Juin 2020

Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier

Document de réflexion

AUTEURS
Laurent Dupont, Olivier Fliche, Su Yang
Pôle Fintech-Innovation, ACPR

ACPR
BANQUE DE FRANCE

Table des matières

En première page de ce document : l'ordinateur CSIRAC, l'un des cinq premiers mis en service, sous la supervision de son co-concepteur Trevor Pearcey (photographie d'archives du 5 novembre 1952).

1. Résumé

Ce document de réflexion s'inscrit dans le cadre des travaux menés par l'ACPR sur l'intelligence artificielle (IA) depuis 2018. En mars 2019, après un premier rapport et une première consultation publique, l'ACPR a lancé des travaux exploratoires avec quelques acteurs du secteur financier afin d'éclairer les enjeux d'explicabilité et de gouvernance de l'IA – au sens essentiellement de *Machine Learning* (ML). Composés d'entretiens et d'ateliers techniques, ils couvraient trois domaines : la lutte contre le blanchiment et le financement du terrorisme (LCB-FT), les modèles internes et en particulier le *scoring* de crédit, et la protection de la clientèle. Deux axes d'étude en sont ressortis : ceux de l'évaluation et de la gouvernance des algorithmes d'IA.

Évaluation

Notre analyse conduit à identifier, dans l'évaluation des algorithmes et des outils d'IA en finance, quatre critères interdépendants :

1. **Le traitement adéquat des données** est un principe essentiel de tout algorithme. Il conditionne non seulement sa performance mais en assure également la conformité réglementaire et doit prendre en compte les considérations éthiques (telles que l'équité dans les traitements ou l'absence de biais discriminatoire).
2. **La performance** d'un algorithme de ML est une notion couverte par un ensemble de métriques suffisant pour évaluer l'efficacité d'un algorithme en finance selon les critères techniques ou fonctionnels souhaités. Il est parfois nécessaire de faire un arbitrage entre ces critères de performance et l'exigence d'explicabilité souhaitée.
3. **La stabilité** décrit la robustesse et la résilience du comportement d'un algorithme de ML au cours de son cycle de vie. Il convient notamment de garantir le caractère généralisable de l'algorithme lors de sa mise en œuvre et de détecter en permanence les risques de dérive des modèles déployés en production.
4. **L'explicabilité**, liée aux concepts de transparence ou d'interprétabilité algorithmique, est une notion qu'il convient de replacer chaque fois dans un contexte particulier pour en préciser la finalité. Une « explication » du résultat ou du fonctionnement d'un algorithme peut s'avérer nécessaire pour les utilisateurs finaux (clients ou utilisateurs internes) ; dans d'autres cas, elle sera destinée aux responsables de la conformité et de la gouvernance de ces algorithmes. L'explication fournie peut ainsi viser à éclairer le client, à garantir la cohérence des processus dans lesquels des humains prennent des décisions, ou encore à faciliter la validation et la surveillance des modèles de ML. Nous proposons quatre niveaux d'explication (observation, justification, approximation, réplication) afin de clarifier les attendus en matière d'explicabilité de l'IA en finance en fonction de l'audience visée et du risque associé au processus métier.

Gouvernance

L'inclusion d'IA dans les processus métiers en finance influe nécessairement sur leur gouvernance. Aussi nous recommandons de porter l'attention, dès la phase de conception des algorithmes, sur les aspects suivants.

Intégration dans les processus métiers. Il convient en particulier de déterminer si le composant d'IA remplace une fonction ayant un caractère critique (en raison de son rôle opérationnel ou du risque de conformité associé), et si son industrialisation est techniquement satisfaisante, selon une méthodologie appropriée au cycle de vie du ML (de sa conception à son *monitoring* en production).

Interactions entre humain et algorithme. Elles peuvent nécessiter une forme d'explicabilité particulière, soit à destination des utilisateurs internes en charge de confirmer les décisions de

l'algorithme, soit pour les clients qui doivent pouvoir être éclairés sur les décisions qui les concernent ou les propositions qui leur sont faites. En outre, l'intervention humaine parfois prévue dans les processus pour mettre en œuvre ou corriger les résultats des algorithmes, bien que souvent nécessaire et bénéfique, est source potentielle de nouveaux risques : ainsi, des biais peuvent être introduits dans une explication des résultats fournis par la machine, ou encore un humain peut avoir un sentiment de responsabilité plus fort lorsqu'il contredit l'algorithme que lorsqu'il le suit dans ses décisions.

Sécurité et externalisation. Les modèles de ML sont exposés à de nouveaux types d'attaques. Par ailleurs, les risques associés à l'externalisation des modèles, de l'hébergement ou des compétences techniques doivent également être évalués, de même plus généralement que les risques de tiers.

Processus de validation initiale. Les fonctions de validation initiale doivent souvent être repensées lors de la conception d'un algorithme basé sur l'IA et destiné à compléter ou modifier un processus existant. Par exemple, selon les cas, le schéma de gouvernance applicable à la ligne métier peut être conservé ou amendé pour la mise en production d'un outil d'IA.

Processus de validation continue. Une fois un algorithme de ML déployé en production, sa gouvernance présente aussi des enjeux nouveaux. Par exemple, son contrôle permanent nécessite une expertise technique en IA et un outillage dédié au *monitoring* de cette technologie, afin de garantir le respect continu des principes d'évaluation exposés plus haut : traitement adéquat des données, performance prédictive, absence d'instabilité, et validité des explications des décisions du système.

Audit. Quant aux missions d'audit – interne ou externe – de systèmes basés sur l'IA en finance, qui constituent une part essentielle de leur gouvernance, les travaux exploratoires menés par l'ACPR suggèrent l'adoption d'une approche duale.

- Le premier volet, analytique, allie analyse du code logiciel et des données utilisées, et méthodologie de documentation (si possible standardisée) des algorithmes, des modèles prédictifs et des jeux de données.
- Le second volet, empirique, repose sur l'utilisation de méthodes explicatives adaptées à l'IA (qui permettent de justifier une décision individuelle ou le comportement général de l'algorithme) et fait appel d'autre part à deux techniques permettant d'éprouver un algorithme en « boîte noire » : l'emploi de données d'évaluation dites de *benchmarking*, et la mise en concurrence du modèle étudié par un modèle dit « *challenger* » conçu par l'auditeur.

Si une telle approche est utilisable tant par un auditeur interne que par l'autorité de supervision, celle-ci fait face à des défis particuliers, en raison de l'étendue du périmètre de sa mission. Elle pourra les relever en acquérant une expertise théorique et pratique en science des données, et en se dotant d'un outillage approprié aux missions de supervision de l'IA.

Consultation

L'analyse exposée dans le présent document de réflexion est soumise à consultation publique. Le but est de recueillir l'avis des acteurs financiers et autres parties concernées par le sujet (chercheurs, prestataires, autorités de contrôle, etc.) sur les pistes de recommandations esquissées mais aussi, plus largement, tout commentaire utile, y compris sur l'adaptation des bonnes pratiques du superviseur.

2. Introduction

2.1. Méthodologie

Dans le prolongement de travaux initiaux suivis d'une consultation fin 2018 sur la place de l'Intelligence Artificielle (IA) en finance, le pôle Fintech-Innovation de l'ACPR a réalisé depuis mars 2019 des travaux de nature exploratoire avec quelques entreprises volontaires afin d'éclairer les enjeux d'explicabilité et de gouvernance de l'IA utilisée dans le secteur. Ce document présente les pistes de réflexion issues de ces travaux exploratoires. Le périmètre des technologies considérées dans ces travaux – et, partant, dans le présent document de réflexion – est précisé dans l'annexe « Périmètre technologique ».

Les acteurs du secteur financier sont, comme l'a montré la première consultation de l'ACPR, particulièrement demandeurs d'un éclairage réglementaire concernant ces nouvelles technologies[1]. Celles-ci sont en effet sources d'opportunités mais aussi de risques – opérationnels ou autres. C'est l'un des rôles des autorités de supervision que de fournir ces éclairages, accompagnés de recommandations pratiques de mise en œuvre, visant à un équilibre entre liberté d'innovation d'une part, conformité réglementaire et gestion raisonnée des risques d'autre part.

2.2. Travaux exploratoires

L'objectif principal des travaux exploratoires était de proposer des éléments de réponse à trois thèmes, tous en lien avec les principales missions de l'ACPR et détaillés dans ce qui suit.

Sur chaque thème, le pôle Fintech-Innovation a mené une exploration approfondie avec des acteurs volontaires, sous une double forme :

- dans tous les cas, entretiens de présentation des algorithmes d'IA en œuvre ainsi que des enjeux principaux d'explicabilité et de gouvernance ;
- dans le cas des ateliers dits « principaux », une phase plus technique impliquant les Data Scientists de part et d'autre, interagissant sur les méthodes et techniques mises en œuvre, et se concluant par des expérimentations et analyses du code logiciel développé par l'acteur.

Les ateliers de travail sont ici présentés succinctement ; ils sont détaillés (sous forme anonymisée) en annexe.

2.2.1. Thème 1 : Lutte contre le blanchiment et le financement du terrorisme (LCB-FT)

La question centrale sur ce thème était de savoir si l'IA peut améliorer la surveillance des transactions, en complément ou en substitution des règles de seuils et de gestion traditionnelles.

Pour ce faire, les acteurs ayant participé aux ateliers ont introduit des algorithmes de ML permettant de générer des alertes (en complément des systèmes classiques déjà en place, basés sur des seuils prédéfinis), alertes directement transmises au niveau 2 (équipes Conformité) pour analyse, ce qui permet de fluidifier et sécuriser le processus de traitement manuel. Le gain opérationnel est démontré, avec un impact notable sur la gouvernance du processus de déclaration de soupçon ou de gel des avoirs – impact lié aux changements du mode d'intervention humaine dans les processus LCB-FT et à la nécessité de surveiller le comportement du système au fil du temps.

[1] Voir également Cambridge Judge Business School, 2020.

2.2.2. Thème 2 : Modèles internes en banque et assurance

La question centrale sur ce thème était d'étudier comment et à quelles conditions l'IA peut être utilisée dans les modèles internes.

Plutôt que l'étude des modèles internes dans leur ensemble, les ateliers sur ce thème se sont centrés sur les modèles d'octroi de crédit, qui leur sont reliés dans la mesure où les scores produits par ces modèles sont utilisés pour construire les classes de risque sur lesquelles sont calculés les actifs pondérés en fonction des risques (ou *RWA*).

Ces travaux ont été menés avec deux acteurs différents : un grand groupe bancaire réalisant en interne la conception et la mise en œuvre de ses modèles de *scoring* de crédit, et un cabinet de conseil proposant une plateforme de construction de modèles avancés hybrides (ici testée sur le calcul de probabilités de défaut). Les deux scénarios ont montré l'impact de l'introduction de ML en termes de gouvernance : technicité accrue du processus de validation initiale, mise en place d'outils de monitoring pour le contrôle interne, et intégration de méthodes explicatives afin d'assurer non seulement le contrôle permanent mais aussi les missions d'audit.

2.2.3. Thème 3 : Protection de la clientèle

La question centrale de ce thème était de s'assurer que les algorithmes d'IA utilisés pour le conseil ou l'assistance à la vente de produits d'assurance non-vie garantissent la bonne prise en compte de l'intérêt du client.

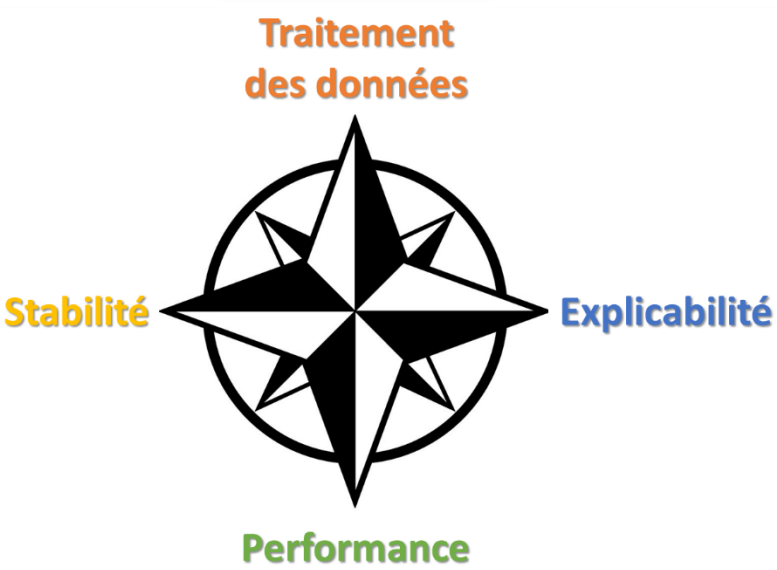
Le modèle de ML étudié concernait la préparation de devis pré-remplis en assurance habitation, avec comme exigences de conformité principale la bonne exécution du devoir de conseil en vue d'éclairer la décision du client, ainsi que la proposition d'un produit d'assurance non-vie cohérent avec les exigences et besoins exprimés par le client.

3. Principes de développement des algorithmes d'IA

Nous proposons ici quatre principes d'évaluation des algorithmes et outils d'IA :

- Traitement adéquat des données ;
- Performance ;
- Stabilité ;
- Explicabilité.

Ces principes représentent des objectifs liés entre eux par des relations nécessitant un arbitrage : il est en général impossible de maximiser les quatre objectifs simultanément. Ces quatre principes constituent donc en quelque sorte des points cardinaux permettant d'orienter la conception et le développement d'un algorithme d'IA :



4

3.1. Principe de traitement adéquat des données

Bien que l'évaluation de la conformité d'un algorithme et de sa mise en œuvre couvre un champ plus vaste, les traitements de données intervenant à chaque étape de la conception et de l'industrialisation d'un algorithme d'IA en constituent le point nodal, notamment sous les aspects suivants :

Données traitées

Il convient d'examiner quelles données de référence, d'apprentissage et d'évaluation sont utilisées pour nourrir l'algorithme. Cet aspect est parfois régi par une réglementation sectorielle, par exemple les exigences de complétude et de qualité des données de risques édictées par la norme prudentielle bancaire BCBS 239.

La gouvernance des données (Dai, 2016) constitue une fonction essentielle dans toute organisation dont certains processus métiers sont axés sur l'exploitation de données (« *data-driven processes* »). La mise en place d'une bonne gouvernance d'un algorithme d'IA est illusoire si les données traitées par cet algorithme échappent elles-mêmes à une gouvernance adéquate : il en est ainsi lorsque des algorithmes sont construits sur des données parcellaires, anecdotiques, ou manipulables, que l'établissement financier ne contrôle pas et dont la pérennité n'est pas assurée.

Prétraitements

L'évaluation doit aussi porter sur les traitements réalisés sur les données préalablement à l'application de ML proprement dit. Ces prétraitements peuvent avoir un impact sur la performance de l'algorithme (par exemple via le sous- ou sur-échantillonnage des données d'apprentissage), mais aussi sur son acceptabilité éthique (par exemple via l'exclusion de variables protégées).

Post-traitements

L'évaluation doit enfin inclure les traitements réalisés sur les prédictions ou les décisions du modèle résultant de l'apprentissage automatique. Ces traitements peuvent eux aussi avoir un impact significatif : c'est le cas des post-traitements opérant sur les modèles une fois entraînés et visant à supprimer ou réduire les biais[2] discriminatoires, par exemple en neutralisant la dépendance des prédictions d'un modèle probabiliste vis-à-vis de variables sensibles (Kamishima, 2012)[3].

3.1.1. Conformité réglementaire

Les aspects de conformité réglementaire comprennent :

- la conformité aux réglementations relatives à la protection de la vie privée ou des données personnelles, à commencer par le RGPD ;
- également la prise en compte des contraintes réglementaires spécifiques à un cas d'usage. Par exemple dans le domaine de l'assurance, l'interdiction d'orienter le processus de vente en fonction de la capacité à payer : l'offre doit au moins être cohérente avec les exigences et besoins du client, et non dictée par une possibilité d'optimisation du chiffre de vente de produits d'assurance.

La première catégorie peut être évaluée par des méthodes classiques et bien éprouvées : détection, prévention ou remédiation des biais (méthodes elles-mêmes applicables en prétraitement, post-traitement ou aux données sources), suppression des variables interdites (présentes explicitement ou implicitement), etc. La seconde catégorie de contraintes réglementaires, celles spécifiques à un secteur, dépassent souvent le cadre du traitement classique des données : ainsi de l'obligation de moyens en LCB-FT ou de résultat pour les gels des avoirs et embargos, qui nécessite des méthodes explicatives adaptées.

Un autre exemple permettra de préciser les enjeux de la réglementation sectorielle : celui d'un système de ML mis en œuvre dans un organisme d'assurance et visant à cibler les clients à contacter en priorité par les équipes marketing afin de leur proposer un contrat d'assurance multirisques. La DDA (Directive sur la distribution d'assurances) impose des principes assez proches de ceux d'équité évoqués dans la section suivante : les distributeurs de produits d'assurance doivent agir de manière honnête, impartiale, et au mieux des intérêts du client. Partant, le ML peut être autorisé pour du ciblage, mais les critères employés doivent être fondés sur les besoins auxquels répond le produit

[2] Il convient de noter le caractère polysémique du terme « biais ». Il désigne tantôt un biais statistique qui caractérise objectivement un modèle prédictif ou toute autre méthode d'estimation, tantôt un défaut d'équité ou une inégalité de traitement dont la polarité et l'importance sont subjectives et d'ordre éthique ou social. La présence d'un biais statistique peut engendrer un biais d'équité, mais ce n'est une condition ni nécessaire ni suffisante.

[3] La problématique de biais discriminatoire n'est en outre pas spécifique aux techniques d'IA. Ce risque existe dans tout modèle statistique, documenté par exemple dans la littérature déjà ancienne sur le « *redlining* » en économie bancaire. Toutefois le risque est parfois amplifié par l'utilisation d'un algorithme de ML, de plus certaines méthodes de détection et de mitigation de ce risque sont aussi spécifiques au ML.

proposé et non sur la capacité du client à y souscrire[4]. L'enjeu est donc pour le processus considéré de bien apprécier les besoins d'assurance de la clientèle prospective. Ces besoins sont bien évidemment plus difficiles à évaluer pour un algorithme que pour un humain, ce qui nécessite dans le cas du ML d'employer une profondeur et une variété de données plus importantes, engendrant ou accentuant les risques liés aux données : corrélations implicites (et souvent difficiles à détecter) avec la capacité à souscrire, ou plus généralement biais non souhaités (là aussi présents de façon implicite voire latente, cf. section suivante). La mise en place d'algorithmes de ML aux fins de ciblage marketing est donc conditionnée à la maîtrise de ces risques, et passe par la mise en place d'outils pour les détecter et les pallier.

3.1.2. Éthique et équité

Outre les contraintes imposées par les réglementations sectorielles et transverses, les enjeux éthiques sont au cœur de l'introduction croissante d'IA dans des processus métiers impactant des individus et groupes d'individus. Ces enjeux éthiques incluent :

- en général, les questions sociétales et d'éthique au sens large ;
- en particulier, les problématiques d'équité (terme que nous utiliserons en lieu et place de l'anglais *fairness*) induites par tout processus de décision automatique ou d'assistance par une machine à la prise de décision.

Pour illustrer à la fois l'importance des questions éthiques et les frontières floues qu'elles partagent avec les autres principes décrits dans cette section, il est intéressant de lister les recommandations en matière d'éthique publiées par le groupe d'expertise en IA de la Commission européenne (European Commission High-Level Expert Group on AI, 2019) :

1. action humaine et contrôle humain ;
2. robustesse technique et sécurité ;
3. respect de la vie privée et gouvernance des données ;
4. transparence ;
5. diversité, non-discrimination et équité ;
6. bien-être sociétal et environnemental ;
7. responsabilité.

Ces recommandations montrent le large spectre des enjeux liés à l'éthique et à l'équité en IA. Or un point d'attention en matière d'équité algorithmique concerne l'étude des biais, notamment à caractère discriminatoire, qui constitue un domaine de recherche actuellement très actif. Schématiquement, il s'agit :

- tout d'abord de bien définir les biais de nature problématique – qu'ils soient biais de classification ou de prédiction, ou biais statistiques non souhaités déjà présents dans les données – et les métriques permettant de caractériser et quantifier ces biais, y compris par des méthodes explicatives (Kamishima, 2012) ;
- de déterminer dans quelle mesure les biais présents dans les données sont reflétés, voire renforcés, par les algorithmes d'IA ;

[4] En effet l'élaboration d'un produit d'assurance passe précisément par la définition d'un marché cible, fondé sur les caractéristiques du groupe de clients pour lesquels le produit répond à des besoins.

- enfin d'appliquer une éventuelle remédiation soit au niveau des données, soit au niveau de l'algorithme.

Toutefois les travaux exploratoires réalisés par l'ACPR, même complétés par une étude plus générale du secteur financier, ont montré que seuls quelques acteurs du secteur financier avaient commencé à aborder la question de la détection et remédiation des biais de modèles. L'accent est pour l'instant mis sur la validation interne des solutions ainsi que sur leur conformité réglementaire, sans pousser l'analyse de l'équité algorithmique plus loin qu'elle ne l'était avec les méthodes traditionnelles – et notamment en ignorant souvent le renforcement potentiel des biais inhérents aux données. Cela ne fait toutefois que refléter le manque relatif de maturité de l'IA, ainsi que la priorité donnée jusqu'à présent aux processus non critiques (y compris en termes de risques éthiques et d'équité) : on peut dès lors prévoir que l'incorporation croissante d'IA en finance bénéficiera de la recherche en cours sur ces sujets.

TRAITEMENT ADÉQUAT DES DONNÉES

L'ensemble des traitements des données se doit d'être aussi bien documenté que le reste du processus de conception de l'IA (code source des algorithmes, performance des modèles produits, etc.).

Cette analyse permet d'évaluer les risques associés de conformité réglementaire et d'ordre éthique, et le cas échéant de mettre en œuvre des techniques de détection et de remédiation des biais algorithmiques.

