requises à cette échelle. Dans ces conditions, il est probable que nous assistions à l'émergence de grandes plateformes capables de regrouper les données de plusieurs participants, d'agréger une demande suffisante pour justifier les investissements importants requis et, en fin de compte, de mettre en œuvre un modèle de gestion où les logiciels VA peuvent être proposés en tant que service aux constructeurs automobiles et exploitants de parcs automobiles.

Opportunité : Résoudre des problèmes complexes pour une adoption en masse d'AV

Les systèmes d'aide à la conduite sont de plus en plus répandus et peuvent désormais garer un véhicule, effectuer un freinage d'urgence, changer de voie, etc. Lorsque les véhicules seront entièrement autonomes, les applications seront nombreuses, qu'il s'agisse de transport de machandises, de transport en commun ou encore le transport à la demande, comme les robots-taxis. Le marché mondial des systèmes de véhicules autonomes était de 82 milliards USD en 2021 et il devrait atteindre 770 milliards USD d'ici 2030, soit un taux de croissance annuel moyen de 39,1 % (de 2022 à 2030).[20]

Défi : Puissance de calcul et volumes de données énormes

Les besoins en données des véhicules entièrement autonomes sont énormes du fait de l'éventail des tâches de planification et de contrôle à effectuer, telles que la détection des piétons, des marquages au sol et des feux de signalisation. Ces fonctions doivent être robustes dans des conditions environnementales changeantes, dont la météo et les environnements locaux sont, par exemple, des caractéristiques. Elles doivent aussi pouvoir gérer des événements temporaires ou rares (par exemple, une coupure inattendue). Les systèmes actuels améliorent le nombre et l'efficacité de ces fonctions, mais leur fiabilité en termes de réduction du nombre de décès et de blessures est loin d'être démontrée. L'entreprise RAND estime que 11 milliards de kilomètres d'essais seraient nécessaires pour égaler un taux d'erreur humain, ce qui représente le test de 100 véhicules en continu pendant plus de 500 ans.[21] L'entreprise NVIDIA elle-même estime que de bonnes performances sur certaines tâches VA nécessitent des exemples d'entraînement de l'ordre d'un million de scènes.

Dans la mesure où chacune de ces scènes impliquera des données provenant de nombreux capteurs, le défi informatique pour les VA est énorme. Pour créer les modèles de perception nécessaires à une pile VA complète, NVIDIA estime qu'une équipe de développement performante pourrait nécessiter environ 5 000 GPU dédiées.[22] L'exécution d'un modèle peut prendre de trois à six jours sur 32 GPU, et il peut exister entre 25 et 50 expériences d'apprentissage en profondeur pour chaque tâche. Les constructeurs automobiles n'ont généralement pas les ressources en termes de compétences, d'expérience, de matériel et de données pour développer seuls ces systèmes.

Solution : Une plateforme partagée de calcul et de données entre plusieurs clients

NVIDIA relève ces défis des façons suivantes :

- Extension d'une plateforme de données commune à plusieurs clients : la mutualisation des données entre plusieurs constructeurs automobiles augmente la quantité de données disponibles pour l'entraînement et peut assurer de meilleures performances du modèle, en particulier avec les cas extrêmes. La qualité des données peut être assurée par une architecture de référence qui définit les normes des spécifications et de placement des capteurs.
- Simulation pour l'entraînement et le test : des centaines de millions de scénarios de conduite peuvent être simulés pour compléter les données du monde réel et lancer des modèles pour des essais sur la voie publique et l'itération, en exécutant l'IA dans des véhicules pour comparer ce qu'elle aurait fait par rapport au comportement réel du conducteur.
- Traitement commun des tâches visuelles : NVIDIA a réduit le calcul requis en entraînant conjointement plusieurs tâches sur une architecture de modèle unique basée sur ResNet. Une fois le modèle complet entraîné, les têtes (couches ultérieures) du modèle peuvent être optimisées pour chacune des tâches données, sans qu'il soit nécessaire d'entraîner à nouveau le tronc (couches antérieures) du modèle. Le fait que le calcul ne soit pas beaucoup plus important que celui requis pour une tâche unique suggère qu'un large calcul commun est possible, ce qui se comprend intuitivement pour le domaine de vision par ordinateur.