3.2. Principe de performance

La performance d'un algorithme de ML peut être classiquement évaluée :

- soit par des métriques de performance prédictive : par exemple valeur AUC (complétée après fixation du seuil par le score F1 ou *a minima* par la matrice de confusion) pour un algorithme prédictif du risque de défaut de crédit d'une personne morale ou physique (on parle alors de KRI pour *Key Risk Indicators*) ;
- soit par des métriques de performance commerciale (KPI pour *Key Performance Indicators*) pour peu que ces métriques soient cohérentes avec les objectifs assignés à l'algorithme et compatibles avec les impératifs de conformité[5].

La performance algorithmique n'est toutefois pas un objectif indépendant des autres, et doit être mise en regard du principe d'explicabilité de l'algorithme d'IA. On verra dans les sections suivantes que le niveau d'explication requis dépend, pour un scénario donné, de plusieurs facteurs et des personnes destinataires de l'explication. Ce niveau d'explication induit des contraintes sur les choix techniques appropriés, notamment sur la « simplicité » de l'algorithme choisi.

Une présentation de l'arbitrage essentiel guidant ce choix technique figure dans l'annexe « Arbitrage simplicité/efficacité ».

[5] Ainsi, il est fort probable que l'optimisation du revenu de la vente de produits d'assurance soit une métrique inadaptée pour un algorithme : le risque serait dans ce cas d'introduire dans le processus algorithmique les « conflits d'intérêt » que la réglementation vise à prévenir.

PRINCIPE DE PERFORMANCE

Sélectionner les métriques de performance, permettant d'évaluer l'efficacité technique ou commerciale de l'algorithme de l'IA, en considérant l'arbitrage nécessaire entre simplicité et efficacité de l'algorithme.

3.3. Principe de stabilité

Le principe de stabilité consiste à garantir la persistance de la qualité et des caractéristiques d'un algorithme au cours du temps. Les attentes en termes de stabilité sont d'autant plus importantes dans le cas du ML que le nombre de dimensions capturées afin de constituer l'ensemble des variables prédictives est en général bien plus important que dans des modèles prédictifs ou décisionnels traditionnels[6].

Trois sources majeures d'instabilité sont identifiées dans cette section. Il convient de noter qu'à ce stade, ces sources d'instabilité sont rarement prises en compte en tant que telles et dans leur ensemble par les algorithmes d'IA en production dans le secteur financier : cela est vraisemblablement dû au manque de maturité relatif des processus d'ingénierie et d'exploitation associées. Les causes d'instabilité du ML ne doivent toutefois pas être négligées, en raison des risques opérationnels et de conformité qu'elles engendrent, aussi quelques méthodes de remédiation sont suggérées ci-dessous pour chacune d'elles.

3.3.1. Dérive temporelle

La stabilité d'un algorithme de ML s'entend avant tout comme stabilité temporelle. En effet, la distribution des données peut changer suffisamment au fil du temps pour dégrader la performance de l'algorithme si celui-ci n'est pas ré-entraîné périodiquement, ainsi que d'autres caractéristiques (conformité réglementaire, absence de biais, etc.)

Cette dérive temporelle peut être détectée au moyen de méthodes de *monitoring* et de lancement d'alerte assez classiques, reposant toutefois sur une infrastructure éprouvée et des indicateurs de déviation adéquats. Un point important à cet égard est que la dérive temporelle d'un modèle est souvent liée à l'évolution de la base de données à partir de laquelle le modèle est entraîné, c'est pourquoi la première étape de la mise en place d'un outil de détection – avant même d'analyser le traitement algorithmique – consiste à détecter les changements structurels dans les données d'apprentissage.

[6] C'est même l'une des caractéristiques du *Big Data*, ainsi qu'une situation où des techniques telles que les réseaux neuronaux sont particulièrement efficaces. De façon générale, il est démontré que le pouvoir prédictif d'un modèle de classification croît avec le nombre de variables utilisées jusqu'à un certain point avant de se dégrader – phénomène appelé pic de Hughes (Koutroumbas, 2008) et associé à celui de « *curse of dimensionality* » (fléau de la dimension). La réduction de dimensionnalité est d'ailleurs un enjeu fréquent en ML (Shaw, 2009).

3.3.2. Généralisation

Le manque de stabilité peut aussi s'entendre comme défaut de robustesse, au sens où le modèle souffre d'un manque de pouvoir de généralisation[7]. La présence de ce défaut peut notamment n'avoir pas été détectée au moment de sa validation, par exemple parce que les jeux de test et de validation – aussi décorrélés soient-ils du jeu d'apprentissage (*out-of-time testing*, *out-of-distribution testing*) – peuvent en effet diverger des données réelles sur lesquelles l'algorithme est éprouvé en production.

Si ce défaut de généralisation peut être détecté et partiellement pallié durant la conception et le paramétrage du modèle, il est toutefois nécessaire de soumettre l'algorithme à un *monitoring* continu tout comme pour la détection de dérives temporelles, car rien ne garantit que la performance escomptée sera généralisée à des données qui n'ont encore jamais été rencontrées.

3.3.3. Réapprentissage

Enfin, le réapprentissage, périodique voire quasi-permanent, d'un algorithme ne résout pas tous les problèmes de stabilité, car il peut *a minima* conduire à une non-reproductibilité des décisions prises initialement sur une donnée. Cette source d'instabilité du modèle au cours de son cycle de vie et de réapprentissage a pour conséquence principale un défaut de déterminisme du système. Ce dernier peut s'avérer problématique lorsqu'une décision doit être reproduite (par exemple afin de satisfaire les droits à l'information et à l'opposition tels que prévus par le RGPD), éventuellement accompagnée d'une explication (par exemple au moyen d'une méthode explicative décrite dans la suite de ce document).

Cette source d'instabilité, lorsqu'elle ne peut être palliée en garantissant une fréquence de réapprentissage suffisamment basse, peut du moins avoir comme remède un archivage de l'ensemble des versions successives d'un modèle d'IA utilisé en production.

PRINCIPE DE STABILITÉ

Identifier les différentes sources d'instabilité susceptibles d'affecter les algorithmes d'IA développés dans l'entreprise au cours du temps.

Pour chaque source, déterminer les risques associés (opérationnels, de conformité ou autres) et mettre en place des méthodes proportionnées de détection et de remédiation.

3.4. Principe d'explicabilité

Le critère d'explicabilité est celui qui distingue tout particulièrement l'IA des processus métiers traditionnels.

[7] Le pouvoir de généralisation est, avec le biais prédictif, l'un des deux critères à arbitrer dans la conception et l'ajustement d'un modèle prédictif. Il est inversement proportionnel à la variance du modèle, on parle donc d'arbitrage biais-variance : un biais faible est généralement associé à une performance élevée sur les données d'entraînement et de test, tandis qu'une faible variance signifie que le modèle est généralisable à de nouvelles données.

3.4.1. Terminologie

Les notions d'explicabilité algorithmique, de transparence, d'interprétabilité et d'auditabilité sont intimement liées :

- La transparence n'est qu'un moyen (certes le plus radical) de donner à comprendre des décisions algorithmiques : elle traduit une possibilité d'accéder au code source des algorithmes, aux modèles qu'ils produisent. Dans le cas extrême d'une opacité totale, l'algorithme est dit fonctionner en « boîte noire » ;
- L'auditabilité caractérise la faisabilité pratique d'une évaluation analytique et empirique de l'algorithme, et vise plus largement à obtenir non seulement des explications sur ses prédictions, mais aussi à l'évaluer selon les autres critères indiqués précédemment (performance, stabilité, traitement des données) ;
- La distinction entre explicabilité et interprétabilité est âprement débattue, et ces débats sont évoqués dans l'annexe à ce document : le terme d'explicabilité est souvent associé à une compréhension technique et objective du fonctionnement d'un algorithme (et serait donc adapté à la perspective d'une mission d'audit), tandis que l'interprétabilité semble davantage liée à un discours de nature moins technique (et viserait donc avant tout le consommateur ou l'individu impacté par l'algorithme).

3.4.2. Objectifs

L'explication algorithmique vise généralement à répondre aux questions suivantes :

- Quelles sont les causes d'une décision ou prédiction donnée ?
- Quelle est l'incertitude inhérente au modèle ?
- L'algorithme fait-il les mêmes erreurs que l'humain ?
- Au-delà de la prédiction du modèle, quelle autre information est utile (par exemple pour assister l'humain dans la prise de décision finale) ?

Les objectifs sont donc multiples, car dépendants des parties prenantes :

- rassurer les experts métiers et les équipes en charge de la conformité ;
- faciliter la validation du modèle par les équipes de conception et de validation ;
- garantir la confiance des individus impactés par les décisions ou prédictions de l'algorithme.

Une présentation de l'arbitrage essentiel guidant le choix technique d'un algorithmique en fonction des propriétés de ses explications figure dans l'annexe « Arbitrage sobriété/fidélité ».

3.4.3. Caractérisation

Une explication idéale posséderait les qualités suivantes :

- **précise** : elle décrit aussi précisément que possible le cas considéré (pour une explication locale) et le fonctionnement exact de l'algorithme (qu'elle soit locale ou globale) ;
- **complète** : elle couvre l'ensemble des motifs et caractéristiques de la ou des prédictions en question ;
- **compréhensible** : elle ne nécessite pas d'effort exorbitant pour être correctement comprise par l'audience à qui elle est destinée ;
- **succincte** : elle est assez concise pour être assimilée en un temps raisonnable, en fonction des contraintes de temps ou de productivité du processus où elle s'inscrit ;
- **actionnable** : elle permet une ou plusieurs actions de la part d'un humain, par exemple infirmer la prédiction en question ;

- **robuste** : elle demeure valable et utile lorsque les données sont changeantes et bruitées ;
- **réutilisable** : elle peut être personnalisée selon le type d'audience.

Bien entendu, certains de ces objectifs sont dans la pratique souvent mutuellement irréconciliables. En outre et comme détaillé par la suite, ils devront être mis en balance avec les autres principes – notamment celui de performance. Aussi ces objectifs serviront plutôt de critères de comparaison entre les explications fournies par différentes méthodes afin de choisir la méthode la plus appropriée à un cas d'usage bien spécifique.

3.4.4. Niveaux d'explication

On adoptera ici par souci de simplicité le terme d'explicabilité (plutôt qu'interprétabilité) comme le plus générique. Il consiste à comprendre, ou donner à comprendre :

- d'une part *comment* fonctionne un algorithme (ce qui correspond à l'acception courante de transparence) ;
- d'autre part *pourquoi* l'algorithme prend telle ou telle décision, autrement dit l'interprétation desdites décisions.

Un enjeu majeur de la question « comment » est l'auditabilité d'une solution algorithmique. Quant à la dimension explicative du « pourquoi », les enjeux associés incluent :

- la compréhension du comportement du système par les opérateurs humains qui interagissent avec lui ;
- la compréhension par le client qui est le sujet auquel s'appliquent les décisions ou prédictions de l'algorithme ;
- l'acceptabilité sociale ou éthique de la solution considérée, par exemple afin de prouver l'absence de biais (implicites ou explicites) à caractère discriminatoire dans les décisions prises par l'algorithme.

Le concept de niveau d'explication tente de résumer la profondeur d'une explication dans une seule métrique[8]. Cette métrique existe selon un continuum, au sein duquel nous retiendrons ici une échelle de quatre niveaux qualitativement distincts, figurés dans les encadrés suivants.

Explication de niveau 1 : observation

Elle répond sous un angle technique à la question : *« Que fait l'algorithme ? »*, ou sous un angle plus fonctionnel : *« À quoi sert l'algorithme ? »*. Ce niveau d'explication peut être obtenu :

- de façon empirique, par une observation des résultats produits par l'algorithme (individuellement ou en agrégat) en fonction des données d'entrée et de l'environnement ;
- de façon analytique, par une fiche descriptive de l'algorithme, des modèles produits et des données utilisées (voir annexe « Documentation des jeux de données »), sans nécessiter une inspection du code ni des données elles-mêmes.

[8] Ce concept est donc par définition simplificateur. Son bénéfice consiste à guider les concepteurs d'IA vers un niveau cible d'explicabilité, sans se substituer à une analyse multidimensionnelle des explications fournies.

Explication de niveau 2 : justification

Elle répond à la question : *« Pourquoi l'algorithme donne-t-il tel résultat (en général ou dans une situation précise) ? »*. Ce niveau d'explication peut être obtenu :

- soit par la présentation simplifiée d'éléments explicatifs issus de niveaux plus élevés (3 et 4), éventuellement assortis d'explications contrefactuelles (voir annexe « Explications contrefactuelles ») ;
- soit par la génération par l'algorithme lui-même de justifications obtenues par apprentissage (voir annexe « Méthodes explicatives conjointes à la modélisation »).

Explication de niveau 3 : approximation

Elle fournit une réponse, souvent inductive, à la question : *« Comment fonctionne l'algorithme ? »*. Ce niveau d'explication peut être obtenu, en sus des méthodes des niveaux 1 et 2 :

- par l'emploi de méthodes explicatives opérant sur le modèle étudié (voir annexe « Méthodes explicatives post-modélisation ») ;
- par une analyse structurelle de l'algorithme, des modèles et des données. Cette analyse sera d'autant plus fructueuse si l'algorithme procède par composition de plusieurs briques de ML (techniques ensemblistes, ajustement automatique ou manuel des hyperparamètres, méthodes de *Boosting*, etc.).

Explication de niveau 4 : réplication

Elle fournit une réponse démontrable à la question : *« Comment prouver que l'algorithme fonctionne correctement ? »*.

Ce niveau d'explication peut être obtenu, en sus des méthodes des niveaux 1 à 3, par une analyse détaillée de l'algorithme, des modèles et des données. Dans la pratique, cela n'est possible que par une revue ligne à ligne du code source, une étude exhaustive des jeux de données utilisés, et un examen de l'ensemble des paramètres du modèle.

Il est important de noter que chaque niveau caractérise une explication ou un type d'explication, et non un algorithme ou un modèle de ML. *Stricto sensu*, il s'agit du niveau d'intelligibilité de l'explication fournie, et non du niveau d'explicabilité intrinsèque de l'algorithme. Ainsi, un modèle très facile à expliquer tel qu'un arbre de décision peut se prêter à une explication de niveau 4 (en détaillant tous ses embranchements) mais aussi le cas échéant à une explication de niveau 1 (en expliquant qu'il s'agit d'un arbre de décision opérant sur tel ensemble de variables prédictives) s'il n'est pas utile ou souhaitable de dévoiler le fonctionnement détaillé du modèle.

Sous un angle plus technique, l'annexe « Obtention d'une explication de niveau élevé » examine plus en détail la faisabilité pratique d'obtention d'une explication de niveau élevé (3 ou 4), en présentant un obstacle important (les dépendances logicielles) ainsi qu'une piste pour atteindre le niveau 4 (réplication).

Nous décrivons dans la suite deux facteurs (parmi d'autres) influant sur le niveau d'explication requis d'un algorithme d'IA, particulièrement dans le secteur financier : d'une part l'audience à laquelle s'adresse l'explication, d'autre part le risque (nature et degré) associé au processus considéré. Ainsi, un même algorithme sera soumis à un niveau d'explication plus élevé lorsque l'explication devra être étayée par une compréhension interne du modèle et/ou que le contexte d'analyse sera particulièrement sensible.

3.4.5. Audience de l'explication

Le premier facteur influant sur le niveau d'explication attendu est l'audience à qui s'adresse une explication de l'algorithme. En effet, les différences de sophistication technique ou métier, mais aussi les motivations intrinsèques à un destinataire particulier du discours explicatif, influent sur la forme de l'explication qu'il est pertinent de proposer.

C'est pourquoi un même algorithme pourra être soumis à différents niveaux d'exigence selon que l'on considère un utilisateur final (qui cherche à s'assurer qu'il n'a pas été lésé ou traité injustement par le système, et pour qui une explication devra être directement intelligible) ou un auditeur (qui doit comprendre en détail le fonctionnement technique du système et qui est soumis à des exigences réglementaires fortes).

Une typologie de l'audience d'une explication, indiquant la forme d'explication à privilégier, est proposée ci-dessous.

Le client ou consommateur : explication simple

Par exemple, le devoir de conseil en matière de vente de produits d'assurance impose que les motifs de la proposition de tel ou tel produit, centrés sur la cohérence du contrat (en assurance non-vie) ou son caractère approprié (en assurance-vie), soient exposés au client prospectif.

La nature et les termes de cette explication doivent donc être intelligibles et satisfaisants vis-à-vis du consommateur (duquel il ne peut être exigé de maîtriser ni les arcanes du processus métier, ni l'implémentation de l'algorithme sous-jacent).

Le contrôleur interne : explication fonctionnelle

Les équipes de contrôle interne doivent par exemple s'assurer de l'efficacité du modèle vis-à-vis de la spécification d'objectifs métier.

L'accent est mis sur la performance du processus plutôt que sur sa mécanique interne, et l'explication associée doit être avant tout fonctionnelle.

L'auditeur : explication technique

L'auditeur doit s'assurer de la cohérence de l'implémentation de l'algorithme vis-à-vis de sa spécification, y compris la conformité réglementaire et le respect du cahier des charges techniques.

Cela consiste par exemple à valider comment est produit un modèle de ML, mais aussi de vérifier l'absence de biais à caractère discriminatoire dans le modèle obtenu, l'explication associée doit donc être techniquement détaillée et rigoureusement fidèle au modèle audité.

3.4.6. Risque associé

Le second facteur d'influence sur le niveau d'explication exigé est le nature et le degré de risque associé au remplacement partiel ou total d'un processus humain par de l'IA.

Ce risque associé est éminemment variable, par exemple :

- **En LCB-FT** : un processus tel que le processus de gel des avoirs, soumis à obligation de résultats, porte un niveau de risque accru non seulement en raison de sa criticité, mais aussi car son évaluation dépend de l'efficacité comparée des humains et de l'algorithme. On peut toutefois supposer que ce risque sera particulièrement élevé dans une phase de contrôle interne ou d'audit amenée à jauger de cette efficacité comparée, mais plus modéré pour un utilisateur quotidien de l'algorithme qui continue à opérer les mêmes vérifications que dans les processus traditionnels sur la base d'algorithmes usuels ;
- **En modèles internes** : l'introduction de ML dans le calcul des ratios de solvabilité d'un établissement bancaire a un impact direct sur l'évaluation de son risque de solvabilité, aussi les équipes en charge des modèles internes en attendront-elles un niveau d'explication satisfaisant ;
- **En assurance** : le processus de vente de contrats d'assurance est soumis à une réglementation propre, impliquant entre autres un devoir de conseil et des exigences de motivation personnalisée le cas échéant. À l'inverse, la segmentation de la clientèle *ex ante* en assurance repose essentiellement sur des objectifs d'efficacité, sans la même exigence d'explicabilité associée.

PRINCIPE D'EXPLICABILITÉ

Pour chaque cas d'usage, il importe de préciser le ou les processus métier impactés et la fonction remplie par le composant d'IA considéré.

Les différents types d'audiences visées par une explication de l'algorithme pourront alors être déterminés, et la nature des risques associés identifiée.

De ce contexte pourront être déduits le niveau et la forme d'explication requise – en concertation avec les parties prenantes à la gouvernance des algorithmes d'IA.

3.4.7. Exemples de niveaux d'explication par cas d'usage

Nous proposons ici d'illustrer ces définitions des niveaux d'explication et de leurs facteurs d'influence sur quelques cas d'usage concrets, tous adoptés et mis en production par des acteurs du marché – et certains étudiés lors de nos travaux exploratoires.

Pour chaque cas d'usage, le tableau suivant propose une position du niveau d'explication exigible en fonction des critères exposés (audience visée et risque associé). Ces suggestions sont basées sur une première analyse du marché – analyse dont ce document vise à valider ou corriger les résultats (voir « Consultation publique »).