Résultats : Un nouveau modèle de gestion qui permet la concurrence à différents niveaux de la pile

Une telle centralisation de la gestion des données ouvre de nouvelles possibilités pour la technologie VA. Selon leurs besoins et leurs fonctionnalités existantes, les constructeurs automobiles participants peuvent soit louer du matériel VA pour entraîner leurs propres modèles sur la base d'un jeu de données plus large, soit utiliser des modèles VA pré-entraînés de NVIDIA. Dans les deux cas, au lieu d'investir massivement dans des matériels et des fonctionnalités de développement, les constructeurs automobiles peuvent comptabiliser la technologie VA dans leurs dépenses d'exploitation et bénéficier d'améliorations à mesure que les matériels et les logiciels deviennent plus performants.

Il s'agit aussi du début d'une nouvelle dynamique de marché. D'un côté, des constructeurs automobiles verticalement intégrés (tels que Tesla), qui peuvent co-concevoir leurs logiciels et matériels pour des expériences plus fluides, de l'autre, des constructeurs toujours plus modularisés, en concurrence au niveau de la qualité de leur matériel et qui achètent leurs logiciels auprès d'acteurs centralisés comme NVIDIA (ce qui réduit considérablement le coût d'entrée sur le marché VA et est susceptible d'engendrer une plus grande concurrence). Le succès de l'un ou l'autre de ces deux paradigmes dépend de l'importance de la qualité des logiciels VA en termes d'expérience globale.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Suncor

Performance et explicabilité ne sont pas toujours une question de compromis

Résumé

L'apprentissage en profondeur est généralement performant dans les processus de prévision non linéaires (de petits changements peuvent avoir un impact considérable) et fortement couplés (beaucoup de dépendances entre les facteurs). Généralement, l'adoption de l'apprentissage en profondeur améliore suffisamment les performances pour que les entreprises acceptent que leurs modèles soient moins explicables (boîte noire). Mais comme le montre l'expérience de Suncor, lorsque les enjeux sont suffisamment élevés, l'explicabilité est primordiale.

Opportunité : prévisions plus précises des problèmes pour gérer la qualité finale

Suncor Energy est spécialisé dans la production de pétrole brut synthétique à partir de sables bitumineux. Pour le diesel en particulier, il s'agit d'éliminer le soufre et l'azote par hydrotraitement, en mélangeant le diesel de distillation directe avec de l'hydrogène (et un catalyseur métallique solide tel que le cobalt) à haute température et haute pression. Le processus est complexe, impliquant plusieurs variables (pression, température, débits) qui interagissent pour influer sur la qualité du produit final. Ces facteurs doivent être étroitement surveillés et contrôlés, afin de réduire les « perturbations », lorsque la qualité des produits s'écarte d'un prix acceptable et que les produits ne peuvent pas être vendus. Avec une production de diesel de 43 000 barils par jour en moyenne, il existe une forte incitation commerciale à éviter autant que possible les perturbations.

Défi : Les décisions à fort impact nécessitent une haute explicabilité

La responsabilité finale de la qualité résultante incombe aux responsables des sites qui supervisent les sites de production et prennent des décisions opérationnelles clés qui l'influencent. Savoir les ajustements à effectuer et connaître les impacts induits relèvent généralement de l'expérience et du jugement de chacun. Pour être défendables, ces décisions à fort impact doivent être clairement justifiées, ce qui signifie que toute technique analytique d'aide à la décision doit être transparente et bien comprise.

Solution : L'analyse en composantes principales a une plus grande explicabilité et des performances comparables

Suncor entreprit d'améliorer ses fonctionnalités de prévision en développant un modèle de signalement des perturbations qui permet de déceler les conditions non optimales émergentes suffisamment à l'avance pour procéder à des mesures correctives. L'équipe de science des données identifia 11 facteurs associés à la qualité des produits qui devaient être évalués en temps réel et créa des modèles intégrant 30 mesures différentes issues de données de capteurs. Dans un premier temps, elle explora un large éventail de techniques sophistiquées d'apprentissage automatique, mais difficiles à interpréter, notamment des réseaux de neurones, la mémoire longue à court terme, les forêts aléatoires, l'amplification de gradient et les arbres de décision, XGBoost (une technique d'ensemble combinant des arbres de décision et l'amplification de gradient), offrant les meilleures performances.

Cependant, lorsque l'équipe compara cette performance à celle des techniques statistiques traditionnelles plus simples, elle constata des performances nettement meilleures que prévu ; par exemple, l'analyse en composantes principales (ACP) n'a présenté qu'un déficit de performance de 10 % par rapport à XGBoost, tout en étant beaucoup plus facile à interpréter.

Résultats : Avertissement en amont avec justification transparente et défendable

Après avoir testé minutieusement les deux approches avec ses principaux intervenants à l'échelle du site, Suncor décida que le plus haut niveau d'explicabilité l'emportait largement sur le compromis de performance dans ce cas. Combinée aux prévisions, PCA a permis de lire les pondérations associées derrière chaque facteur, comme le classement des facteurs les plus importants dans la prévision. Le système résultant put prévoir des événements de perturbation jusqu'à une heure à l'avance, avec une exactitude de 80 %, toutes les cinq minutes.

Annexe
BESTSELLER
Boston Scientific
Crédit Mutuel
Global Bank
IFFCO-Tokio
KPMG
Marketing Platform
McCormick
Navtech
NVIDIA
Suncor
Zzapp

Zzapp Malaria

Apprendre des images satellites pour lutter contre le paludisme

Résumé

Avec l'imagerie satellite, la vision par ordinateur est couramment utilisée pour identifier des objets visibles, les réseaux de neurones convolutifs (CNN) étant généralement le choix par défaut. Même lorsque les objets eux-mêmes ne sont pas clairement visibles, les modèles prédictifs utilisant des réseaux CNN peuvent parfois déduire leur présence en fonction d'autres caractéristiques (par exemple, la zone entourant l'objet en question). Cependant, comme le montre l'expérience de ZzApp Malaria, ce n'est pas toujours le cas, et dans de telles situations, des techniques traditionnelles, telles que la régression linéaire, peuvent suffire.

Opportunité : Prévenir le paludisme grâce au traitement des eaux stagnantes

Le paludisme a causé environ 627 000 décès en 2020, dont 96 % en Afrique. La lutte antivectorielle est le principal moyen d'éviter la transmission par les piqûres de moustiques anophèles porteurs du paludisme. Jusqu'à maintenant, les principaux modes de prévention sont des produits tels que des moustiquaires ou la pulvérisation d'insecticides en intérieur, mais ils ne sont que partiellement efficaces et ne fonctionnent pas en extérieur. Une autre approche consiste à traiter directement l'eau stagnante au sein de la communauté (où les moustiques se reproduisent et se multiplient), mais ces programmes ne sont pas suffisamment systématiques ou complets pour être efficaces à grande échelle.

Défi : Petits plans d'eau non visibles sur les images satellites

Dans le traitement de l'eau, la difficulté est d'identifier l'eau stagnante, afin de pouvoir la gérer. Les grands plans d'eau sont facilement visibles sur l'imagerie satellite, et des algorithmes de vision par ordinateur permettent de les identifier automatiquement. En revanche, les plans plus petits sont difficiles à détecter, même avec des techniques sophistiquées d'imagerie par satellite, et ils peuvent être recouverts ou n'exister que d'une saison à l'autre. Si les plans d'eau stagnante pouvaient être mieux identifiés, il serait alors possible de diriger plus efficacement les pulvérisations et de mieux contrôler la population de moustiques.

Solution : Déduire la présence d'eau stagnante grâce à la topographie

L'entreprise Zzapp Malaria a été créée pour gérer ce problème et a commencé par étudier les zones critiques du paludisme à Sao Tomé. Elle créa une application qui permet aux inspecteurs sur le terrain d'enregistrer l'emplacement des plans d'eau rencontrés, de suivre les traitements de l'eau dans le temps et d'établir un ensemble d'exemples positifs d'entraînement, tels que les emplacements où l'eau stagnante est présente. Elle collecta également des images satellites (photographies, infrarouge et radar) et les utilisa pour entraîner un algorithme de détection d'objets basé sur un réseau CNN, qui fonctionne bien pour les grands plans d'eau, mais mal pour les petits (en particulier lorsqu'ils sont obscurcis).

Pour résoudre ce problème, l'équipe extraya des images 50 caractéristiques topographiques et basées sur des images et les utilisa dans une approche traditionnelle reposant sur la régression linéaire pour déterminer la probabilité d'eau stagnante dans chaque segment d'une carte. Avec une précision de 75 %, la performance est équivalente à celle du réseau CNN, mais offre une transparence nettement plus grande sur les facteurs qui déterminent la prévision. L'équipe constata également que les déterminants topographiques dépendent fortement de l'environnement local et que la régression linéaire est plus facilement transférable à d'autres environnements locaux que les approches basées sur des réseaux de neurones.

Résultat : Une approche transparente et transférable à d'autres environnements locaux

La performance relativement élevée des modèles de régression peut s'expliquer par le fait qu'ils ont pu tirer parti des caractéristiques sur la façon dont l'eau stagne en fonction des caractéristiques topologiques des données, au lieu d'avoir à les inférer avec un réseau CNN. Dans tous les cas, la transparence et la transférabilité supplémentaires du modèle de régression sont essentielles à la volonté de Zzapp d'étendre son approche au-delà de Sao Tomé et à d'autres pays, au Ghana, au Zanzibar et au-delà, où le terrain peut être significativement différent.