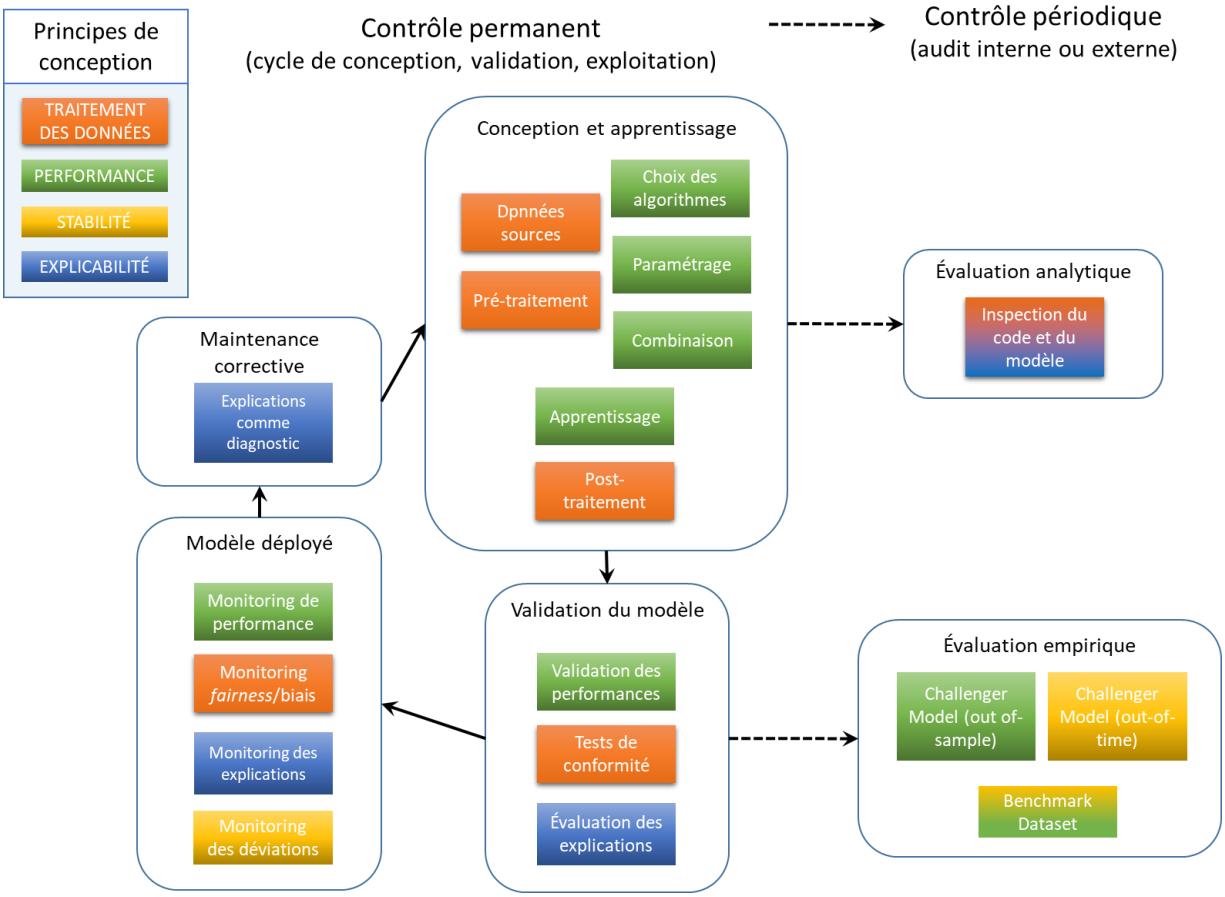
Cas d'usage			Critères d'explicabilité			Niveau d'explication requis
Domaine	Processus métier	Fonctionnalité de l'IA	Audience de l'explication	Contexte	Risque associé	
Contrats d'assurance	Gestion de contrat	Propositions d'indemnisation	Client	Processus d'indemnisation	Risque opérationnel (insatisfaction du client)	1
			Contrôleur interne	Vérification au quotidien du bon fonctionnement du processus	- Risque opérationnel - Risque de conformité (respect du contrat) - Risque financier	2
			Auditeur	Évaluation de l'algorithme	- Risque opérationnel - Risque de conformité (respect du contrat) - Risque financier	3
	Proposition de vente	Pré-remplissage de devis	Client	Demande de devis en ligne	Risque de conformité (mauvaise information du client, manquement au devoir de conseil, présence de biais discriminatoire...)	2
			Contrôleur ou auditeur interne	Évaluation de la conformité	Risque de conformité (mauvaise information du client, manquement au devoir de conseil, présence de biais discriminatoire...)	3
Modèles internes	Conception du modèle	Calcul des ratios de solvabilité	Équipe de validation	Validation des modèles, et de la politique de changement de modèle	- Risque de modèle (de solvabilité) - Risque de conformité	4
			Organes d'administration, de gestion ou de surveillance	Approbation	- Risque de modèle (de solvabilité) - Risque de conformité	2
Sécurité financière	Gel des avoirs	Remontée d'alertes	Agent en niveau 2	Analyse des alertes	Néant (si le comportement de l'analyste n'est pas modifié par l'existence de l'algorithme)	1
			Contrôleur de l'algorithme	Contrôle permanent	- Risque opérationnel (faux positifs et faux négatifs) - Risque de conformité (obligation de résultat)	2
			Auditeur	Contrôle périodique	- Risque opérationnel (faux positifs et faux négatifs) - Risque de conformité (obligation de résultat)	3

Exemples de niveaux d'explication (par exigence croissante : 1. observation, 2. justification, 3. approximation, 4. réplication). Cf. section « *Niveaux d'explication* ».

4. Évaluation des algorithmes d'IA

Le diagramme suivant représente l'ensemble des étapes du cycle de vie d'un algorithme d'IA, tant du point de vue de sa conception que de sa validation, permanente ou périodique, réalisée en interne ou en externe. Ce cycle de vie va donc de la phase de conception et d'apprentissage jusqu'à l'exploitation du modèle déployé en production, avec un retour itératif à l'apprentissage suite notamment à une maintenance corrective.

Il vise à montrer comment chacune des étapes du cycle de vie bénéficie d'un processus d'évaluation adéquat, reposant sur les quatre principes de traitement des données, de stabilité, de performance et d'explicabilité exposés précédemment. Il illustre aussi la combinaison d'une évaluation analytique et d'une évaluation empirique, approche multifactorielle détaillée dans la section « Audit des algorithmes d'IA ».



ÉVALUATION DE L'IA

Le cycle de vie des algorithmes d'IA introduits dans chaque processus métier mérite d'être détaillé afin de préciser, à chaque étape, quels principes de conception (traitement des données, performance, stabilité, explicabilité) s'appliquent particulièrement et selon quelle méthode se doit d'être évaluée la phase considérée.

5. Gouvernance des algorithmes d'IA

L'introduction d'algorithmes de ML dans le secteur financier vise, tant par des méthodes descriptives que prédictives, à automatiser ou améliorer – notamment en l'individualisant – la prise de décision auparavant effectuée par des humains. Partant, leur gouvernance nécessite de réexaminer la validation de ces processus décisionnels : en particulier, les impératifs de conformité réglementaire et les objectifs de performance ne sont respectés que par l'atteinte d'un niveau minimal d'explication et de traçabilité.

Dès la phase de conception des algorithmes, les questions de gouvernance méritent d'être prises en compte, notamment[9] l'intégration de l'IA dans les processus traditionnels, l'impact de cette intégration sur le contrôle interne (en particulier le rôle confié aux humains dans les nouveaux processus), la pertinence d'externaliser une partie de la conception ou de l'exploitation, enfin les fonctions d'audit interne ou externe.

5.1. Principes de gouvernance dans le secteur financier

Les principes généraux de gouvernance dans un établissement financier incluent la description de la politique de « culture du risque » déployée au sein de l'établissement, la présentation des normes éthiques et professionnelles promues par l'établissement, ou la description du dispositif mis en œuvre pour s'assurer de la bonne application de ces normes, et du processus en cas de manquement. S'y ajoute la documentation d'autres procédures telles que celle mise en place pour gérer et prévenir les conflits d'intérêts.

Dans ce contexte, les éléments de gouvernance les plus pertinents dans le cadre de l'introduction d'IA dans les processus métiers semblent les suivants : les procédures opérationnelles au sein de ces processus ; l'extension de la séparation des fonctions à la gestion de ces algorithmes d'IA ; et la gestion des risques associés à l'IA. On en rappelle ci-dessous rapidement les principes.

5.1.1. Procédures opérationnelles

Les procédures opérationnelles doivent être formalisées, adaptées aux différentes activités et régulièrement actualisées, sous forme par exemple de manuels de procédures. L'objet de ces procédures est de déterminer notamment les différents niveaux de responsabilités, les attributions dévolues et les moyens affectés au fonctionnement des dispositifs de contrôle interne, de décrire les systèmes de mesure, de limitation et de surveillance des risques et le mode d'organisation du dispositif de contrôle de la conformité. Ces procédures décrivent également les règles relatives à la sécurité des systèmes d'information et de communication et aux plans d'urgence et de poursuite de l'activité.

5.1.2. Séparation des fonctions

Il n'existe pas de norme de structure organisationnelle en termes de contrôle interne et de gestion des risques, tout au plus des méthodologies éprouvées lors de la mise en place de telles fonctions (COSO, Cobit, Risk IT, etc.). Néanmoins le dispositif standard du contrôle interne doit comporter *a minima*

[9] Une question de gouvernance non traitée dans le cadre de ce document est celle qui précède toute décision d'adopter ou non un outil, indépendamment de son utilisation d'IA et du domaine métier concerné, à savoir l'analyse bénéfice/coût. On s'attache ici aux enjeux de gouvernance spécifiques à l'IA dans le secteur financier.

plusieurs niveaux de contrôle permettant de répondre au principe du « double regard », soit de façon classique[10] :

- un contrôle de 1er niveau, au sein des différents services qui conduisent leurs activités et/ou exercent leurs responsabilités de façon maîtrisée ;
- un contrôle de 2e niveau, exercé par les responsables de services ou de directions, ou dans les organisations plus complexes par les services spécialisés de contrôle interne (également appelé contrôle permanent) ;
- un contrôle de 3e niveau, exercé par la direction de l'audit interne, qui a pour objet de s'assurer de la bonne mise en œuvre des dispositifs de contrôle en évaluant de façon périodique leur efficacité opérationnelle.

Il doit exister une stricte séparation entre les services qui engagent les opérations, et ceux qui enregistrent et contrôlent les opérations au fil de l'eau.

5.1.3. Identification et suivi des risques

Les entreprises doivent par ailleurs établir une cartographie des risques régulièrement actualisée et évaluée, permettant d'avoir une vue cohérente et globale des risques. Elles doivent aussi définir et promouvoir continuellement une culture du risque solide et cohérente traitant de la connaissance du risque et du comportement à adopter en matière de prise de risque. Elles doivent enfin mettent en place des systèmes et procédures assurant une analyse transversale et prospective des risques.

5.2. Intégration dans les processus

L'un des enjeux majeurs liés à la gouvernance des algorithmes d'IA est leur intégration dans les processus existants. Les facteurs essentiels à prendre en compte dans cette intégration sont le rôle joué par les algorithmes, le degré d'industrialisation obtenu, et la typologie des utilisateurs directs.

5.2.1. Rôle de l'IA

Le rôle joué par les briques d'IA dans le processus métier est très variable.

Il peut être d'ordre opérationnel, et donc jouer une fonction critique, comme illustré par l'atelier LCB-FT : le rôle du ML consiste à router certaines alertes dont le niveau de risque est particulièrement élevé directement vers le niveau 2 (Conformité), avec un risque opérationnel en cas de surcharge de ces équipes Conformité. La criticité de cette brique fonctionnelle est encore accrue par la contrainte d'opération en temps réel, la détection de transactions suspectes devant être signalée avec une latence aussi faible que possible.

À l'inverse, l'introduction de ML dans un cas d'usage comme celui de la sélection de clients prospectifs pour réaliser du démarchage commercial ou de la vente croisée n'a pas de caractère véritablement disruptif, et n'engendre pas de changement au cœur du processus métier.

[10] On se réfère ici à l'arrêté du 3 novembre 2014 pour les institutions financières. Des dispositions spécifiques concernent, par exemple, le contrôle interne de la gestion des placements par un organisme d'assurance (article R336-1 du Code des assurances).

5.2.2. Industrialisation de l'IA

Les objectifs d'industrialisation du ML diffèrent selon la nature du processus métier et l'utilisation faite des modèles. Deux exemples illustrent cette diversité de situations :

- Lorsque du ML est utilisé dans les processus de marketing (cas fréquent, même s'il n'a pas fait l'objet des ateliers décrits dans le présent rapport), une marge de manœuvre importante est accordée à la construction des modèles, qui est souvent itérative car les modèles sont déployés pour une utilisation ponctuelle afin d'alimenter par exemple une campagne marketing.
- À l'inverse, le ML utilisé dans les cas d'usage examinés lors des ateliers requiert une méthodologie de construction plus systématique, aboutissant en principe à une industrialisation au sens classique de l'ingénierie logicielle : automatisation de la construction, garantie de reproductibilité, processus d'assurance-qualité, et suivi des modèles mis en production (en particulier leur stabilité dans le temps). Les objectifs d'industrialisation sont donc plus ambitieux dans ce second cas.

En conséquence, le mode de conception du système d'IA peut aller du développement d'un outil à usage ponctuel (« *one-shot* »), en passant par un mode itératif, jusqu'à un processus continu (et donc complètement automatisable par une méthodologie d'intégration continue et de déploiement continu – couramment qualifiée de *CI/CD* dans l'industrie logicielle[11]).

Quant au mode de livraison du système d'IA, il peut lui-même varier entre une livraison manuelle où seuls les artéfacts finaux (à savoir les modèles de ML sélectionnés) seront retenus et exploités en production, et à l'extrême inverse une livraison de l'ensemble des données et résultats intermédiaires utilisés lors de l'exécution de l'algorithme et la construction des modèles. Un compromis entre ces deux approches est celui de service outillé proposé par le cabinet de conseil ayant participé au Sous-atelier « Probabilité de défaut » : il comporte d'une part la chaîne de conception qui adhère à une méthodologie systématique (quoique non automatisée) mais reste à la main du fournisseur de solution, d'autre part une plateforme de partage d'informations qui permet au client utilisateur du modèle de ML de procéder à une revue complète du processus de conception, et qui fournit une piste d'audit indépendamment de l'exécution de ce processus.

INDUSTRIALISATION DE L'IA

Planifier l'industrialisation de l'ensemble du cycle de vie des algorithmes d'IA. Selon le mode d'utilisation de l'IA, cette industrialisation peut nécessiter une méthodologie systématique, respectant les principes d'automatisation de la construction des modèles, de reproductibilité, de suivi d'un processus d'assurance-qualité, et du *monitoring* de la chaine de conception.

Dans tous les cas, il convient d'assurer la traçabilité de la chaîne de conception du système d'IA.

[11] CI/CD désigne une méthode générale de conception logicielle basée sur l'automatisation de toute la chaîne d'ingénierie logicielle, permettant de livrer au client des produits de manière plus fréquente que ne le permettent les méthodes traditionnelles sans sacrifier leur qualité. Cette méthodologie récente est étroitement liée aux méthodes agiles ainsi qu'à la tendance DevOps qui associe les fonctions de développement et d'exploitation informatique.

5.2.3. Prise en compte des utilisateurs directs de l'IA

L'impact de l'introduction d'IA sur le processus métier considéré dépend avant tout des utilisateurs directs de ses résultats (qui peuvent être des utilisateurs internes tels que les équipes de gestion ou de marketing, ou externes tels que les clients et prospects de l'organisation) – par opposition aux équipes chargées du contrôle interne ou externe du fonctionnement du processus, et qui interviendront par la suite.

En particulier, en fonction des étapes du processus organisationnel prises en charge par l'IA, le maintien de la qualité du processus nécessite de se demander si une forme particulière d'explication doit être fournie aux utilisateurs finaux, afin d'éclairer et motiver les décisions ou prédictions les concernant.

Typologie des utilisateurs directs

Ainsi dans le cas de l'intégration d'une brique de ML dans le processus LCB-FT (voir les détails sur cet atelier en annexe), les utilisateurs sont les équipes de niveaux 1 et 2 :

- il s'agit pour les équipes Conformité de réaliser un contrôle adapté à cette nouvelle approche (et donc avec une maîtrise de la technologie sous-jacente) ;
- la validation du modèle doit être bien plus fréquente que dans le cas des modèles internes de fonds propres par exemple, car les dérives peuvent ici avoir lieu en temps réel (par exemple un taux de faux positifs qui diverge de la norme), le *monitoring* du modèle doit donc lui-même être possible en (quasi) temps réel.

Dans le cas de l'Atelier « Protection de la clientèle », la fourniture de devis pré-remplis de souscription à un contrat d'assurance-habitation se destine aux consommateurs eux-mêmes, avec la contrainte d'accompagner les produits offerts de l'exposé de motifs, en adéquation avec les exigences et besoins du client.

Interactions humain/algorithme

Il est important pour l'utilisateur direct des résultats de l'algorithme, du moins lorsqu'il s'agit d'un utilisateur interne en charge de garantir l'exactitude et la qualité d'un processus métier, de conserver une indépendance vis-à-vis de la machine. En effet, l'expert humain est susceptible de repérer les erreurs manifestes de l'algorithme, ce qui a entre autres avantages celui de pouvoir contribuer à la performance et à la stabilité de celui-ci (deux des quatre principes de développement exposés dans ce document).

Il convient par ailleurs de noter que l'IA fournit un levier supplémentaire afin de vérifier l'absence de biais systématique ou de dérive dans les décisions prises, les propositions faites ou les conseils émis : c'est une situation où l'introduction de ML dans les processus permet de réduire le risque opérationnel.

Également, l'intervention humaine dans un processus décisionnel confié à un logiciel n'est pas anodine en ce qu'elle introduit un risque nouveau : le « revers de la médaille » de confier à un intervenant humain la possibilité de valider les décisions d'un algorithme est que cet humain peut voir sa responsabilité engagée, notamment dans les cas où il contredit le résultat de l'algorithme plutôt que de le confirmer. À l'inverse, l'humain modifie aussi parfois son comportement en présence de la machine : il pourrait avoir tendance à suivre systématiquement les instructions de l'algorithme,

préférant suivre ses erreurs – qui ne lui sont pas imputables – plutôt que d'engager sa responsabilité en le contredisant. Ces aspects d'indépendance et la responsabilité vis-à-vis de l'algorithme sont bien sûr liés au principe d'explicabilité, car il est nécessaire que l'humain puisse comprendre les principaux ressorts de la décision algorithmique pour lui opposer le cas échéant une autre décision, suffisamment éclairée.

Enfin, l'intervention d'un opérateur humain risque d'introduire un biais explicatif, souhaité ou non, lorsque cet opérateur fournit une explication déconnectée des facteurs sous-jacents à la décision ou au conseil prodigué par un algorithme : l'explicabilité de l'algorithme s'en trouve déformée ou neutralisée, en outre la transparence algorithmique n'opère plus, ce qui peut masquer certaines de ses défaillances. Une recommandation simple consiste alors à ne pas faire intervenir d'humain dans la détermination et la formulation des explications de l'algorithme.

PRISE EN COMPTE DES UTILISATEURS DIRECTS DE L'IA

Bien définir le périmètre et les modalités d'intervention humaine dans les processus utilisant de l'IA. En particulier, prévoir l'intégration de l'IA dans les processus métiers en fonction des utilisateurs finaux. Si ceux-ci incluent à la fois des utilisateurs internes et externes, expliciter les différences en termes de formes d'explication à fournir pour les résultats de l'algorithme.

Les résultats de l'algorithme peuvent devoir être soumis à validation humaine. Cette validation doit être encadrée par des règles métier documentées dans les procédures de contrôle interne, d'une part car elle engage la responsabilité humaine et d'autre part car l'algorithme peut induire un changement dans les comportements et jugements humains.

5.3. Contrôle interne

L'autre impact majeur de l'introduction de l'IA concerne la validation continue des algorithmes, et notamment les processus de contrôle interne.

5.3.1. Organisation du contrôle interne

Ainsi la surveillance des performances de l'algorithme, et notamment la détection de potentielles dérives, impose une conception différente de la chaîne de validation humaine : l'atelier LCB-FT a par exemple illustré comment le remplacement d'une partie des opérateurs de niveau 1 par un algorithme de ML peut induire une perte en capacité d'évaluation future de l'efficacité du processus, du moins en termes de faux négatifs (alertes non remontées par l'algorithme, correspondant donc à des transactions qu'un humain n'aura pas l'occasion d'analyser) : c'est pourquoi une partie de ces opérateurs ont été affectés à une fonction d'annotation manuelle en parallèle de l'algorithme, fournissant ainsi en continu de nouvelles données d'apprentissage.

En termes d'organisation du contrôle interne, un algorithme basé sur l'apprentissage est souvent introduit afin de remplacer tout ou partie des tâches accomplies par une équipe de niveau 1 (soumise au contrôle hiérarchique) et/ou de niveau 2 (réalisant des contrôles de conformité) – probablement pas celle de niveau 3 (en charge des audits internes), même s'il n'est pas exclu d'en arriver à ce stade d'automatisation dans un futur plus lointain. Il apparaît donc que l'utilisateur des résultats de l'algorithme n'est pas le responsable de son bon fonctionnement, encore moins son concepteur.

ORGANISATION DU CONTRÔLE INTERNE ET IA

Le contrôle interne de processus métiers impliquant des algorithmes d'IA doit autant que possible combiner des profils d'experts techniques et d'experts métiers : le contrôle du comportement de ces algorithmes nécessite à la fois la validation technique des composants en jeu au moment de leur mise en production, leur surveillance en continu, et la gestion appropriée des risques de conformité induits ou renforcés par une approche basée sur le ML.

5.3.2. Validation fonctionnelle initiale

Dans le cas de l'Atelier « Modèles de crédit », un processus de validation préalable à la mise en production d'un modèle a été défini, impliquant les équipes techniques de Conception et de Validation des modèles (au niveau local et global dans le groupe bancaire considéré) ainsi que les services Conformité et Risques.

En particulier, toute mise en production d'un modèle (nouveau modèle, ou résolution d'un problème identifié sur un modèle en production) nécessite de passer en Comité Risques au niveau du groupe bancaire, notamment afin d'approuver la stratégie choisie. L'utilisation de ML dans ces modèles doit donc être prise en compte et évaluée à travers l'ensemble de l'organisation, tant au sein des équipes techniques et des experts métiers que dans les comités de validation au niveau groupe.

VALIDATION FONCTIONNELLE INITIALE

Définir, au moment de la conception d'un algorithme d'IA, l'impact qu'il aura lors de sa mise en production sur le processus de validation.

L'ensemble des équipes de validation doivent être impliquées, depuis les fonctions conception et de validation technique jusqu'aux comités transverses concernés par les processus impactés par l'algorithmes, tels que le Comité Risques de l'entreprise dans l'exemple précédent.

Cette approche paraît intéressante et applicable à d'autres cas d'usage, selon des modalités variables en fonction des processus impactés et de leur criticité.

5.3.3. Validation fonctionnelle continue

Les travaux exploratoires relatifs au thème LCB-FT, où l'algorithme de ML effectue une détection d'anomalies aux niveaux 1 et 2, ont particulièrement bien montré comment il doit être soumis à un contrôle permanent plus sophistiqué que les méthodes traditionnelles, incluant notamment la vérification du bon calibrage de l'algorithme au fil de l'eau : volume d'alertes remontées au niveau 2, taux de faux positifs filtrés par la suite, etc.

Cette mise à niveau du contrôle permanent nécessite donc de la part des équipes qui en sont responsables :

- *a minima* la mise en œuvre et la maîtrise d'outils de monitoring du comportement des algorithmes en mode opérationnel (et éventuellement en temps réel) ;

- l'expertise adéquate pour détecter un dysfonctionnement de l'algorithme en amont, idéalement aussi pour le diagnostiquer et y remédier.

VALIDATION FONCTIONNELLE CONTINUE

La validation fonctionnelle d'algorithmes d'IA au fil de l'eau requiert à la fois un outillage dédié (tel que des tableaux de bord permettant aux responsables des processus impactés de veiller à leur bon fonctionnement général), et une interaction étroite avec les experts techniques responsables de leur conception et de leur validation initiale.

5.3.4. Le cas particulier des changements de modèle interne

Un aspect crucial des enjeux de validation induits par le recours à l'IA est celle des éléments déclencheurs d'une revalidation d'un modèle. En effet, contrairement à des systèmes experts essentiellement basés sur des moteurs de règles et autres paramètres de configuration prédéfinis, qui devront être revalidés lorsque ces paramètres auront été sciemment modifiés ou seront eux-mêmes jugés obsolètes, les modèles basés sur du ML sont essentiellement invalidés par suite d'un changement dans les données qui les alimentent. (On notera toutefois que ces modèles ne sont pas exempts de paramètres de configuration prédéfinis, notamment les hyperparamètres des algorithmes de ML, qui peuvent alors être soumis au même traitements que les modèles traditionnels.)

Dans le cas particulier des modèles internes (dits « bâlois » dans le secteur bancaire), une déclaration aux autorités de supervision s'impose lorsqu'un ajustement de paramètres dans ces modèles engendre un dépassement de seuil ou un critère de type similaire, clairement documenté et explicité dans la politique de changement de modèle soumise par l'établissement. Cet ajustement de paramètres étant généralement décidé et réalisé par des experts métiers en charge du modèle interne, on peut se demander ce que devient le facteur déclencheur d'une déclaration de changement de modèle si celui-ci est basé sur de l'apprentissage automatique.

Les modèles internes « classiques » supposent le calibrage de paramètres à partir de données, ce qui peut être rapproché de la notion d'apprentissage pour un modèle de ML.

Par ailleurs, la réglementation exige la mise en place d'une gouvernance couvrant notamment les aspects relatifs au *backtesting* et à l'actualisation des paramètres des modèles, l'actualisation étant de fait souvent liée aux résultats du *backtesting* et donc déclenchée par une évolution constatée des données.

La déclaration de changement de modèle doit être faite dès lors que le changement induit est jugé matériel par l'établissement, qui doit donc définir dans sa gouvernance le processus d'évaluation de la matérialité d'un changement, indépendamment de l'utilisation de ML ou non.

Enfin, la plupart des acteurs bancaires optent pour une stratégie de calibrage à date ; toutefois certains modèles internes « classiques » réalisent une réactualisation régulière et systématique de leurs paramètres (par exemple paramètres de volatilité dans un modèle de marché), de façon analogue là aussi aux modèles de ML. Ainsi, du point de vue de la politique de changement de modèle, il semblerait que les modèles de ML puissent être traités exactement comme des modèles internes classiques.

5.3.5. Validation technique

Une expertise technique est nécessaire à la validation de l'IA, typiquement sur toute la chaîne de Data Science :

- rôles de *Data Owner* et *Data Steward*, responsables respectivement de la gouvernance et de la qualité des données utilisées par les algorithmes ;
- *Data Engineers* et *Data Scientists*, garants du bon fonctionnement opérationnel des composants logiciels implémentant l'algorithme en question ;
- enfin *Data Analysts* en charge d'assurer la validation initiale et continue des résultats produits par l'algorithme.

Les étapes à couvrir incluent en effet :

- l'apprentissage des algorithmes ;
- de façon plus complexe, leur validation au fil de l'eau (non-régression de l'algorithme, absence de dérives du modèle, etc.) ;
- enfin la fonction (à valeur ajoutée bien plus élevée) de repérage des nouvelles sources de données ou caractéristiques à prendre en compte.

VALIDATION TECHNIQUE

La validation technique de l'introduction d'algorithmes d'IA dans les processus métiers nécessite le développement interne à l'organisation d'une large expertise en Data Science, qu'elle soit répartie au sein des autres équipes (modèle organisationnel bâti sur des profils à « double compétence ») ou qu'elle forme un pôle dédié.

Cette expertise technique doit être transverse (allant idéalement du Data Engineering jusqu'aux techniques de ML les plus avancées) et à plusieurs étages : compétences généralistes, application au secteur financier, et maîtrise des processus propres à l'organisation considérée.

5.3.6. Gestion des risques liés à l'introduction de l'IA

Le dispositif de contrôle interne sera nécessairement impacté par l'utilisation d'IA et ses évolutions seront intimement liées aux risques induits, selon le type d'intégration dans le processus (décision automatique ou aide à la décision) et la nature de ces risques (risque réglementaire ou juridique, opérationnel, ou financier).

Considérons par exemple le cas de l'IA au service du traitement des déclarations de sinistres, en aval de la chaîne de valeur de la distribution de produits d'assurance, scénario non étudié dans nos travaux exploratoires mais dans lequel le ML a connu un essor récent. Le cas d'usage typique consiste à introduire un algorithme de filtrage des déclarations, qui détectera des cas de fraude probable ou appliquera des critères d'exclusion. Ainsi les cas passant ce crible automatique donneront lieu à une proposition d'indemnisation (produite automatiquement ou pas), tandis que les cas refusés seront aiguillés vers des chargés de dossiers de sinistres. Il faut noter que l'enjeu principal est ici l'efficacité du processus de gestion des sinistres :

- le risque est ici de nature financière pour l'assureur, si le taux de déclarations donnant lieu à une proposition d'indemnisation augmente de façon indue ;

- il est de nature opérationnelle dans le cas où le nombre de dossiers refusés par l'algorithme augmente au point de surcharger les équipes de 2nd niveau en charge de traiter ces dossiers ;
- enfin, il peut être risque de conformité si cette surcharge est telle que le taux de contentieux augmente drastiquement.

À l'inverse, dans le cas de l'Atelier « Protection de la clientèle », l'enjeu est l'explicabilité du processus de conseil pour que le consommateur à qui s'adresse la proposition de produit d'assurance, soit éclairé avant sa décision.

Un aspect transverse à ces différentes considérations est la nécessité d'un mécanisme de secours ou de remédiation (dit *« Fallback Plan »*) spécifique pour remédier à un incident, dysfonctionnement majeur ou panne du composant d'IA, allant jusqu'aux mécanismes de continuité d'activité :

- si l'intégration dans les processus initiaux s'est faite de façon suffisamment modulaire et robuste, ce plan de remédiation peut simplement consister à revenir au processus initial le temps de réparer la défaillance ;
- à l'inverse, si l'introduction du ML a plus largement modifié la chaîne de traitements, le plan de secours sera nécessairement plus sophistiqué (et souvent plus complexe à mettre en œuvre car il doit lui-même avoir été validé).

GESTION DES RISQUES LIÉS À L'INTRODUCTION DE L'IA

Bien définir la nature des risques associées au rôle de l'IA dans le processus : risque opérationnel purement lié aux systèmes d'information, risque financier, risque de conformité, etc.

Inclure ces risques dans la cartographie des risques exigée par les principes de gouvernance, jusqu'aux mécanismes de remédiation d'incidents et de continuité d'activité.

5.4. Sécurité et externalisation

La sécurité de solutions reposant sur des algorithmes de ML nécessite la prise en compte de deux types de risques peu ou pas présents dans les solutions traditionnelles : le risque algorithmique spécifique au ML (qu'il s'agisse de disponibilité ou d'intégrité de l'algorithme), et celui associé au traitement des données par du ML. Une considération supplémentaire est la potentielle externalisation de la conception/réalisation ou de l'hébergement des solutions en question, laquelle présente des enjeux de sécurité spécifiques au ML.

5.4.1. Sécurité du ML

Les enjeux de sécurité du ML sont similaires à ceux de la sécurité des systèmes d'informations (SSI) traditionnels : ils ressortissent en général aux trois catégories de confidentialité, d'intégrité et de disponibilité. Ils doivent toutefois être traités de façon spécifique dans le cas du ML. En revanche, le traitement de ces risques n'a rien de spécifique dans le secteur de la finance, d'autant que la surface d'exposition y est bien moindre que dans d'autres secteurs : en effet la SSI y est généralement bien prise en compte, en outre le code en *open source* et surtout les données publiques y sont plus rarement employés.

On ne sécurise pas un modèle de Machine Learning comme on sécurise un service web proposé via une API REST[12], ni comme on sécurise les données sources sous-jacentes. Ces trois cibles d'attaque potentielles se situent sur trois couches d'architecture différentes (tout en étant mutuellement imbriquées) soit, de la couche la plus basse à la plus élevée : données, modèle, application.

La description exhaustive des failles potentielles d'un modèle de ML ainsi que des moyens de s'en prémunir sort du cadre de ce document. Une catégorisation des principales attaques possibles est toutefois proposée dans l'annexe « Recension des attaques contre un modèle de ML ».

SÉCURITÉ DU ML

Lors de l'étude de la sécurité associée aux processus utilisant de l'IA, il convient de ne négliger ni l'analyse de la sécurité des systèmes d'information (SSI) au sens classique, ni les failles et techniques spécifiques à ce domaine.

En particulier, toute étude d'impact, plan de remédiation ou audit de sécurité doit prendre en compte les attaques des modèles de ML, celles des données sous-jacentes (en termes de confidentialité, d'intégrité, de disponibilité), et celles des modules prédictifs ou d'aide à la décision dans leur ensemble.

5.4.2. Risque de tiers et externalisation

Les institutions financières ont recours à différents types de tiers prestataires pour développer leur IA : la conception (maîtrise d'ouvrage) et la réalisation (maîtrise d'œuvre) peuvent être confiés à une société externe ; certaines prestations d'achat sont de plus en plus couramment observées en IA ; enfin, l'hébergement et l'exploitation de services d'IA peuvent être externalisés chez un hébergeur traditionnel ou un fournisseur de solutions Cloud.

Enjeux associés

Les enjeux classiques liés à l'externalisation des compétences, de la maîtrise d'œuvre, et de la maîtrise d'ouvrage sont particulièrement prégnants en IA.

Ils constituent autant de défis difficiles à relever en pratique s'ils n'ont pas été suffisamment anticipés. C'est pourquoi la première des bonnes pratiques à respecter avant d'externaliser est de réaliser une analyse de risques *ex-ante* prenant en compte les aspects décrits ci-après.

Réversibilité

La réversibilité des solutions externalisées constitue une source potentielle de vulnérabilité importante – et non spécifique à l'IA – des institutions financières aujourd'hui : les orientations des autorités européennes de supervision (ABE, 2019 ; AEAPP, 2020) en témoignent.

[12] REST (pour *Representational State Transfer*) est une méthode architecturale couramment choisie pour les services exposés sur le Web. Une API (*Application Programming Interface*) REST est ainsi un moyen simple, standardisé et facile à sécuriser de concevoir et déployer un service Web.

L'usage de l'IA peut toutefois encore accroître l'importance de ces questions. La maîtrise de toute la chaîne de technologies, lorsque celle-ci est externalisée, représente en effet un véritable défi, notamment les points suivants :

- ML avancé de type prédictif (expertise en Data Science) ;
- code logiciel pas forcément ouvert ni bien documenté avec de multiples points d'intégration (expertise en conception et architecture logicielle) ;
- infrastructure hétérogène mélangeant serveurs dédiés et Cloud (et requérant une expertise en *DevOps*).

Même quand l'ensemble des compétences sont présentes, par exemple dans un grand groupe bancaire, elles peuvent se trouver dispersées dans des départements trop hermétiques entre eux pour être en mesure de ré-internaliser ce qui avait été livré (souvent d'un seul bloc) par des prestataires.

Prestations de conception et réalisation

L'externalisation de la conception et de l'implémentation d'une brique d'IA, comme dans le cas du Sous-atelier « Probabilité de défaut », induit de nombreux changements dans le processus métier. Les défis qui en résultent sont entre autres :

- la possibilité (en termes de ressources humaines et technologiques) de réaliser une validation technique sur du code non écrit par les développeurs internes à l'organisation ;
- la rédaction et la mise à jour d'une documentation aussi exhaustive que possible, satisfaisant notamment les exigences des processus de contrôle interne ;
- la prise en compte, dès la conception de la brique d'IA, de l'auditabilité de la solution en production, ce qui requiert une architecture et des modalités d'intégration logicielle particulièrement bien conçus ;
- enfin, la mise à disposition d'outils explicatifs sophistiqués et eux-mêmes bien documentés, afin de permettre l'explicabilité de résultats produits par un outil réalisé en externe.

Prestations d'achat

Les prestations d'achat présentent des risques similaires à l'externalisation, notamment les suivants : risque de dépendance, risque de non-reproductibilité des résultats, sécurité informatique du prestataire, capacité de service après-vente, capacités d'audit (pour autant qu'il soit pertinent, c'est-à-dire quand il s'agit d'achats réguliers à un même prestataire).

Le Sous-atelier « Probabilité de défaut » a permis d'aborder la question du risque de dépendance vis-à-vis du fournisseur de solutions ou de briques d'IA : le risque semble dans ce cas maîtrisable dans la mesure où le fournisseur propose à ses clients d'accéder à toutes les étapes, y compris intermédiaires, du modèle qui lui est livré. Il n'en revient pas moins au client de maîtriser les technologies en jeu pour limiter le risque de dépendance. On peut en outre mettre en garde contre le recours à des prestations d'achat de solutions ne répondant pas au risque de dépendance de façon satisfaisante, en particulier si le client n'est vraiment maître que du modèle final qui lui a été livré, et non de l'ensemble de la chaîne permettant de le régénérer – voire lorsque cette chaîne est non documentée ou même inaccessible. Ce manque d'information, tant sur les conditions de conception de l'algorithme que sur son fonctionnement, constitue non seulement un risque opérationnel mais peut en outre engendrer des difficultés de contrôle interne et d'audit.

On notera enfin qu'un risque tel que celui lié l'absence de maîtrise de la reproductibilité d'un modèle n'est ni nouveau ni spécifique à l'IA : ainsi, les modèles traditionnels auxquels l'IA vise à se substituer partagent parfois ce caractère non-reproductible des solutions d'IA externalisées, surtout s'ils utilisent des méthodes stochastiques de type simulations de Monte Carlo.

Hébergement sur le Cloud

Un mode d'externalisation de plus en plus fréquent dans le secteur financier est le recours à un Cloud public. Pour accompagner cette tendance, de premières orientations ont été publiées par les Autorités européennes de supervision dans les domaines bancaire (ABE, 2019) et assurantiel (AEAPP, 2020).

Ces orientations couvrent peu ou prou le même périmètre dans les deux domaines, à savoir : l'analyse de la criticité des processus métiers (et l'analyse d'impact), des exigences de documentation, le devoir d'information envers le superviseur, les droits d'accès et d'audit par l'institution financière mais aussi par le superviseur, la sécurité des systèmes d'information, les risques associés à la localisation des données, la sous-contractualisation, la planification d'urgence (y compris les plans de continuité des opérations) en cas d'incident et la stratégie de sortie de l'accord d'externalisation.

RISQUE DE TIERS ET EXTERNALISATION

Toute décision d'externaliser (la conception/réalisation, l'hébergement ou l'exploitation) ou de faire appel à un tiers pour tout autre type de prestation (achat ou fourniture de service) doit être précédée par une analyse de risques *ex-ante*, et prendre en compte les résultats de cette analyse, en particulier en matière de réversibilité.

Il importe en outre de respecter dans la relation avec le tiers quelques principes de gouvernance essentiels :

- assurer la documentation et la traçabilité des travaux réalisés afin de pouvoir les auditer si nécessaire ;
- garantir à l'institution financière la possibilité (technique et pratique, mais aussi juridique c'est-à-dire sur le plan de la propriété intellectuelle) d'accès au code source et aux modèles, même lorsque ceux-ci sont développés ou hébergés en externe ;
- offrir la même garantie au superviseur afin de rendre possible un audit couvrant les systèmes, le code logiciel et les données.

5.5. Audit des algorithmes d'IA

Les principes d'évaluation des algorithmes d'IA, exposés précédemment pour le contrôle permanent, restent valables pour un audit ponctuel de ces algorithmes, qu'il soit réalisé en interne (dans le cadre de contrôles périodiques) ou en externe (lors d'une mission de contrôle de l'autorité). Ainsi, un audit devra prendre en compte le contexte précis dans lequel a été développé l'algorithme et, en particulier, les processus métiers dans lesquels il s'intègre ou qu'il impacte d'une façon ou d'une autre. En fonction de ce contexte et de leurs objectifs, les auditeurs devront opérer les arbitrages déjà évoqués entre les différents critères d'évaluation des algorithmes ; et les méthodes d'évaluation elles-mêmes doivent être adaptées à des systèmes auto-apprenants. Pour ce faire, le niveau d'expertise en IA de l'équipe d'audit doit être suffisant, tout comme dans le cas, décrit plus haut, des fonctions de contrôle permanent.

5.5.1. Approche multifactorielle

La multiplicité des situations induites par les paramètres décrits dans ce qui précède (combinatoire entre type d'algorithme, audience visée et cas d'usage), ainsi que par les circonstances concrètes de chaque procédure de validation (accès au code et aux données plus ou moins possible, ressources techniques de validation mobilisables ou non, etc.), incitent à adopter une approche multifactorielle pour l'évaluation de solutions algorithmiques de ML.

Dans cette approche, méthodes analytique et empirique se conjuguent utilement.

Évaluation analytique

Un algorithme d'IA peut être caractérisé par son niveau maximal d'explicabilité, selon les quatre niveaux identifiés précédemment (« Niveaux d'explication »).

Les niveaux 1 (observation) et 2 (justification) ne permettent pas, en principe, de comprendre et d'interpréter le fonctionnement de l'algorithme « de l'intérieur », en analysant ses composants à différents niveaux de granularité : seule une interprétation non basée sur l'inspection de l'architecture interne est alors possible. En revanche, les niveaux 3 (approximation) et 4 (réplication) reposent sur une analyse structurelle ou détaillée de l'algorithme, et plus précisément du code source et du modèle produit.

Si la politique de risque de l'établissement concerné a pris en compte de façon adéquate le niveau d'explicabilité requis par le contexte d'utilisation de chaque algorithme d'IA, les audits dédiés à ces algorithmes devraient logiquement porter sur des algorithmes à enjeu significatif et donc à niveau d'explication élevé (3 ou 4). Dans ce cas, une évaluation analytique est possible et l'accessibilité de l'ensemble du code logiciel et de la documentation associée en est le prérequis le plus important.

Dans le cas où l'audit (tant interne qu'externe) d'un processus englobe un algorithme faiblement explicable (niveau 1 ou 2), la première étape de l'audit – en ce qui concerne l'algorithme - sera de vérifier que le faible niveau d'explicabilité est compatible avec le niveau de risque et les exigences de conformité du processus. Il pourra également évaluer l'algorithme et son impact sur l'efficacité du processus par les méthodes empiriques décrites ci-dessous.

ÉVALUATION ANALYTIQUE DU ML

Mettre en place, dès la conception des algorithmes d'IA dont l'enjeu le justifie, les méthodes et outils d'évaluation analytique adaptés. Ces méthodes peuvent recourir à des standards de documentation relatifs aux algorithmes, aux modèles produits et aux jeux de données utilisés, ainsi qu'à une revue aussi exhaustive que possible du code et des données.

Évaluation empirique

La solution algorithmique est dans ce cas considérée comme une « boîte noire » afin d'évaluer son fonctionnement depuis l'extérieur, c'est-à-dire en observant son comportement en fonction de diverses données en entrée (*input data*). Plusieurs approches sont possibles, nous en décrivons trois dans ce qui suit.

Méthodes explicatives post-modélisation. Ces méthodes, dites globales ou locales selon qu'elles visent à expliquer une décision ou l'ensemble des décisions possibles, opèrent sur les modèles pré-entraînés embarqués dans les algorithmes d'IA. Elles restent valables même lorsqu'il est impossible d'accéder à la documentation, au code et au modèle en question, et s'appliquent ainsi particulièrement aux algorithmes dont le niveau maximal d'explicabilité requis est le niveau 1 ou 2. Un panel non exhaustif de ces méthodes est fourni en annexe. Les ateliers réalisés par l'ACPR avec des acteurs du secteur, également détaillés en annexe, ont montré que ces méthodes explicatives étaient déjà couramment utilisées dans les phases de validation interne des algorithmes de ML, avant tout pour justifier de leur bon fonctionnement autrement que par de simples métriques d'efficacité. Il semble logique que ces méthodes explicatives soient aussi mises à disposition des acteurs externes responsables de l'évaluation de ces systèmes.

Jeu de données *« benchmarking »*. On fournit un jeu de tests destiné à éprouver l'algorithme. Ce jeu de données peut être soit synthétique, soit constitué de données réelles (anonymisées le cas échéant), soit hybride (typiquement au moyen d'un modèle génératif qui permet d'augmenter un échantillon de données initial dit de « *bootstrapping »*).

D'un point de vue technique, toute évaluation empirique de ce type nécessite des ressources dédiées en Data Science, et plus spécifiquement des profils de Data Engineering à même de bâtir des jeux de données et des infrastructures de *benchmarking*.

Mise en concurrence de modèles. On fournit un modèle de ML dit *« challenger »*, dont les prédictions ou décisions sont destinées à être comparées à celles produites par le modèle étudié.

Un point d'attention concernant la mise en concurrence de modèles de ML est sa faisabilité pratique : le développement de modèles *challengers* nécessite en effet une allocation importante de ressources (humaines et matérielles) et de temps. Ces contraintes sont notamment difficilement compatibles avec les missions d'audit de modèles telles qu'actuellement réalisées par le superviseur, qui consistent à analyser les propriétés des résultats du modèle en place et à vérifier leur cohérence vis-à-vis des attendus réglementaires. Elles n'ont donc pas pour but de construire un modèle de ML alternatif[13]. La section suivante suggère quelques pistes donnant au superviseur les moyens de déployer ce type de méthode, ambitieuse et complexe, d'évaluation empirique.