À propos de Research Insights

Research Insights fournit aux dirigeants des entreprises des informations stratégiques factuelles sur des questions essentielles dans le secteur public comme dans le secteur privé. Ces informations reposent sur les conclusions de l'analyse de nos propres travaux de recherche de base. Pour plus d'informations, contactez l'IBM Institute for Business Value à l'adresse iibv@us.ibm.com.

Votre partenaire dans un monde qui change

IBM collabore avec ses clients en réunissant les informations métier, la recherche avancée et les technologies, afin de leur apporter un avantage distinct dans l'environnement actuel qui évolue rapidement.

IBM Institute for Business Value

Depuis deux décennies, l'IBM Institute for Business Value est le groupe de réflexion d'influence d'IBM. Produire des informations stratégiques fondées sur la recherche et la technologie qui aident les dirigeants à prendre des décisions opérationnels plus judicieuses est ce qui nous motive.

Chaque année, grâce à notre positionnement unique au carrefour des affaires, de la technologie et de la société, nous sommes en mesure d'interroger des milliers de dirigeants, de consommateurs et d'experts, et d'interagir avec eux. Nous pouvons ainsi synthétiser leurs points de vue dans des informations crédibles, inspirantes et exploitables.

Inscrivez-vous sur ibm.com/fr-fr/ibv pour recevoir le bulletin d'information de l'IBV par courrier électronique et rester connecté et informé. Vous pouvez également nous suivre sur Twitter (@IBMIBV) ou nous retrouver sur LinkedIn (https://ibm.co/ibv-linkedin).

Remarques et sources

1 Sources: “Fast Start in cognitive innovation: Top performers share how they are moving quickly.” IBM Institute for Business Value. Janvier 2017. https://www.ibm.com/blogs/internet-of-things/fast-start-cognitive/ Unpublished data. C&A8. In general, where is your organization in its adoption of cognitive computing? Select the most advanced level for your organization; “Shifting toward Enterprise-grade AI: Confronting skills and data challenges to realize value.” IBM Institute for Business Value. Septembre 2018. https:// www.ibm.com/thought-leadership/institutebusiness-value/report/enterpriseai Données non publiées. AI1. In general, where is your organization in its adoption of artificial intelligence? Select the most advanced level for your organization; “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. Novembre 2020. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic# Unpublished data. S6. In general, where is your organization overall and your particular function in terms of adoption of artificial intelligence? 2022 Omdia AI Market Maturity survey https://omdia.tech.informa.com/OM023919/AI-Market-Maturity-Survey--2022-Database Q1. What is the state of AI deployment in your company? The rating scale in Omdia survey has been assumed to equivalent to IBM IBV rating scale in the following way: investigating technology and use cases = considering; Identified at least one use case and developing pilot = Evaluating; Currently piloting AI in at least one function or business = Piloting; Live AI deployment in at least one function or business unit = Implementing; Scaling AI deployment across multiple business functions or units = Operating/optimizing.

2 “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. 2020. https://ibm.co/ai-value-pandemic

3 “McCarthy, J; M.L. Minsky; N. Rochester; C.E. Shannon. “A proposal for the Dartmouth summer research project on artificial intelligence.” Consulté le 13 juillet 2022. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

4 Goodfellow, Ian; Yoshua Bengio, Aaron Corville. “Deep Learning.” The MIT Press. 2016. https://www.deeplearningbook.org

5 LeCun, Yann; Yoshua Bengio; Jeffrey Hinton. “Deep Learning.” Nature. 28 mai 2015. https://www.nature.com/articles/nature14539.pdf

6 Burns, Ed. “Timeline of AI winters casts a shadow over today’s applications.” TechTarget. Consulté le 13 juillet 2022. https://www.techtarget.com/searchenterpriseai/infographic/Timeline-of-AI-winters-casts-a-shadow-over-todays-applications

7 Thompson, Neil C.; Kristjan Greenewald; Keeheon Lee; Gabriel F. Manso. “Deep Learning’s Diminishing Returns.” IEEE Spectrum. 24 septembre 2021. https://spectrum.ieee.org deep-learning-computational-cost

8 Ibid.

9 Ibid.

10 World Health Organization malaria fact sheet. 6 avril 2022. https://www.who.int/news-room/fact-sheets/detail/malaria

11 “Fast Start in cognitive innovation: Top performers share how they are moving quickly.” IBM Institute for Business Value. Janvier 2017. https://www.ibm.com/blogs/internet-of-things/fast-start-cognitive/ Unpublished data. QC&A10 What are the important value drivers for cognitive computing? Select the top 5.“Shifting toward Enterprise-grade AI: Confronting skills and data challenges to realize value.” IBM Institute for Business Value. Septembre 2018. https:// www.ibm.com/thought-leadership/institutebusiness-value/report/enterpriseai Données non publiées. AI2. What are the important value drivers for artificial intelligence/cognitive computing? Select top 5. “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. Novembre 2020. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic# Unpublished data Q8. What are the most important value drivers for artificial intelligence? Select top 5.