Il faut enfin noter que les métriques d'évaluation empirique peuvent être multiples et dépendre des cas d'usage, tant pour le *benchmarking* que pour les modèles *challengers*. Elles mettront l'accent tantôt sur les métriques d'efficacité (afin d'évaluer la performance algorithmique), tantôt sur le traitement réservé à certains segments de population (afin de détecter les biais discriminatoires), parfois encore sur les décisions concernant un point de données particulier, etc.

[13] Pour mettre en perspective l'effort demandé par cette construction de modèle alternatif, le développement d'un modèle de crédit par un organisme bancaire s'étale typiquement sur plusieurs années et implique des dizaines de personnes en interne, alors même que son périmètre se limite aux données de l'organisme.

ÉVALUATION EMPIRIQUE DU ML

Des méthodes d'évaluation empirique devraient être mises en place dès la conception des algorithmes d'IA, et incluses dans le processus d'assurance-qualité des modèles produits (tests de non-régression, tests fonctionnels, tests d'intégration).

Les méthodes explicatives sont un outil essentiel à l'évaluation d'algorithmes et de modèles de ML. Elles peuvent être implémentées dès la conception d'un algorithme, ou à l'inverse opérer sur les modèles préalablement entraînés ; en outre, certaines méthodes sont applicables à l'évaluation d'un modèle en « boîte noire ». Le choix d'une méthode explicative devrait prendre en compte le type d'algorithme, l'audience visée pour les explications (utilisateurs finaux externes, responsables du contrôle permanent, auditeurs ou superviseurs), et le risque inhérent au processus métier.

Les missions d'audit interne ou de supervision peuvent aussi employer des méthodes d'évaluation empirique telles que le *benchmarking* au moyen de leurs propres scénarios et jeux de données, ou la comparaison de leurs propres modèles prédictifs à ceux conçus par l'entreprise. Pour faciliter ces missions, il est souhaitable de prévoir des schémas de données et une architecture logicielle aussi modulaires et bien documentées que possible – ce qui constitue en soi une bonne pratique de développement logiciel.

5.5.2. Enjeux spécifiques aux superviseurs

L'approche multifactorielle de l'évaluation des algorithmes d'IA, telle que décrite ci-dessus, nécessite que le superviseur adapte ses outils, méthodes et données. En effet, une brique de ML ne s'analyse pas comme un processus traditionnellement confié à l'humain ou à des algorithmes procéduraux, et nécessite non seulement une expertise minimale en la matière (cf. section « Formation »), mais également des ressources significatives consacrées au développement d'outils – tels que les modèles *challengers* – et à la maintenance de jeux de données permettant de réaliser des contrôles efficaces.

Par ailleurs, la diversité des cas d'usage de l'IA et, pour un même cas d'usage, la diversité des modélisations possibles[14], imposent au superviseur de trouver, dans la mise au point de ses méthodes d'évaluation, un équilibre entre flexibilité, nécessaire pour supporter la diversité inhérente aux modélisations rencontrées, et formalisme, rendant possible une approche systématique (capacité de brancher un modèle « challenger » sur les données de l'organisme, et inversement de tester le modèle en place sur un jeu de données externe, etc.).

Travail sur les outils

Ce travail consiste à construire des briques logicielles et de Data Science afin de faciliter et accélérer les missions de contrôle.

Ces outils devraient permettre la production de modèles *challengers* décrits ci-dessus, à confronter à ceux fournis par les acteurs supervisés. Un obstacle à prendre en compte spécifiquement par le superviseur est la dépendance vis-à-vis de schémas de données hétéroclites entre acteurs de

[14] Dans le cas des modèles internes de risques, par exemple, la diversité des modélisations peut être vue comme un facteur de prévention des risques « moutonniers » et donc du risque systémique (cf. document de réflexion « Intelligence artificielle : enjeux pour le secteur financier », 2018)

l'industrie : la méthode de mise en concurrence impose de bâtir des modèles pouvant ingérer des données suivant un schéma particulier pour chaque acteur. La difficulté est similaire en sortie d'algorithme : ainsi parmi les approches décrites dans l'atelier LCB-FT, certaines produisent une classification catégorielle (niveau de risque faible/moyen/élevé) quand d'autres produisent un score numérique de suspicion.

Travail sur les données

Ce travail consiste à mettre l'autorité de contrôle en mesure de disposer et de traiter les données de différentes sources ouvertes ou fermées (données publiques, réglementation, rapports de contrôle, etc.) à différents niveaux (superviseur national ou européen).

Ces données devraient permettre la constitution de jeux de données pour le *benchmarking* des modèles produits par l'industrie : il s'agit de mesurer la performance des modèles, d'évaluer leur explicabilité sur des données quelconques, de détecter leurs dérives, etc.

La problématique liée au *benchmarking* rejoint celle exposée pour les modèles *challengers* : en l'absence de standard, le *benchmarking* nécessite de produire des données dans un format satisfaisant les schémas de données en vigueur chez chacun de acteurs supervisés, imposant là aussi un surcoût d'adaptation technique et méthodologique.

Formation

Afin de permettre et d'accompagner cette adaptation des méthodes, outils et données du superviseur, se pose la question de la pertinence d'une formation en Data Science : soit chez le superviseur – ce qui est en l'occurrence envisagé au sein de l'ACPR –, soit dans des instituts spécialisés (par exemple dans le domaine de la conformité en banque, de l'actuariat, etc.).

6. Consultation publique

Les répondants sont invités à illustrer leur réponse aux questions posées dans cette consultation dans les cas d'usage de l'IA, en particulier du ML, en vigueur dans leur organisation.

6.1. Contexte

QUESTION 1 : EXPÉRIENCE EN ML

- Quelle est la nature de votre connaissance ou expérience de l'IA en général, et du ML en particulier (recherche, Data Science, connaissance opérationnelle, etc.) ?
- Si vous répondez au nom d'une entreprise du secteur financier, quel est le niveau de familiarité et d'expertise de vos équipes, aussi bien techniques que métiers, avec l'IA ?

QUESTION 2 : MISE EN ŒUVRE DU ML *(QUESTION DESTINÉE AUX ENTREPRISES DU SECTEUR FINANCIER)*

- Quels algorithmes de ML sont en place dans votre organisation ?
- Pour chaque type d'algorithme, préciser les cas d'usage où ils sont utilisés, et le type d'environnement (développement, pré-production, production) ?
- Pour chaque cas d'usage, sur quels critères et quelle méthode d'évaluation a été fait le choix de l'algorithme retenu (performance pure, compromis explicabilité/efficacité, etc.) ?
- Quels sont les rôles respectifs des différentes équipes dans la conception et la mise en œuvre d'algorithmes de ML dans votre organisation (experts techniques, maîtrise d'ouvrage informatique traditionnelle, experts métiers, responsables conformité, etc.) ?

6.2. Principe d'explicabilité

Ce document a dégagé des exigences d'explicabilité sur la base des trois thèmes explorés, ainsi que sur celle d'une étude plus générale en matière d'IA – tant dans le secteur financier que plus généralement sur l'état de l'art de la recherche et dans les domaines les plus pertinents.

Plusieurs points restent à confirmer quant à la pertinence de ces exigences, qui forment l'objet de cette partie de la consultation.

QUESTION 3 : DÉFINITION DES NIVEAUX D'EXPLICATION

Les quatre niveaux d'explication ressortant de cette analyse (1 : observation, 2 : justification, 3 : approximation, 4 : réplication) sont-ils clairement définis ? Si non, préciser les motifs d'incompréhension.

QUESTION 4 : ADÉQUATION DES NIVEAUX D'EXPLICATION

Ces niveaux d'explication constituent-ils un ensemble de niveaux adéquat aux sens suivants :

- Ces niveaux couvrent-ils selon vous le spectre des applications actuelles ou futures de l'IA en finance, depuis une transparence totale (et donc une auditabilité facilitée) jusqu'à des algorithmes fonctionnant en boîte noire ?
- Le choix de quatre niveaux semble-t-il approprié (si non, en faudrait-il plus ou moins) ?

QUESTION 5 : EXEMPLES PRATIQUES DE NIVEAUX D'EXPLICATION

Le tableau présenté dans la section « Exemples de niveaux d'explication par cas d'usage » donne, pour quelques cas d'usage de l'IA dans le secteur financier, une suggestion de niveau d'explication adapté.

- Ces suggestions vous paraissent-elles adaptées? Si non, pour quelle raison ?
- Sont-elles adaptées à vos propres scénarios d'utilisation de l'IA (préciser ces scénarios) ? Si non, dans quel sens ?

6.3. Principe de performance

QUESTION 6 : MESURES TECHNIQUES DE LA PERFORMANCE

Quels commentaires suscitent de votre part les métriques techniques de la performance couramment utilisées (AUC ou score F1, score GINI, etc.), et notamment :

- Leur adéquation aux différents type d'algorithme d'IA ?
- Les méthodes adéquate de choix/sélection de ces métriques ?
- Les usages qui en sont faits (validation du modèle, choix du point de fonctionnement, détection de dérive, etc.) ?

QUESTION 7 : MESURES FONCTIONNELLES DE LA PERFORMANCE

- Quelles métriques fonctionnelles (ou « KPI ») vous paraissent-elles pertinentes ? Prennent-elles en comptes les aspects de conformité spécifiques aux processus concernés ?
- Par qui ces métriques devraient-elles être définies (équipes techniques, experts métiers, avec ou sans interaction avec les équipes conformité ou gestion des risques, …) ?

6.4. Principe de stabilité

QUESTION 8 : DÉRIVE TEMPORELLE DES MODÈLES

- Quels risques induit, selon vous, la dérive temporelle des modèles de ML ?
- Quelles sont les méthodes utilisées pour y remédier ou du moins les circonscrire (*out-of-time testing*, déclenchement d'alertes en cas de dérive, etc.) ?

QUESTION 9 : GÉNÉRALISATION DES MODÈLES

- Quelles sont les limites identifiées au pouvoir de généralisation des modèles de ML, qu'elles soient liées au sur-apprentissage (*overfitting*) ou à des limites inhérentes au modèle ?
- Comment peuvent-elles être traitées (*out-of-sample testing*, etc.) ?

QUESTION 10 : INSTABILITÉ DUE AU RÉENTRAÎNEMENT

- Les phases de réentraînement (en mode périodique ou continu) sont-elles une source d'instabilité des modèles, selon votre expérience ?
- Quelles sont les techniques utilisées pour la limiter (jeux de données ou tests de non-régression, etc.) ?

6.5. Principe de traitement adéquat des données

QUESTION 11 : CONFORMITÉ RÉGLEMENTAIRE EN MATIÈRE DE DONNÉES

Quelles méthodes ou techniques vous paraissent-elles indiquées pour s'assurer du respect des différentes contraintes réglementaires, notamment relatives au traitement des données :

- RGPD ?
- autres réglementations transverses ?
- réglementations sectorielles telles que la DDA ?

Préciser à quelle(s) phase(s) du processus (conception/apprentissage/prédiction) interviennent ces méthodes et techniques.

QUESTION 12 : DÉTECTION ET REMÉDIATION DES BIAIS

Quelles méthodes vous paraissent-elles indiquées pour analyser les biais, respectivement :

- dans les données sources utilisées par vos modèles de ML ?
- dans les algorithmes eux-mêmes ?
- dans les modèles produits par ces algorithmes et les décisions qu'ils engendrent ?

Plus précisément, quelles métriques d'équité (*fairness*) permettent d'identifier les biais, par exemple ceux à caractère discriminatoire ?

Quelles méthodes pour remédier aux biais non souhaités ainsi identifiés ?

6.6. Intégration dans les processus

QUESTION 13 : RÔLE DE L'IA

- Quelles sont ou devraient être, selon vous, les grandes lignes d'une méthode d'analyse des composants d'IA selon leur intégration dans les processus métier ?
- Que devrait-elle permettre d'évaluer : criticité de leur fonction, caractère disruptif vis-à-vis du processus traditionnel, gain de productivité, types d'interactions humain/algorithme, etc. ?
- Que pensez-vous du maintien de processus « parallèles » confiés à des humains pour évaluer en continu (ou corriger) les résultats de l'algorithme ?

QUESTION 14 : MÉTHODOLOGIE DE CONCEPTION DE L'IA

- La méthodologie de conception des algorithmes d'IA devrait-elle différer des modèles traditionnels, et notamment des méthodes standards d'ingénierie logicielle ? Si oui, en quoi ?
- Comment faut-il, selon vous, que la chaîne de conception, de sélection et de validation des algorithmes de ML prenne en compte l'intégration de ces algorithmes aux processus métier?

6.7. Contrôle interne

QUESTION 15 : GESTION DES RISQUES

- Quel est l'impact de l'introduction d'IA dans les différents processus métiers en matière de gestion des risques : présence de nouveaux risques associés ou amplification de risques existants (préciser de quelle nature : opérationnelle ou financière, juridique, etc.) ?
- Ces risques appellent-ils des mesures de gestion des risques spécifiques à l'IA (par exemple, calibrage des algorithmes de ML pour limiter l'exposition à tel ou tel risque) ?

QUESTION 16 : VALIDATION FONCTIONNELLE

- Quel doit être le processus de validation fonctionnelle avant mise en production ?
- La validation fonctionnelle devrait-elle être réitérée en cas de déploiement d'une version corrective ? Préciser si la réponse dépend du type de mise à jour.
- Comment doit être assuré le monitoring du ML en production sur le plan des risques métier ?

QUESTION 17 : POLITIQUE DE CHANGEMENT DE MODÈLE *(MODÈLES DE RISQUES INTERNES)*

- Selon vous, à quelles conditions les algorithmes de ML peuvent-ils être utilisés pour les modèles Bâlois ou les modèles internes en assurance ?
- Comment peuvent-ils être pris en compte de façon adéquate dans la politique de changement de modèle de l'établissement concerné ?

QUESTION 18 : VALIDATION TECHNIQUE

- Quel doit être le processus de validation technique avant mise en production ?
- Comment le monitoring du ML peut-il être assuré en production sur le plan technique ?

6.8. Sécurité et externalisation

QUESTION 19 : EXTERNALISATION

L'usage de l'IA introduit-t-il des enjeux ou des risques spécifiques lorsque son développement ou son exploitation sont externalisés ? Si oui, lesquels ?

QUESTION 20 : SÉCURITÉ

- Quel est l'impact de l'introduction du ML sur la sécurité des systèmes d'information (SSI) au sens classique ?
- Quels types d'attaques des modèles de ML vous semblent-ils les plus importantes, tant en termes de probabilité de survenue que d'impact en cas de succès (attaques causatives, attaques par substitution de modèle, attaques « adversarielles », etc.) ? Préciser selon le type de modèle de ML, selon le cas d'usage, et selon l'environnement (serveurs dédiés ou infrastructure externalisée sur le Cloud, etc.)

6.9. Approche multifactorielle de l'évaluation

Ce document de réflexion suggère la mise en place d'une approche multifactorielle dans l'audit des processus mettant en œuvre de l'IA. Plusieurs questions permettront ici de préciser cette approche.

QUESTION 21 : ÉVALUATION ANALYTIQUE

- Quels éléments parmi les suivants sont, selon votre expérience, disponibles pour l'évaluation des algorithme d'IA : le code source ? la documentation ? les modèles résultants ? les données d'apprentissage et de validation ? Précisez si la réponse dépend du type d'algorithme, du cas d'usage étudié, ou du contexte de l'évaluation (validation interne, audit externe, etc.)
- Quelles méthodes (standards ou non) de documentation du ML sont-elles utilisées pour décrire les algorithmes, les modèles et/ou les jeux de données ?

QUESTION 22 : ÉVALUATION EMPIRIQUE

- Laquelle des deux approches de Benchmarking ou de mise en concurrence de modèles (cf. section « Évaluation empirique ») vous semble-t-elle la plus appropriée ?
- La nature des architectures de Data Science au sein des organismes concernés est-elle suffisamment modulaire pour permettre ce genre de test fonctionnel au niveau des données ou de l'algorithme ?
- Le schéma de données est-il assez flexible pour se prêter à du *Benchmarking* sans faire reposer de contrainte d'intégration des données sur le superviseur ?
- De façon analogue, ce schéma est-il assez ouvert et documenté pour que le superviseur puisse mettre son propre modèle en concurrence sans pâtir d'une asymétrie informationnelle ?

QUESTION 23 : MÉTHODES EXPLICATIVES

- Quelles méthodes explicatives post-modélisation parmi celles décrites dans l'annexe 11 (recension des méthodes explicatives en IA) sont actuellement en production parmi les divers cas d'usage de l'IA dont vous avez connaissance ?
- Avez-vous connaissance d'autres méthodes d'explication algorithmique que celles décrites ici ? Si oui, lesquelles ? Ont-elles déjà été mises en pratique ?
- Les méthodes explicatives utilisées varient-elles en fonction du type d'algorithme ?
- Varient-elles selon l'audience de l'explication et, si oui, comment ?
- Varient-elles selon le niveau de risque associé au processus et, si oui, comment ?

Annexes

7. Périmètre technologique

L'IA, domaine vaste et dont la définition – basée sur les travaux de recherche et les pratiques industrielles – évolue au fil du temps, est ici considéré uniquement dans ses aspects pertinents pour le secteur de la finance, à la fois sous sa forme actuelle et dans ses évolutions plausibles à un horizon plus ou moins proche.

7.1. Distinction entre ML et IA

Ainsi, il s'agit toujours d'apprentissage automatique (ou *Machine Learning*, abrégé par la suite en ML), qui est par ailleurs le sous-domaine de l'IA le plus étudié de nos jours, par opposition aux autres formes d'IA : robotique, théorie des jeux, optimisation sous contraintes, systèmes multi-agents, représentation des connaissances (*Knowledge Representation and Reasoning*) ou encore planification automatique.

Parmi les méthodes de ML utilisées dans le secteur financier et considérées dans ce document, on peut sans être exhaustif citer les catégories suivantes :

- les méthodes d'apprentissage non supervisé (notamment techniques de *clustering*), qui ont leur application dans des scénarios de détection de fraudes ou d'anomalies ;
- les modèles prédictifs considérés comme « traditionnels », tels qu'arbres de décisions et régressions logistiques ou linéaires, ou plus sophistiqués et couramment mise en œuvre, tels que techniques ensemblistes à base d'arbres de décision (*Random Forests*, *Gradient Tree Boosting, etc.*) ;
- le NLP (*Natural Language Processing*, pour Traitement Automatique du Langage), employé pour classifier et analyser toute donnée textuelle ;
- le *Deep Learning* (réseaux de neurones profonds), employé dans plusieurs cas d'usage, y compris celui où ces algorithmes excellent le plus, à savoir la CV (*Computer Vision*, pour reconnaissance d'images) – même si ce cas d'usage est moins courant en finance que dans d'autres secteurs.

7.2. Distinction entre modèles et algorithmes

Un autre point terminologique important est la distinction entre un algorithme d'IA et le modèle produit par cet algorithme. Un algorithme de ML (comme indiqué plus haut, c'est le type d'IA considéré dans ce document) est une procédure exécutable et représentée sous forme de code logiciel, comme tout algorithme ; elle présente la particularité d'opérer sur des données en entrée (avant tout d'apprentissage, mais aussi de test) et de fournir en sortie un modèle de ML. Ce modèle est, de façon générique, constitué lui-même d'un algorithme prédictif et des données du modèle. L'algorithme est typiquement une procédure d'optimisation qui minimise une métrique d'erreur du modèle sur les données d'apprentissage.

Quelques exemples illustrent les relations entre modèle et algorithme de ML :

- un algorithme de régression linéaire produit un modèle constitué d'un vecteur de coefficients ;
- un algorithme de construction d'arbre de décision produit un modèle qui est un arbre dont les nœuds internes sont des conditions logiques impliquant les variables prédictives, et dont les feuilles des valeurs prédites ;
- un algorithme de réseau neuronal (basé par exemple sur une méthode et *back-propagation* et un algorithme de descente de gradient) produit un modèle qui est une structure de graphe dont les nœuds sont des vecteurs de coefficients.

Dans ce document, on utilise parfois indifféremment modèle ou algorithme lorsque le contexte est dénué d'ambiguïté et que l'on sous-entend à la fois le processus de construction du modèle par l'algorithme et le processus de prédiction mettant en jeu le modèle préalablement construit.

8. Présentation détaillée des ateliers

Cette annexe présente, pour chaque travail exploratoire mené sur l'un des trois thèmes retenus :

- une description de l'intérêt de l'exercice ;
- les objectifs du modèle présenté par l'acteur ;
- quelques détails techniques sur la méthode et l'algorithme ;
- le processus de validation de la méthode par l'acteur ;
- les enjeux de gouvernance engendrés par l'introduction d'IA dans le processus métier ;
- les méthodes et les enjeux d'évaluation de l'algorithme mis en œuvre, selon les quatre critères que nous avons retenus (explicabilité, performance, stabilité, traitement des données) ;
- le degré d'industrialisation obtenu pour le composant d'IA étudié.

Les développements ci-dessous ne constituent en aucune façon une évaluation des algorithmes étudiés au cours des ateliers, ni des processus dans lesquels ils s'insèrent[15]. Ils ont pour objectif de fournir des éléments de contexte factuels au lecteur, pour éclairer les enseignements généraux tirés par l'ACPR dans le présent document de réflexion.

8.1. Atelier « LCB-FT »

8.1.1. Contexte réglementaire

La réglementation LCB-FT en vigueur prévoit que les organismes financiers mettent en place un dispositif de gestion des risques leur permettant de détecter les personnes politiquement exposées (PPE), les opérations effectuées avec des personnes entretenant des liens avec un pays à haut risque listé par le GAFI ou par la Commission européenne, et les opérations incohérentes et inhabituelles selon la connaissance de la clientèle, pouvant faire l'objet d'une déclaration de soupçon (DS).

Les réglementations européenne et nationale relatives au gel des avoirs imposent également que les organismes financiers mettent en place une organisation pour la mise en œuvre des mesures de gel des avoirs et d'interdiction de mise à disposition des fonds.

Ces réglementations n'imposent pas l'utilisation d'un dispositif informatisé, mais en pratique, compte tenu de leur taille et de leur activité, la plupart des organismes en sont dotés.

A fortiori, la réglementation ne prévoit pas de disposition spécifique à l'utilisation de l'intelligence artificielle.

8.1.2. Intérêt de l'exercice

L'objectif de l'atelier LCB-FT, enrichi de séances complémentaires, a consisté à :

- comprendre quelles sont les applications potentielles d'IA en matière de LCB-FT ;
- s'acculturer autour des technologies d'IA sous-jacentes ;

[15] L'appel à candidature publié en mars 2019 précisait : *« Les travaux envisagés **ne s'inscrivent en aucune façon dans une démarche de contrôle de l'ACPR** : ils ne donneront donc pas lieu à des « suites » de la part de l'Autorité. Réciproquement, les candidats retenus s'abstiennent de se prévaloir d'une quelconque approbation de l'ACPR sur les algorithmes étudiés. »*

- réfléchir sur des adaptations éventuelles des superviseurs en vue de contrôler un processus où l'algorithme d'IA sera amené à être déployé.

8.1.3. Objectifs du modèle

Le principal projet étudié dans le cadre de cet atelier consiste à introduire des outils de *ML* pour assister le filtrage des messages transactionnels – en d'autres termes, concevoir des algorithmes afin d'assister les agents en charge de faire la part entre les faux positifs et les personnes figurant sur les listes de gels ou d'embargo, dans les listes d'alertes produites par un outil de marché basé sur des règles.

Dans le processus actuel, des opérateurs réalisent une revue des alertes issues du dispositif de criblage afin de déterminer s'il s'agit de faux positifs ou de personnes physiques ou morales visées par des mesures restrictives. Ces opérateurs sont répartis sur deux niveaux. Une équipe de niveau 1 est chargée du traitement initial des alertes, conformément à une grille de décisions. Les alertes qui ne sont pas résolues par le niveau 1 doivent être remontées au niveau 2 qui est autorisé à libérer le paiement, le rejeter, ou déclarer une homonymie auprès de l'autorité administrative en charge du gel des avoirs.

Le rôle du modèle développé est d'aider à la décision et de permettre d'adresser directement au bon niveau d'analyse les messages en fonction de leur pertinence, c'est à dire que les messages les plus sensibles seront traités directement au niveau 2, fluidifiant et sécurisant le processus. Le niveau 1, n'ayant plus en charge le traitement initial d'une partie des alertes, pourra alors absorber des augmentations de volume d'alertes. C'est le modèle TPA (*True Positive Acceleration*) développé aujourd'hui par l'institution.

8.1.4. Détails techniques

L'algorithme est basé sur un modèle de réseau de neurones qui s'alimente de caractéristiques de complexité variable : caractéristiques du message, distances phonétiques, décomposition des adresses (reconnaissance d'entités nommées) et sémantique des champs libres. Ces variables sont donc tirées des messages envoyés par l'outil de filtrage et ne contiennent pas de données personnelles contrairement aux messages envoyés initialement.

L'IA peut ainsi considérablement rationaliser ce travail. La possibilité de discriminer rapidement des ensembles de messages volumineux ne libère pas uniquement les analystes pour des tâches à plus forte valeur ajoutée, l'analyse des retours de l'IA leur permet également d'effectuer leur travail de manière plus approfondie. Le processus de prévision du risque devient beaucoup plus précis à mesure que la quantité de données analysées augmente. L'allégement des tâches répétitives sans forte valeur ajoutée et l'apport aux analystes de travaux stratégiques engageants amélioreront également la rétention du personnel.

Enfin, on peut même envisager le cas où l'IA contribue à améliorer les décisions d'un analyste par une analyse *a posteriori* des alertes libérées ou remontées par l'analyste, pour lui fournir un moyen de vérifier ses résultats afin d'ajuster les décisions sur les futures alertes.

8.1.5. Processus de validation

Le point de départ de l'acteur participant à cet atelier était de tirer parti des approches de validation de modèles existantes en matière de gestion des risques : le modèle pourrait ainsi constituer un outil pour le contrôle interne.

Les cadres usuels de ces modèles de gestion des risques comprennent des équipes de validation des modèles et des équipes de révision des modèles. Ces deux équipes sont indépendantes entre elles : une revue indépendante augmente le potentiel d'efficacité de l'algorithme et réduit son risque opérationnel.

L'objectif est de pratiquer une validation formelle annuellement et à chaque fois que le modèle subit une variation significative. Dans le même temps, des systèmes experts – éventuellement humains –, comme il est envisagé pour le projet étudié, fondés sur des règles devraient être utilisés pour fournir une base de référence permanente à laquelle comparer le modèle, et ainsi aider à identifier les cas où les décisions relatives à l'IA s'écartent des normes attendues.

La particularité des processus de validation dans le cadre du déploiement du ML est son caractère continu. En effet, il faut distinguer :

- d'une part, la validation de l'intégration au processus qui se fera en une seule fois, selon des schémas de validation conformes à la gouvernance générale de l'entreprise ;
- d'autre part, la validation statistique du modèle qui, en plus de la cohérence exigée avec la première forme de validation, demandera une approche continue dans le temps.

En d'autres termes, la notion de validation *a priori* est à réajuster puisque des cycles de validation plus courts sont nécessaires, ce qui rend la dichotomie entre validation initiale et contrôle interne permanent moins pertinente dans le cas d'un algorithme d'IA.

Dans tous les cas, le processus de validation doit être proportionnel aux risques, notamment règlementaires.

8.1.6. Enjeux de gouvernance

Le choix de gouvernance fait par l'acteur participant à cet atelier était d'assurer une double présence humaine dans la surveillance et la garantie du bon fonctionnement de l'algorithme : d'une part au niveau 2 qui intervient *a posteriori* et permet de s'assurer du bon fonctionnement de l'algorithme, d'autre part au niveau 1 pour annoter les transactions en parallèle de l'algorithme, ce qui alimente celui-ci en données d'apprentissage supplémentaires. Cette dernière approche n'a pas été retenue par l'ensemble des acteurs interrogés ayant introduit de l'IA dans leur processus de filtrage LCB-FT (voir « Atelier complémentaire ») ; elle permet toutefois de valider la performance de l'algorithme et sa stabilité au cours du temps, même en présence de changements majeurs dans le profil des données transactionnelles.

En termes de risque opérationnel, un point d'attention particulier provient de la baisse sensible de la charge du niveau 1 (de l'ordre de 10% en moyenne) engendrée par le recours au modèle de ML. Il est ainsi nécessaire d'anticiper le risque opérationnel induit par une potentielle interruption de l'algorithme ou par une défaillance plus générale du système : ce risque est critique car ses conséquences sont majeures, le composant d'IA contribuant ici à une obligation de résultat. Il convient

notamment de s'assurer que les équipes restent en mesure d'absorber, sans dégradation de la qualité du service rendu, le traitement de l'intégralité des messages en niveau 1 si nécessaire.

8.1.7. Évaluation : méthodes et enjeux

Explicabilité

Les contraintes d'explicabilité de l'algorithme sont ici différentes des deux autres ateliers (concernant l'octroi de crédit et la conception ou la distribution d'un produit d'assurance).

En effet, elles n'imposent pas de motiver les décisions envers un individu. Le contrôle de la pertinence de l'alerte remontée par l'algorithme est par ailleurs relativement simple à faire par l'analyste : pour faire son travail, et comparer l'alerte aux listes de sanction, celui-ci n'a pas besoin de connaître les « raisons » pour lesquelles l'alerte lui est remontée.

L'explicabilité présente ici avant tout un intérêt métier: il s'agit de faciliter l'analyse des comportements de traitement capturés par l'algorithme (et qui constituent ses données d'apprentissage). Cette aide à la compréhension des traitements réalisés par les analystes vient en appui de l'évaluation continue de la performance et de l'efficacité de l'algorithme, un enjeu crucial dans un domaine soumis à une obligation de résultat (en d'autres termes, il s'agit de ne pas omettre de « vrais positifs »).

Performance

Les performances statistiques du modèle prédictif et leur impact opérationnel sur le processus de traitement des alertes ont pu être évalués, et il en ressort que :

- Sur le plan statistique, le modèle présente un léger sur-apprentissage, qui ne semble toutefois pas de nature à induire un risque métier, compte tenu de la façon dont s'insère l'algorithme dans la chaîne de traitement (dans le plus mauvais des cas, les messages ne seront pas automatiquement accélérés, mais tout de même adéquatement traités par les équipes de niveau 2 le cas échéant);
- L'impact de l'algorithme sur le processus se traduit bien par une baisse sensible du nombre d'alertes à traiter par les équipes de niveau 1, ainsi que par une augmentation marginale du nombre d'alertes à traiter par les équipes de niveau 2 du fait de la meilleure expertise du modèle.

Stabilité

Le comportement du modèle semble stable dans le temps, au sens où l'impact relatif de l'accélération des messages sur la charge des équipes de niveau 1 et de niveau 2 est lui-même stable dans le temps.

Néanmoins dans ce cas d'usage, la qualité et la complétude des données sont essentielles, et leur « fraîcheur » indispensable pour s'assurer que le modèle qui les utilise ait un sens. Deux approches peuvent être utilisées pour s'en assurer :

- tenir compte de la temporalité qui joue un rôle primordial sur le sens des données, et la faire apparaître dans l'algorithme : les données utilisées en LCB-FT doivent en effet être revues afin de tenir compte des nouvelles méthodes employées par les individus mal intentionnés ;
- construire des variables génériques indépendantes du temps : par exemple, au lieu d'utiliser une variable « pays », utiliser « pays appartenant à une certaine liste de sanctions », qui est une caractéristique intemporelle en lien direct avec le problème analysé.

Traitement des données

Ce projet est directement issu de la direction de la Conformité, toutefois comme indiqué plus haut, une spécificité liée à l'utilisation de ML est que le processus de validation, outre l'équipe Conformité, repose à la fois sur l'expert métier et sur l'expert en données.

8.1.8. Industrialisation

Le projet – entrepris en mode agile – était au moment de l'atelier toujours en R&D. Comme indiqué dans la section « Processus de validation » , il paraît important de ne pas exiger à ce stade, même sur un sujet aussi sensible que la LCB-FT, un circuit de validation trop large qui impliquerait d'autres directions et inhiberait son industrialisation.

8.1.9. Atelier complémentaire

Un atelier a été mené sur le thème LCB-FT avec un autre acteur du secteur, également un groupe bancaire.[16] Seules les différences notables vis-à-vis du premier atelier sont ici résumées.

Objectifs du modèle

Le processus métier où intervient le modèle de ML est ici le filtrage de messages transactionnels, non pas pour les passer au crible de listes de sanction (conduisant à un éventuel rejet de paiement ou à un gel d'avoirs), mais pour détecter les transactions suspectes afin d'établir, le cas échéant, une déclaration de soupçon (DS). Cette tâche est accomplie par un progiciel spécialisé dans le filtrage des transactions financières, qui opère sur la base de règles métiers préconfigurées : ces règles sont appliquées à chaque transaction, afin de calculer un score. Ce score sert ensuite à aiguiller les transactions au-dessus d'un seuil de suspicion vers les équipes chargées d'analyser les alertes, décomposées selon la norme en niveaux 1 et 2 : les alertes au-dessus d'un premier seuil sont destinées au N1 (au niveau des agences), celles dépassant un seuil encore plus élevé au N2 (les correspondants Tracfin du groupe bancaire).

Dans la nouvelle approche, un modèle de ML est entraîné sur une base d'apprentissage constituée pour moitié environ d'alertes validées issues d'une déclaration manuelle, et pour le reste d'alertes générées par le progiciel sur la base de règles métier. Il est intéressant de noter que pour une proportion non négligeable des déclarations réalisées manuellement, le score généré par le moteur de règles métier est nul.

L'intégration du ML diffère de l'atelier LCB-FT principal en ce que le modèle de ML est ici introduit *en complément* du progiciel métier, sur plusieurs points :

- Le modèle de ML de l'atelier principal était en charge de remonter une partie des alertes déclenchées par les règles métier depuis le niveau N1 vers le niveau N2. Dans le cas présent, le modèle de ML produit des alertes supplémentaires envoyées directement en N2. Le ML intervient donc dans un processus parallèle, et non dans un processus en série où l'application des règles métier est suivie de la prédiction par le ML. Ainsi, plutôt qu'un outil de classification d'alertes déjà déclenchées, le groupe bancaire a mis en œuvre un outil de détection d'alertes validées qui s'applique à l'ensemble du flux de transactions.

[16] Cet atelier est présenté comme atelier complémentaire car il fut réalisé plus tardivement, en outre le cas d'usage de l'IA ainsi que sa mise en œuvre technique sont relativement similaires au premier atelier.

- Également, un filtre a été mis en place afin que, lorsqu'une transaction obtient un score élevé par le modèle de ML, une alerte ne soit générée que si aucune alerte n'a été levée par le moteur de règles métier sur ce même client dans les trois mois précédents. Autrement dit, une alerte déclenchée par l'outil de ML est une alerte ayant obtenu un score élevé et qui est passée sous les seuils de détection du moteur de règles métier dans les trois mois précédents.
- Enfin, contrairement au moteur de règles métier, le modèle de ML considère des éléments complémentaires aux données de transaction : il combine les valeurs statistiques des transactions conjointement avec des variables dites statiques (soit des mesures directes telles qu'ancienneté du client ou valeur du patrimoine, soit des variables construites telles que type de produits et contrats) sur une fenêtre temporelle glissante.

Enjeux de gouvernance

Contrairement à l'atelier LCB-FT principal, les équipes de niveau 1 ne procèdent pas ici à une annotation en parallèle de l'algorithme afin de détecter les faux négatifs laissés de côté par le ML : en effet selon eux, tout échantillon pertinent (ayant un nombre suffisant de faux négatifs) serait trop volumineux. Deux méthodes d'analyse des faux négatifs seraient envisageables, à savoir soit fixer un seuil plus bas au déclenchement d'alertes soit envoyer systématiquement les n cas les plus suspects, ce qui induirait à leur sens dans les deux cas une surcharge de traitement humain trop élevée. Par ailleurs, certains faux négatifs sont dus à des variables non observables.

Il convient de noter que ce mode d'introduction du ML dans le processus (en complément du progiciel métier), et le routage des alertes qu'il génère vers le niveau 2, entraînent bien sûr une surcharge des équipes N2. C'est pourquoi les équipes ont entamé un autre projet d'utilisation de l'IA dont le but est de router certaines alertes déclenchées par le progiciel depuis N2 vers N1, afin de diminuer cette surcharge induite.

Également en lien avec ces changements de processus métier, l'organisme bancaire participant à cet atelier complémentaire a adopté une organisation de son expertise dans le domaine LCB-FT centrée sur les « doubles compétences », maîtrisant à la fois le ML (y compris les questions du traitement des données) et les enjeux métiers ou risques.

Ces choix de gouvernance distincts entre les deux ateliers LCB-FT sont particulièrement intéressants à observer : chacune de ces options est probablement appropriée à son contexte particulier, et les retours d'expérience en la matière constitueront certainement un savoir-faire précieux concernant les arbitrages possibles entre la performance prédictive d'un algorithme de ML, sa stabilité au cours du temps, et la charge de travail dédiée à l'annotation manuelle par un humain.

Explicabilité

Les exigences d'explicabilité s'adressent à différent types d'utilisateurs du système. La réflexion conjointe des équipes techniques, conformité et MOA (maîtrise d'ouvrage informatique) du groupe a conduit à proposer des formes d'explicabilité adaptées à ce que chaque type d'utilisateur souhaite observer et dans quel contexte (en phase avec l'approche décrite dans la section « Audience de l'explication »). Ainsi :

- les équipes techniques (profils de type *Data Scientist*) utilisent les explications en phase de conception des modèles et non à des fins de *monitoring* en continu. Des valeurs de SHAP (*Shapley Additive Explanations*) sont la forme d'explication utilisée pour comprendre la décision sur une transaction donnée ;

- les experts Conformité les utilisent pour étayer leur décision d'abandon ou de validation d'une alerte. Des ateliers ont été organisés avec ces utilisateurs afin de bien cerner leurs besoins (l'exploitation de tableaux constitués de scores de SHAP était inadaptée), conduisant au développement d'une interface graphique leur présentant des explications toujours basées sur les valeurs de SHAP mais plus faciles à interpréter et exploiter ;
- enfin, le groupe vise aussi dans le futur à fournir des explications pertinentes aux auditeurs internes ou externes, incluant (en complément des deux formes d'explications précédentes) une documentation appropriée à leur bonne intelligibilité de l'algorithme.

Performance

L'indicateur de performance principal est le taux de déclaration, à savoir taux d'alertes produites par le système de détection qui donnent lieu à une DS. L'introduction de ML selon l'architecture décrite précédemment a permis de doubler ce taux de déclaration.

Stabilité

Un outil de monitoring a été mis en place dès le passage en production de l'algorithme de ML afin de détecter toute anomalie de fonctionnement ou dérive du modèle. Il suit plusieurs indicateurs concernant le modèle, les données en entrée, la distribution du score en sortie, etc.

Les équipes indiquent qu'il est encore trop tôt pour évaluer si les dérives du modèle de ML sont plus ou moins fréquentes que le besoin de reconfiguration du progiciel. La mise à jour du modèle de ML serait toutefois plus simple que le reparamétrage du moteur de règles métiers, pour plusieurs raisons : il s'agit d'un simple réapprentissage sans ajout de nouvelles variables, intégralement automatisé, et qui ajuste l'ensemble des paramètres du modèle sans intervention humaine. En outre, la durée de mise à jour du modèle de ML, depuis l'apprentissage jusqu'à la mise en production, n'excéderait pas 2 à 3 jours, soit nettement moins qu'une reconfiguration du progiciel.

8.2. Atelier « Modèles de crédit »

Le deuxième atelier concernait les modèles internes de risque et de calcul des fonds propres réglementaires. Dans les faits, les candidats aux ateliers ont proposé des cas d'usage proches mais ne coïncidant pas exactement avec ce thème.

En conséquence, cet atelier a « pivoté » vers la modélisation du risque de crédit, octroyé aux entreprises ou aux particuliers. Il a été scindé en deux sous-ateliers :

- l'un relatif aux modèles de crédit qualifiés « d'acceptation » : il s'agit essentiellement de modèles calculant un score d'octroi. Le sous-atelier a été réalisé avec un grand groupe bancaire ;
- l'autre relatif aux modèles de crédit qualifiés « de comportement » : ces modèles visent à estimer une probabilité de défaut à un horizon temporel donné pour un crédit en cours. Le sous-atelier a été réalisé avec un grand cabinet de conseil qui fournit à ses clients du secteur bancaire une solution de conception de modèles avancés.

8.2.1. Contexte réglementaire

Ces deux sous-ateliers partagent les constats initiaux suivants :

- Les modèles internes classiques sont en général aisément auditables mais peu performants. Des modèles plus avancés ou plus complexes permettraient d'améliorer cette performance, au détriment de l'explicabilité ;
- Or les exigences règlementaires imposées par le superviseur ou les experts métier sont identifiées comme des freins à la mise en œuvre d'algorithmes plus innovants, notamment ceux basés sur de l'apprentissage automatique : ces exigences concernent la stabilité des modèles obtenus, leur auditabilité, mais aussi la transparence et l'explicabilité des algorithmes en jeu ;
- Les obstacles liés à la protection des données personnelles, ainsi qu'à certaines limitations inhérentes aux données (en termes d'accès ou de complétude), rendent difficile l'analyse de corrélations entre les multiples variables caractérisant un client et son comportement.

8.3. Sous-atelier « Scores d'octroi »

8.3.1. Intérêt de l'exercice

Le groupe bancaire considéré a mis en place des recommandations méthodologiques de modèle de score d'octroi de crédit.

L'atelier a permis d'étudier comment les équipes ont pris en compte ces recommandations – définies de façon itérative sur une dizaine d'années – pour élaborer des modèles qui leur soient conformes.

8.3.2. Objectifs du modèle

Dans le domaine du score d'octroi de crédit, des ateliers ont permis d'étudier plusieurs modèles répondant à un double objectif :

- réduire la dépendance vis-à-vis de fournisseurs de données externes (de type *Credit Bureau*) via la prise en compte de nouveaux types de données internes dans les modèles : par exemple, données comportementales en plus des scores de *Credit Bureau* et des données internes traditionnelles telles que l'historique de crédit ;
- plus classiquement, améliorer le pouvoir discriminant des modèles de score d'octroi.

Les trois modèles étudiés concernaient respectivement l'octroi de crédit aux entreprises, le crédit à l'achat de véhicules d'occasion, et le crédit à l'achat d'équipement ménager.

Le modèle *Household Equipment* (concernant l'équipement ménager) est ici décrit plus en détail, les deux autres modèles présentent des enjeux similaires, tant fonctionnels que techniques. Son enjeu commercial est de fournir une réponse à la demande de crédit en moins de 5 minutes.

8.3.3. Détails techniques

Les données sources utilisées sont les suivantes :

- Données sur la demande de crédit
    - Données individuelles sur le demandeur et l'éventuel codemandeur
    - Informations sur le produit (montant, termes du crédit, etc.)
- Données risques liées au contrat
    - Données de calcul des états de défaut
    - Données de calcul des variables comportementales
- Données externes

- Scores de *Credit Bureau*
- Fichiers des banques centrales (pour la Banque de France : FCC, FICP).

Les équipes Data Science de l'acteur rencontré insistent particulièrement sur l'importance d'enrichir les données internes classiquement utilisées, par des données externes qui seront à terme, en plus de celles mentionnées ci-dessus, de nature diverse (texte, séries temporelles, etc.) et parfois issues de sources ouvertes (obtenues par *webscraping*). L'atout du ML consiste en cette exploitation de données (parfois appelées alternatives) et pas uniquement dans l'utilisation de nouveaux types d'algorithmes.

La plupart des modèles mis en œuvre par les équipes utilisent des méthodes de type *Gradient Tree Boosting*, après l'avoir confronté aux autres algorithmes couramment utilisés par le groupe (en particulier, SVM était trop consommateur en ressources de calcul et réseaux de neurones ont été jugés inadaptés au cas d'usage).