12 “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. 2020. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic

13 Payraudeau, Jean-Stéphane; Anthony Marshall; Jacob Dencik. “Unlock the business value of hybrid cloud: How the Virtual Enterprise drives revenue growth and innovation.”IBM Institute for Business Value. 2021. https://ibm.co/hybrid-cloud-business-value. Payraudeau, Jean-Stéphane; Anthony Marshall; Jacob Dencik. “Extending digital acceleration: Unleashing the business value of technology investments.” IBM Institute for Business Value. 2021. https://ibm.co/hybrid-cloud-business-value

14 Fleming, Martin. “Breakthrough: A Growth Revolution.” Business Expert Press. 2022

15 Une bibliothèque NLP open source développée par Facebook AI. https://fasttext.cc

16 XGBoost est un algorithme ML d'ensemble basé sur un arbre de décision qui utilise une infrastructure d'amplification de gradient. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithmlong-she-may-rein-edd9f99be63d

17 McCormick a constaté que chaque tranche de 5 à 10 ans d'expérience réduit de moitié le nombre d'itérations.

18 The Flavor Extract Manufacturer’s Association of the United States. Les valeurs FEMA font référence aux ingrédients généralement reconnus comme sûrs et autorisés aux États-Unis. https://www.femaflavor.org/

19 VGG16 (également appelé OxfordNet) est une architecture de réseau de neurones convolutifs dont le nom est issu du Visual Geometry Group d'Oxford. https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html

20 Communiqué de presse Report Ocean. “Autonomous Vehicle System Market |(CAGR) of 39.1%| by Product Type, End-User, Application, Region – Global Forecast to 2030.” 14 juillet 2022. https://www.marketwatch.com/ press-release/autonomous-vehicle-systemmarket- cagr-of-391-by-product-type-end-user-application-region-global-forecast-to-2030-2022-07- 14?mod=search_headline

21 Kalra, Nidhi and Susan M Paddock. “Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?” Rand Corporation. 2016. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1400/RR1478/RAND_RR1478.pdf

22 Les GPU sont organisées en systèmes d'apprentissage en profondeur spécialement conçus (par exemple, NVIDIA DGX, qui comprend 8 GPU par serveur).

Compagnie IBM France, 17 Avenue de l'Europe, 92275 Bois-Colombes

Produit aux États-Unis Unis, août 2022

Les informations contenues dans ce document étaient à jour à la date de sa publication initiale, et peuvent être modifiées sans préavis par IBM. Les offres mentionnées dans le présent document ne sont pas toutes disponibles dans tous les pays où IBM est présente.

LES INFORMATIONS CONTENUES DANS LE PRÉSENT DOCUMENT SONT FOURNIES «EN L'ÉTAT», SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DÉCLINE NOTAMMENT TOUTE RESPONSABILITÉ RELATIVE À CES INFORMATIONS EN CAS DE CONTREFAÇON AINSI QU'EN CAS DE DÉFAUT D'APTITUDE À L'EXÉCUTION D'UN TRAVAIL DONNÉ. Les produits IBM sont garantis conformément aux dispositions des contrats.

Ce rapport fournit des orientations générales uniquement. Il n'est pas destiné à se substituer à une étude détaillée ou à l'avis d'un professionnel. IBM ne sera en aucun cas responsable de tout dommage résultant de l'utilisation de ce document.

Les données utilisées dans le présent rapport peuvent provenir de sources tierces et IBM ne procède à aucune vérification, validation ou audit indépendants de ces données. Les résultats de l'utilisation de ces données sont fournis « en l'état », sans aucune garantie explicite ou implicite.

Ce document est imprimé sur du papier recyclé post-consommation exempt de chlore sur une imprimante ayant la certification Forest Stewardship Council (FSC) Chain of Custody et utilisant des encres biologiques. L'énergie utilisée pour fabriquer ce papier et réaliser cette impression a été générée par une énergie verte renouvelable. Merci de recycler ce document.

6PQKYZ12-FRFR-01