8.3.4. Processus de validation

Le parcours de validation au sein du groupe bancaire rencontré, pour tout modèle d'octroi développé avec du ML et avant sa mise en production (qu'il s'agisse d'un nouveau modèle ou de la résolution d'un problème identifié sur un modèle en production), est le suivant :

- Les équipes Crédit (locales, ou centrales lorsque les entités locales ne disposent pas des ressources techniques nécessaires, notamment en Data Science), qui ont développé le modèle, envoient à l'équipe Validation un dossier composé d'une documentation technique et de l'ensemble du code source.
- L'équipe Validation analyse la documentation (validation conceptuelle) mais font aussi retourner le code de génération du modèle (apprentissage, test et validation) afin d'en vérifier les résultats et d'apporter un regard critique sur les méthodes utilisées. L'équipe Validation a en effet toutes les compétences nécessaires pour évaluer le modèle selon les principes énoncés dans ce document (traitement des données, performance, stabilité, explicabilité). Cette phase de validation est détaillée ci-dessous.
- Pour certaines des entités seulement, les modèles d'octroi sont utilisés dans les modèles bâlois : dans ce cas, l'équipe Validation présente le modèle au Comité Risques du groupe mère, notamment afin d'approuver la stratégie choisie (à risque constant / diminution de risque / stratégie hybride).
- Le cas échéant, une fois validé par le groupe mère, le dossier est transmis à la BCE pour validation des modèles prudentiels.

Le processus de validation comporte donc des phases conceptuelles et des phases dites « appliquées ».

8.3.5. Enjeux de gouvernance

Cet atelier a présenté un scénario d'introduction d'un composant de ML en tant qu'outil d'aide à la décision. En effet ce composant s'intègre dans un processus comportant plusieurs étapes :

1. Application de règles métier (âge, filtrage, surendettement) préalablement définies par les experts métier en lien avec l'équipe Validation ;
2. Calcul automatique du score d'octroi (dont le poids dans le processus de décision global est moindre que celui des règles métier) ;
3. Intervention humaine par un agent, où celui-ci peut opérer un forçage, aussi bien dans le cas d'un score élevé (acceptation du prêt par la système) que dans celui d'un score faible.

8.3.6. Évaluation : méthodes et enjeux

Explicabilité

Les objectifs d'explicabilité identifiés dans ce cas d'usage sont multiples :

- il s'agit avant tout pour les concepteurs de l'algorithme de valider le bon fonctionnement du modèle obtenu et de faciliter le processus de validation ;
- les explications se destinent aussi aux responsables du *monitoring* du modèle en continu ;
- enfin, elles seront à terme utiles aux agents pour comprendre un résultat négatif produit par l'algorithme avant de prendre une décision sur le refus définitif du crédit ou son octroi par forçage.

La méthode SHAP a été retenue pour ces trois situations (LIME a été aussi évalué) pour plusieurs raisons : elle permet de passer d'une interprétabilité globale (quelle information pèse sur les décisions du modèle) à une interprétabilité locale (quelles valeurs prises par une information influent positivement ou négativement sur la décision), en outre elle a été jugée comme celle produisant les explications les plus proches du modèle traditionnel (de régression logistique), et enfin elle était aisée à mettre en œuvre dans chaque cas de figure.

Une méthode d'explicabilité contrefactuelle (cf. « Explications contrefactuelles ») est néanmoins également à l'étude : elle réclamera un important travail d'IHM si les explications doivent présenter beaucoup d'informations à l'utilisateur. Par ailleurs l'explication doit être aussi intuitive que possible, or l'arbre de décision sous-jacent peut avoir découpé selon des critères peu logiques, par exemple « `âge < 23,5 ans` ».

Performance

Les principales métriques retenues pour évaluer la performance d'un modèle sont la matrice de confusion ou le score F1 pour évaluer rappel et précision, le GINI pour évaluer son pouvoir discriminant, et le coefficient Kappa pour la comparaison entre ancien et nouveau modèles de *scoring*.

Un seuil du score GINI est notamment prédéfini par les recommandations en vigueur dans l'organisme, d'une part pour l'ensemble des modèles d'octroi (le constat actuel étant que ce score est atteignable pour la plupart des modèles refondus à l'exception de certaines sous-populations telles que les tranches d'âge « jeunes »), d'autre part pour l'ensemble des modèles réglementaires (avec un seuil plus élevé).

Le gain de score GINI obtenu en passant des modèles de *scoring* traditionnels au modèle de ML issu du *Gradient Tree Boosting* est assez faible (quelques points de pourcentage) dans le cas détaillé au cours de l'atelier, celui concernant l'équipement ménager. Néanmoins il peut aller jusqu'à 23 points de pourcentage parmi les modèles développés par l'équipe rencontré – ce gain est obtenu pour un modèle ayant initialement un faible pouvoir discriminant. De plus, même un gain de GINI pouvant *a priori* sembler marginal représente en général une réduction significative des pertes de crédit attendues (*Expected Credit Loss*).

Stabilité

La principale métrique de stabilité retenue est basée sur les résultats de la cross-validation (contrôle de l'écart-type sur les différents *folds*).

Plusieurs indicateurs sont également surveillés :

- taux de mutation de la population (*Population Stability Index*) ;
- évolution du profil du portefeuille (taux de demandes, taux d'acceptations, nombre d'impayés sur le trimestre écoulé), en accord avec les pratiques de *monitoring* décrites dans la section « Dérive temporelle » ;
- évolution des métriques de performance métier.

En cas d'alerte sur ces indicateurs, une analyse vise à fournir les causes probables de ces anomalies statistiques, et un plan d'action est produit, incluant éventuellement une refonte du modèle.

L'organisme indique manquer à ce stade de recul sur l'exploitation de cet algorithme (qui est faite en doublure du modèle traditionnel encore utilisé en production) quant à sa stabilité ou à la fréquence nécessaire de mise à jour du modèle.

8.3.7. Industrialisation

Les modèles de score d'octroi développés par l'organisme ne sont pas encore en production. Une méthode d'analyse du risque de crédit des entreprises, en revanche, a été mise en place : elle analyse des données essentiellement ouvertes (en Open Data) pour estimer le risque de défaillance d'une société.

8.4. Sous-atelier « Probabilité de défaut »

8.4.1. Intérêt de l'exercice

Un atelier a été conduit avec le Pôle Crédit d'un cabinet de conseil proposant à ses clients du secteur financier une solution de conception de modèles ML de probabilité de défaut. Cet atelier constituait un bon complément à celui sur les scores d'octroi en raison du caractère générique et surtout externalisé de la solution (adoption par les acteurs bancaires d'un outil développé par un prestataire).

L'approche du cabinet consiste non pas à fournir une solution sur étagère, fonctionnant en boîte noire, mais une boîte à outils permettant la définition d'un modèle en dialogue constant avec le client. En pratique, le modèle proposé en fin de processus est un modèle « hybride », basé en partie sur des algorithmes ML avancés lors de la conception mais traduit en des algorithmes simples et explicables lors du déploiement. La raison de ce choix semble motivée par la nécessité de fournir un modèle documenté, assorti d'une piste d'audit.

La solution actuelle est conçue pour les modèles de score d'octroi de crédit et de probabilité de défaut, mais le fournisseur étudie l'application d'une approche similaire aux modèles internes, à savoir utiliser du ML pour trouver des correctifs aux modèles actuels sous forme de règles expertes.

8.4.2. Objectifs du modèle

Les objectifs attendus sur ce projet étaient principalement les suivants :

- Amélioration de la performance des modèles utilisés pour la prise de décision : en particulier une meilleure discrimination du risque via l'identification des effets non linéaires entre les facteurs de risque, une meilleure classification des individus, et une identification plus rapide des changements dans les comportements sous-jacents du portefeuille
- Une amélioration de la qualité des données, grâce à l'utilisation de techniques de mise en qualité
- Une meilleure estimation du capital réglementaire grâce à des modèles plus précis

- Une transparence et auditabilité des nouveaux modèles.

Il existe un enjeu de données essentiel sur ce sujet, car le volume de données à disposition est très variable en fonction du cas d'usage : peu de données pour le crédit à la consommation, bien plus pour le crédit immobilier.

8.4.3. Détails techniques

Les étapes principales du fonctionnement nominal de la solution étudiée sont les suivantes – elles sont assez classiques à l'exception de la dernière qui fait l'originalité de l'approche retenue :

1. Préparation des données pour la modélisation : contrôle de qualité et mise en qualité traditionnels.
2. Construction d'un modèle de référence (de type « traditionnel ») : en pratique, il s'agit d'une régression logistique.
3. Construction d'un modèle challenger (de type « avancé ») : il s'agit d'algorithmes de ML supervisés plus sophistiqués : typiquement des forêts aléatoires ou réseaux de neurones.
4. Identification de la marge de progression du modèle de référence : dans le cas d'usage retenu, 80% de l'erreur est imputable à 20% de la population considérée, il s'agit donc ici d'identifier les sous-populations incorrectement classées par le modèle de référence.
5. Visualisation de l'explicabilité des décisions prises par les algorithmes de ML : les méthodes utilisées sont classiques (SHAP, LIME).
6. Extraction de règles métier simples et auditables expliquant la différence de performance vis-à-vis du modèle classique, de référence : pour ce faire, des segments de population mal classés par le modèle de référence sont identifiées automatiquement, puis des règles métier sont extraites par un expert humain (typiquement en charge des risques) afin de réduire autant que possible cette différence de performance.
7. Définition du modèle hybride final, combinant modèle de référence et règles métier.

La solution est fournie en mode « service outillé » : outre la chaîne de conception des modèles hybrides décrite ci-dessus, une plateforme de partage d'informations propose une revue du processus de conception complet par le client, de façon indépendante à l'exécution de ce processus.

D'un certain point de vue, la méthodologie de conception adoptée est basée sur les modèles *challengers* mentionnés dans ce document comme une approche possible à l'audit (cf. « Audit des algorithmes d'IA ») : de l'ordre de plusieurs centaines de modèles-types sont ainsi mis en concurrence afin de retenir le meilleur, vis-à-vis duquel la marge de progression du modèle de référence devra être minimisée. La stratégie consiste donc à répliquer la performance des meilleurs modèles *challengers* tout en demeurant dans un cadre opérationnel *a priori* mieux maîtrisable, avec la combinaison d'un modèle (régression logistique) intrinsèquement explicable avec des règles métier en nombre restreint.

Les concepteurs de la plateforme de ML étudiée insistent sur la décision prise dès le départ d'éviter le recours à un modèle de ML « pur », d'une part en raison de la difficulté de mise en œuvre du ML dans ce type de scénario, d'autre part parce qu'un tel modèle masquerait des comportements inhérents à la population modélisée, tels que la transition d'individus entre segments de population au cours du temps dans le cas de modèles d'octroi de crédit.

8.4.4. Processus de validation

La validation fonctionnelle initiale s'appuie sur une documentation de l'algorithme et une présentation des résultats. Elle est effectuée par le fournisseur de la solution en accompagnement de son client,

dans un mode itératif focalisé notamment sur les étapes 4 à 7 décrites précédemment (de l'identification des marges de progression jusqu'à la définition du modèle hybride produit).

Quant à la validation fonctionnelle continue, elle est similaire au *backtesting* classiquement utilisé pour les modèles de crédit, à ceci près que le suivi des sous-populations concernées par les règles métier est requis à une fréquence plus élevée, le but étant d'anticiper la détection d'un biais de modèle. Les résultats du *backtesting* sont par ailleurs présentés au Comité Risques afin de déterminer si un réajustement du modèle s'impose.

8.4.5. Enjeux de gouvernance

La conception de la solution, qui consiste *in fine* à ajuster le modèle de référence par des règles métiers (étapes 6 et 7 ci-dessus), a pour objet de la rendre compatible avec la gouvernance classique des modèles traditionnels. En particulier, le modèle hybride conçu par le participant à l'atelier est assimilable au fonctionnement classique des modèles d'octroi de crédit, avec un processus métier analogue : ainsi un modèle de régression – comparable aux modèles IRB (*Internal Ratings-Based Approach*) – est utilisé, puis un « *override* » (forçage du résultat, similaire au « *notching* » pratiqué par les agences de notation de crédit) peut être appliqué par un opérateur humain s'il identifie une faiblesse dans le résultat produit par le modèle. Outre le fait de maintenir un modèle de gouvernance éprouvé, le bénéfice attendu est une capacité explicative élevée du modèle (voir « Explicabilité » ci-dessous). Le choix d'un modèle hybride a aussi été retenu pour d'autres raisons opérationnelles : capacité d'implémentation facilitée, stabilité et robustesse.

Une autre question de gouvernance soulevée par ce cas d'usage est pour sa part relativement classique : celle de l'externalisation de la conception (et de la mise à jour) du modèle.

8.4.6. Évaluation : méthodes et enjeux

Explicabilité

En faisant le choix technique d'un modèle hybride basé sur des règles de décision, le fournisseur de la solution a mis l'accent[17] sur la production d'explications à destination des utilisateurs et des organes de gouvernance, aussi bien locales que globales.

Ainsi, un critère d'explicabilité important est que les forçages des décisions décrits ci-dessus doivent être motivés. *A fortiori* l'utilisateur du modèle – typiquement un chargé de compte – doit comprendre pourquoi le modèle a produit un score donné. Par ailleurs, comme expliqué dans la section « Interactions humain/algorithme », l'intervention d'un agent humain introduit un risque de biais explicatif par rapport au résultat de nature plus objective fourni par le modèle. Dans le modèle hybride, les règles métier ont été présélectionnées par l'algorithme : le modèle a d'abord été optimisé en tant que régression logistique, puis l'ajout de règles métier a pour but d'optimiser le modèle hybride résultant, toujours en termes de performance globale.

Quant à l'explicabilité locale, l'utilisation de SHAP fournit les motifs d'un score particulier produit par le modèle de régression logistique. Une explication de la décision pour l'ensemble du modèle hybride consiste à compléter ces valeurs de SHAP par la justification d'un forçage éventuel par les règles

[17] Ce souci se retrouve dans le détail de certains choix techniques : ainsi dans l'ajustement des hyperparamètres, le choix d'un algorithme génétique a été fait plutôt qu'une optimisation bayésienne, car plus facile à vulgariser et présentant une performance similaire.

métier, ce qui se traduit (trivialement) par l'appartenance de l'individu concerné à tel ou tel segment de population sur lequel l'algorithme avait optimisé la performance prédictive.

Performance

Les métriques suivantes, combinant performance prédictive et performance commerciale (cf. section « Principe de performance »), sont utilisées pour évaluer la pertinence de l'ensemble de la chaîne de traitement :

- comme métrique de performance prédictive, le gain du score GINI (typiquement de l'ordre de 5% dans les cas étudiés) ;
- comme indicateurs de performance commerciale : d'une part un gain en termes de revenus à appétence de risque équivalente (d'environ 50%), d'autre part une réduction de la perte attendue (*Expected Loss*) dans les modèles internes.

En outre, la réplicabilité du modèle a été étudiée : dans les expérimentations initiales le caractère non reproductible de la conception de modèle était problématique, mais il a été résolu par la suite.

Stabilité

En premier lieu, cet atelier a illustré l'observation faite dans la section « Dérive temporelle », selon laquelle la dérive temporelle d'un modèle prédictif peut être avant tout causée par un changement important dans les données sources, sans même faire intervenir l'algorithme de ML. Ainsi dans le cas des risques de crédit, les changements dans la structure de la population considérée introduisent des biais dans les modèles. Or l'évolution de la base de clients d'un groupe bancaire, par exemple, est rarement prise en compte par les modèles de type IRB. Aussi le cabinet de conseil ayant participé à l'atelier prône-t-il l'adoption par les organismes bancaires d'une solution de monitoring des portefeuilles (de clients, mais aussi d'encours ou d'actifs) afin de détecter ces changements structurels.

Concernant l'étude de la stabilité due cette fois au modèle prédictif lui-même, un sous-projet est en cours chez le concepteur pour fournir des KPI constituant la base d'un protocole de *monitoring* et de *backtesting* des modèles hybrides produits, afin d'identifier les dérives du modèle.

La stabilité du modèle hybride ne diffère de celle du modèle de régression logistique que par le choix des règles métier à inclure dans ce modèle. Ce choix est fait par le client, en interaction avec le fournisseur de la solution pour discuter des implications techniques. Le client peut aussi choisir lors d'une revue du modèle de supprimer une règle, par exemple en fonction de son « appétence aux risques », ou en raison de problèmes de qualité de données sur une variable impliquée dans la règle en question.

De plus l'analyse de stabilité du modèle hybride a montré que l'introduction de règles métier n'enlève rien à la robustesse du modèle, du moment que la preuve est faite que ces règles se cantonnent aux segments de population identifiés. Elles permettent en outre d'assurer un suivi spécifique de ces segments de population.

Enfin, des études initiales suggèrent qu'une périodicité de 6 mois pour la mise à jour du modèle serait adéquate, tant pour les scores d'octroi que pour les probabilités de défaut.

Traitement des données

Dans ce modèle de solution de type plateforme externalisée, la validation des modèles, mais aussi de l'adéquation du traitement des données, est *in fine* à charge des équipes Conformité et Risques du client.

Il n'existe donc pas de délégation de responsabilité, néanmoins les exigences réglementaires d'explicabilité imposées au client final sont par nécessité reportées sur la solution fournie par le prestataire. Celui-ci a d'ailleurs indiqué au cours des ateliers qu'un projet était en cours afin de mettre en place un Comité Éthique impliquant de grandes banques parmi ses clients ; le but ultime est de fournir un cadre de gestion des risques de modèles (*Model Risk Management* ou MRM).

8.4.7. Industrialisation

Le choix d'un processus de constructif itératif et non pas automatisé de bout en bout est délibéré. En effet, la solution proposée par le participant à cet atelier fait intervenir une phase d'intervention humaine dans l'optimisation du modèle hybride : cette approche rend impossible l'automatisation intégrale de la chaîne de conception (mais comporte, selon le participant, des avantages en termes d'explicabilité et de stabilité du modèle résultant – cf. section précédente).

En-dehors de cet absence d'automatisation complète, l'industrialisation de la solution est observable dans ses deux composantes :

- d'une part la chaîne de construction des modèles hybrides suit une démarche systématique et une mise en œuvre industrialisée ;
- d'autre part l'outillage destiné au client prendra la forme d'une plateforme de partage d'informations (modèles, données, résultats) permettant au client d'être dans la boucle des décisions prises et des résultats obtenus lors de la conception des modèles. L'objectif de cette plateforme, encore au stade du développement au moment de la rédaction de ce document, est de fournir une piste d'audit automatique des échanges avec le client.

Cette architecture de solution a pour but de permettre la traçabilité de l'ensemble des étapes de conception de l'algorithme déployé en production, que ces étapes soient réalisées par une machine ou par des agents humains.

Quant aux risques induits par l'externalisation (voir section « Risque de tiers et externalisation »), ils appellent les commentaires suivants :

- le mode de conception du modèle, décrit précédemment, permet respectivement la reproductibilité de sa construction et son auditabilité ;
- la qualité de service relève de la responsabilité du client étant donné qu'il prend en charge sa mise en production et son exploitation ;
- la continuité de service et la réversibilité ne soulèvent pas de difficultés majeure car le client a la capacité de rejouer le modèle de régression, de plus il doit avoir la capacité de suivre l'évolution des règles métier indépendamment de l'externalisation de la conception du modèle ;
- reste le risque de dépendance vis-à-vis du fournisseur de solution, notamment dans son aspect le plus incontournable de maîtrise technologique : il incombe dans ce type de situation au client final de développer et maintenir son expertise et son savoir-faire afin de maîtriser ce risque.

8.5. Atelier « Protection de la clientèle »

Cet atelier a été réalisé avec un organisme d'assurance autour d'un projet de proposition de vente, à savoir un système visant à fournir un devis pré-rempli pour la vente et la souscription d'une assurance-habitation.

8.5.1. Contexte réglementaire

Comme évoqué dans la section « Conformité réglementaire », le devoir de conseil imposé par la DDA oblige à vendre le produit dans le respect de l'intérêt des clients. Ainsi, le but de l'innovation technologique en la matière doit être de proposer une offre cohérente avec les exigences et besoins du client plutôt que de contribuer à développer une demande.

8.5.2. Intérêt de l'exercice

L'enjeu principal de cette expérimentation était d'éclairer, sur un cas d'usage, les enjeux réglementaires de l'utilisation de l'IA dans le domaine de la distribution de produits d'assurance.

8.5.3. Objectifs

Pour un client ayant souscrit un contrat, par exemple d'assurance automobile, le but de l'outil développé est de pré-remplir un devis d'assurance habitation, incluant un tarif « à partir de ».

8.5.4. Données sources

La spécificité de ce cas d'usage est le recours intensif à des données géographiques liées directement à la sociologie immobilière :

- des données carroyées fournies par l'INSEE, incluant des informations telles que le taux de maisons vs. d'appartements, le taux de propriétaires, la surface moyenne, le revenu moyen (aux niveaux IRIS et commune) ;
- des données sur le bâti, commercialisées par un fournisseur de données, qui donnent surface et périmètre des bâtiments, à partir desquelles un format de bâtiment est estimé, puis la probabilité maison vs. appartement en est déduite ;
- le nombre de pièces au niveau commune ;
- un champ d'adresse postale, sur lequel du *text mining* est appliqué afin d'en extraire des *features* discriminantes pour la prédiction maison/appartement) ;
- un champ email, utilisé dans cette même prédiction.

8.5.5. Détails techniques

Le devis est pré-rempli avec les variables cibles suivantes, qui sont prédites en mode itératif (i.e. la 2e est prédite en incluant la 1e variable prédite dans les variables explicatives, la 3e est prédite en incluant les 2 premières, etc.) :

1. type d'habitation : maison ou appartement
2. statut du client : propriétaire ou locataire
3. nombre de pièces
4. assurance optionnelle des objets précieux
5. année de construction

8.5.6. Processus de validation

La validation concernait surtout ici l'équipe de Conformité : celle-ci opère des contrôles de cohérence entre d'une part l'expression des besoins par l'assuré, d'autre part les risques déclarés dans le devis pré-rempli puis amendé par l'utilisateur.

8.5.7. Enjeux de gouvernance

Le devis pré-rempli produit par l'algorithme étudié est exploité par l'organisme d'assurance dans plusieurs cas d'usage : l'envoi par email d'un lien hypertexte vers le devis ; le traitement des appels entrants, afin de faire des ventes croisées sur portefeuille ; ou encore le soutien de campagnes d'appels téléphoniques sortants.

L'enjeu de gouvernance réside principalement dans le respect des obligations de conformité à l'occasion de la vente de produits d'assurance, notamment le devoir de conseil, qui impose que les motifs du choix de tel ou tel produit soient exposés au client prospectif, ainsi que la cohérence entre les exigences et besoins exprimés par le client et les caractéristiques du produit conseillé.

En particulier, il convient de veiller à ce que les interactions clients-machine soient soigneusement étudiées afin que le processus de souscription basé sur le pré-remplissage des informations ne désincite pas le client à exprimer ses besoins[18], ni à vérifier l'exactitude des risques déclarés[19]. Conjointement, la réglementation impose que les amendements éventuels (cochage ou décochage d'une option ou modifications signalées par le client) au devis pré-rempli soient reflétés fidèlement en parallèle, si nécessaire, dans tout autre document formalisant et retraçant le recueil des exigences et besoins auprès du client et la proposition du produit d'assurance.

Une illustration de ces enjeux peut être donnée par certaines mesures adoptées au cours du développement de l'outil. Afin de garantir l'information du client, à l'ouverture du devis pré-rempli une fenêtre informative (*pop-up*) enjoint explicitement le client prospectif à vérifier les informations du devis, et à les corriger le cas échéant. L'exemple de l'option « objets précieux » est parlant : elle était initialement pré-cochée dans tous les devis, or il s'est avéré que 60% des clients décochaient la case. Il a donc été décidé, toujours dans l'optique de fournir un devis ajusté au plus près des intentions

[18] En particulier, un algorithme jugé efficace par les utilisateurs peut être doté par ceux-ci d'un pouvoir « prescriptif » même s'il n'a pas été conçu dans cet objectif.

[19] Le risque ici est celui des contentieux futurs potentiels en cas de « fausse déclaration » dont l'origine se trouverait dans le pré-remplissage.

du client, de cocher cette case en fonction de la prédiction fournie par le modèle sur la variable « objets précieux ».

8.5.8. Évaluation

Explicabilité

L'explicabilité auprès du client d'une prédiction individuelle ne représente pas selon l'organisme un enjeu majeur pour les modèles à destination du marketing : dans le cas présent, plutôt que d'une explication, il s'agit de fournir à l'assuré une demande explicite de validation.

En revanche, il semble important de pouvoir fournir une explication des prédictions du modèle – et surtout de ses erreurs – aux équipes chargées du *monitoring* du système et de la vérification de la conformité du processus[20]. Une erreur du modèle a en effet un impact fort sur le processus de souscription – processus dont la bonne compréhension par le client doit être garantie : un défaut de conseil peut être invoqué, voire la mise en jeu de la responsabilité lorsqu'une prédiction erronée et non corrigée par l'assuré devient une fausse déclaration (certes non intentionnelle). En outre, les erreurs de prédiction à l'avantage de l'assuré engendrent – si elles s'accumulent – un risque supplémentaire.

Performance

La performance prédictive du modèle est ici triviale à évaluer : il s'agit de mesurer la précision de la classification selon chacune des variables-cibles. En retenant les trois premières (avec une marge d'erreur d'une unité sur le nombre de pièces), l'équipe obtient 90% de classifications correctes.

Stabilité

Ce cas d'usage ne présente pas d'enjeu de stabilité, car les données sources sont relativement statiques et que les enjeux mineurs en termes de pouvoir prédictif.

Traitement des données

En termes de traitement des données, il existe des variables interdites pour le *pricing* assurantiel, l'enjeu est donc de vérifier leur absence des modèles produits, ainsi que l'impossibilité de leur inférence à partir d'autres variables prédictives du modèle.

8.5.9. Industrialisation

Les modèles prédictifs ainsi conçus ont été mis en production. Bien qu'ils tournent sur des serveurs d'exploitation, l'utilisation de l'IA ne rentre pas ici dans le cadre d'un processus de décision automatique: il ne s'agit pas de fournir des prédictions en continu car l'exploitation passe par une étape manuelle de collecte des résultats de l'algorithme.

Ainsi, une fois le modèle prédictif validé sur les plans fonctionnel et technique, il s'agit de le faire tourner à intervalle régulier afin de fournir les données utilisées dans les trois situations décrites précédemment (campagnes email, appels entrants, campagnes d'appels sortants).

[20] La possibilité de fournir des explications aux décisions algorithmiques à des fins de contrôle interne ou d'audit n'a toutefois pas été explorée lors de cet atelier.

9. Distinction entre explicabilité et interprétabilité

La distinction entre les deux concepts est présente dans la littérature scientifique, mais elle ne fait pas l'objet d'un consensus.

Définition sans distinction

Burrell (2016) insiste sur le problème d'interprétabilité des résultats des algorithmes sans définir les termes en jeu. Doshi-Velez et Been Kim (2018) échouent à distinguer les deux termes en les définissant l'un par rapport à l'autre. Néanmoins, l'article s'efforce de justifier la nécessité d'une catégorisation des formes d'interprétabilité. De la même façon, Biran et Cottonn (2017) effectuent un cercle entre les deux concepts: « l'explication est reliée à l'interprétabilité: les systèmes sont interprétables si les opérations peuvent être comprises par un humain […] ».

Tout en rappelant l'absence de définition formelle, Bogroff et Guéguan (2016) définissent l'interprétabilité comme l'habilité à expliquer ou à présenter des étapes dans des termes compréhensibles pour l'humain. De son côté, Tim Miller (2018) propose une analyse très complète des deux concepts. L'introduction du concept de « *degree* » permet de donner comme définition de l'interprétabilité « le degré à partir duquel un observateur peut comprendre les causes d'une décision ». Malheureusement, l'explicabilité n'est pas définie suivant ce même concept, et elle est définie comme un mode d'obtention de la compréhension par un agent. Miller insiste sur la nécessité pour le lecteur d'observer les similarités et les différences entre les deux concepts… alors qu'il annonce cinq lignes plus haut que ces concepts seront utilisés de façon équivalente.

Définition par distinction

En reprenant la définition d'interprétabilité de Miller, Molnar (2019) tente d'opérer la distinction entre les deux termes. Il essaye de définir l'explication comme explication des prédictions aux individus. Il introduit la question d'une « bonne explication » dans son ouvrage.

Bryce Goodman et Seth Flaxman, dans «*European Union regulations on algorithmic decision-making and a ''right to explanation''*» (2017), effectuent implicitement la distinction entre les deux concepts dans leur interprétation des articles 13 à 15 du RGPD. Ils indiquent notamment qu'un algorithme effectue des corrélations et des associations, de telle sorte qu'ils réalisent des prédictions sans fournir d'élément explicatif de ces corrélations ou de ces associations. La difficulté soulevée est donc qu'une interprétation est difficile puisque l'algorithme effectue des opérations sans avoir à interpréter ou à expliquer la démarche. Les auteurs identifient une tension entre le droit à l'accès aux informations personnelles collectées (articles 13-15) et les droits et libertés de collection des données (article 22). Une prédominance de l'article 22 conduirait au développement d'une société « black box » (Pasquale, 2015).

L'intervention de Laurent Serrurier à l'IRIT le 12 novembre 2018 consiste à dire que l'explicabilité renvoie au fonctionnement technique de l'algorithme tandis que l'interprétabilité renvoie à une dimension éthique. L'explicabilité est une donnée technique de la complexité de l'algorithme; l'interprétabilité est lié à la notion d'acceptabilité sociale.

De même, l'intervention de Louis Abraham à l'ACPR (2019), après avoir abordé la définition de Biran et Cottonn mélangeant les deux concepts *(« Explanation is closely related to the concept of*

interpretability: systems are interpretable if their operations can be understood by a human, either through introspection of through a produced explanation. »), assimile l'interprétabilité à la question « pourquoi » et l'explicabilité à la question « comment ».

L'article « Vers une intelligence artificielle responsable » d'Aurélien Garivier (2018) propose une distinction explicite en définissant les deux termes. En s'appuyant sur l'article « A Berkeley View of Systems Challenges for AI », une règle de décision est dite interprétable si on comprend comment elle associe une réponse à des observations; elle est dite explicable si on comprend sur quels éléments est basée la décision, éventuellement de façon contrefactuelle.

Sous-distinction au sein de l'interprétabilité

L'article de Lipton (2017) propose l'interprétation la plus satisfaisante des concepts d'interprétabilité et d'explicabilité. En refusant de comprendre l'interprétabilité comme concept monolithique, Lipton tente de penser un continuum de concepts en établissant différents critères logiques : la confiance dans le résultat, la causalité, la transférabilité du savoir, le caractère informatif et équitable de la décision prise. Ce cadre d'analyse permet de proposer une représentation concrète du continuum entre compréhension et explication.

10. Aspects techniques de l'explicabilité

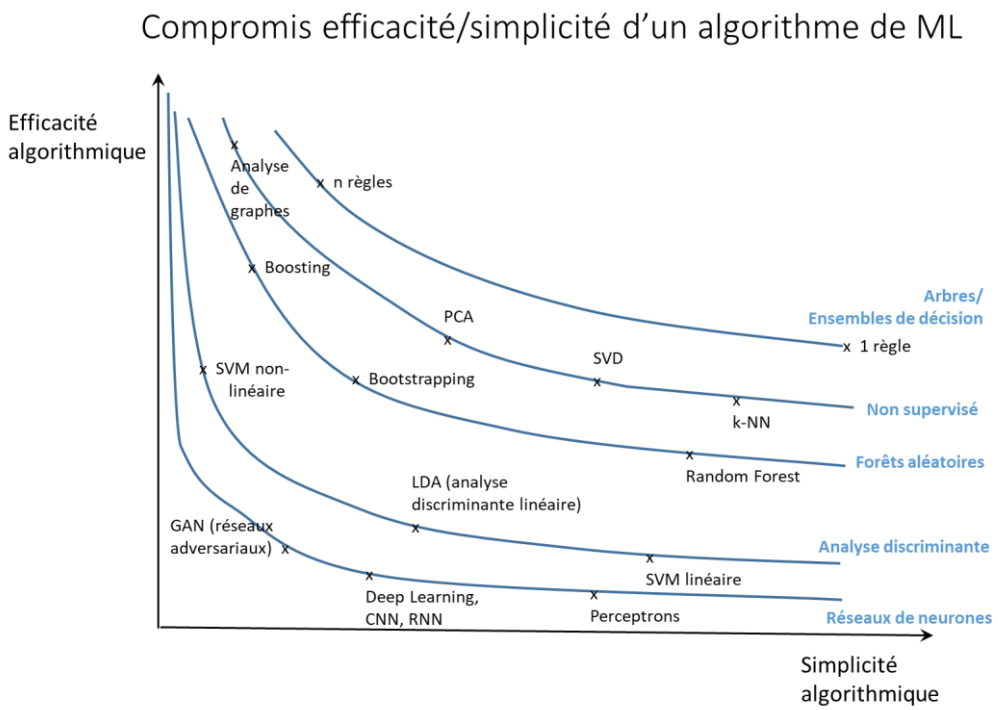
10.1. Arbitrages techniques

Cette annexe décrit les choix techniques découlant du positionnement de l'exigence d'explicabilité lors de l'introduction d'IA dans un processus métier. Cette description est d'ordre général car les considérations ne se limitent pas au secteur financier. Deux arbitrages sont présentés : d'une part entre simplicité et efficacité au niveau de l'algorithme de ML, d'autre part entre sobriété et fidélité de la méthode explicative retenue.

10.1.1. Arbitrage simplicité/efficacité

Un type donné d'algorithme de ML peut être plus ou moins complexe, au sens de se prêter à une inspection de son fonctionnement. Les types d'algorithme varient aussi en efficacité (mesurée comme indiqué précédemment par des métriques de performance prédictive ou commerciale).

Le diagramme suivant tente de schématiser l'arbitrage simplicité/efficacité opéré par les algorithmes de ML les plus courants :



Parmi les nombreuses simplifications et approximations opérées par ce diagramme, il convient de souligner les points suivants.

Métriques de simplicité et d'efficacité

D'une part, l'ordonnancement des types d'algorithme en termes de simplicité est subjectif. En effet, davantage que le type de modèle, sa taille et sa structure influent sur son explicabilité, car il ne sert quasiment à rien de comprendre une seule partie du modèle : ainsi une forêt aléatoire (*Random Forest*) comprenant des milliers d'arbres sera incomparablement plus difficile à comprendre qu'un réseau neuronal comportant une seule couche interne d'une dizaine de neurones.

D'autre part, le caractère déterministe ou stochastique d'un algorithme est aussi un critère fondamental à prendre en compte dans l'évaluation de son efficacité. En effet, les résultats d'un algorithme fondamentalement stochastique sont dépendants de tirages aléatoires : non seulement relativement aux jeux d'apprentissage et d'évaluation, mais aussi intrinsèquement à sa procédure (par exemple par ré-échantillonnage dans les méthodes de type *Bootstrap*).

Noter enfin que l'efficacité d'un type d'algorithme ne se livre pas non plus à une évaluation scalaire, car elle dépend du cas d'usage considéré (nature et volume des données, choix des paramètres, etc.)

Taxonomie (non exhaustive)

D'autre part, les algorithmes de ML sont représentés sur ce diagramme de façon non exhaustive. En particulier, des catégories telles que le *« Reinforcement Learning »* ont été exclues d'emblée car absentes – à notre connaissance - des technologies en œuvre sur le marché.

En revanche, les techniques d'apprentissage non supervisé ne peuvent être ignorées : ainsi la modélisation de réseaux d'interdépendance entre des PME participant à une plateforme de prêts de pair-à-pair (*P2P Lending*) réalisée par une analyse de graphe basée sur la factorisation (de type SVD ou *Singular Value Decomposition*, ou alternativement de *Latent Factor Model*) des caractéristiques de ces entreprises, a démontré une capacité non seulement descriptive, mais aussi prédictive des risques de défaut de crédit sur ce type de plateformes (Ahelegbey, 2019). Or ces risques se prêtent difficilement à une modélisation traditionnelle.

Découplage conception / modélisation

Enfin, la conception d'un algorithme et la structure du modèle résultant peuvent généralement être découplés : c'est la force et, à notre sens, la grande innovation apportée par des modèles hybrides tels que celui décrit dans le Sous-atelier « Probabilité de défaut ».

Cette approche consiste à construire un modèle simple et intuitif, en l'optimisant de façon itérative par comparaison avec les prédictions d'un algorithme bien plus efficace et souvent plus complexe. On combine ainsi le meilleur des deux mondes, à savoir la performance (en termes par exemple de taux de rappel et de précision) d'un algorithme complexe et l'explicabilité (taille maîtrisée et compréhensibilité) du modèle retenu *in fine*.

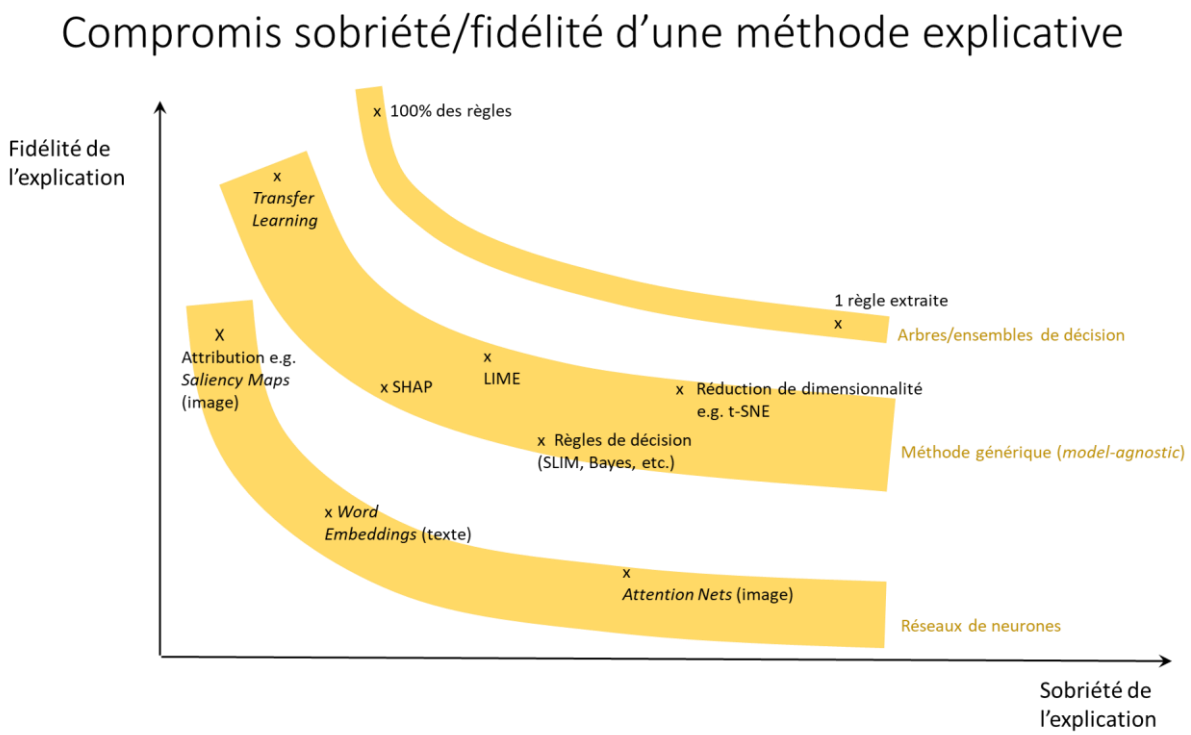
10.1.2. Arbitrage sobriété/fidélité

L'exigence d'explicabilité portée par l'introduction d'un algorithme ou d'une méthode d'apprentissage automatique dans un processus métier ne se limite pas à un équilibre à trouver entre simplicité et efficacité de l'algorithme : *l'explication elle-même* doit s'avérer intelligible et convaincante pour l'audience concernée, adaptée au cas d'usage, et proportionnée au risque porté par le processus.

Or un arbitrage existe là aussi, entre d'une part la fidélité d'une explication à l'algorithme qui l'a produite (imparfaite puisqu'on simplifie forcément l'algorithme en expliquant qu'il a pris telle décision en vertu de certaines caractéristiques de l'individu ou transaction considéré), d'autre part la sobriété de l'explication – c'est-à-dire son caractère intuitif, son intelligibilité par un individu non expert en la matière.

On peut schématiser par le diagramme suivant l'arbitrage sobriété/fidélité d'une explication selon le type d'algorithme de ML considéré et le type d'explication choisi. Y sont figurés quelques « couloirs »

d'arbitrage illustrant que pour un même type d'algorithme, certaines méthodes explicatives vont dévier de la courbe de tendance générale.



10.2. Obtention d'une explication de niveau élevé

10.2.1. Faisabilité de la réplication

Il convient tout d'abord de souligner que le niveau 4 (réplication) vise à reproduire à l'identique *le comportement* du modèle, et non d'en comprendre la mécanique interne dans ses moindres détails – ce qui peut s'avérer impossible pour certains modèles, typiquement les réseaux neuronaux profonds.

On notera aussi que certains acteurs rencontrés lors des travaux exploratoires de l'ACPR ont mis en place une méthode de réplication dès la phase de conception et de validation de leurs algorithmes, donc en amont des missions de contrôle interne ou d'audit : ils ont (ce qui se fait classiquement en ingénierie logicielle sur des composants particulièrement critiques d'un système) implémenté leurs algorithmes de ML dans plusieurs langages.

10.2.2. Le problème des dépendances logicielles

En outre, lorsqu'une revue du code de l'algorithme est pertinente (niveaux 3 et 4), un problème se présente, qui n'est pas spécifique aux algorithmes d'IA mais se pose pour quasiment toute mission d'audit de code bien conçue : de multiples librairies, outils ou composants externes sont utilisés par le code analysé, et ne peuvent que difficilement (dans le cas de code open source) voire pas du tout (dans le cas de code fermé) être eux-mêmes passés en revue. Même dans le cas d'une simple régression logistique ou linéaire, plusieurs librairies sont ainsi invoquées ; le problème se trouve démultiplié dans le cas d'algorithmes d'IA sophistiqués, qui réclament par ailleurs un niveau d'explication plus élevé.

Ainsi, fournir une explication de niveau 3 ou 4 est difficile dans la plupart des cas, et cette difficulté est accrue par plusieurs circonstances : lorsque l'algorithme a recours à des librairies ou produits tiers, mais aussi lorsque la mission d'audit porte sur toute la chaîne de conception du modèle plutôt que sur le seul modèle résultant.

Une piste de réflexion parfois évoquée pour faciliter ce type d'analyse approfondie est de mettre en place une certification des composants de ML « sur étagère », de façon analogue par exemple à des composants de sécurité qui doivent être éprouvés et officiellement validés avant d'être embarqués dans des applications critiques. Quoi qu'il en soit, l'analyse détaillée de l'algorithme évoquée pour le niveau 4 (réplication) porte avant tout sur l'utilisation de ces librairies sur étagère, un point essentiel étant notamment l'optimisation des hyperparamètres en tant qu'elle influe significativement sur la performance de l'algorithme.

11. Recension des méthodes explicatives en IA

Cette recension ne se veut pas exhaustive : elle se restreint comme le reste du document aux cas d'usage de l'IA dans le secteur financier, en outre le but est de dresser un tableau des applications ayant une visée de mise en œuvre pratique – qu'elles soient actuellement en production ou simplement en phase d'expérimentation.

On se place ici dans le cadre spécifique d'un algorithme de ML. Les méthodes d'explication se répartissent alors classiquement en trois catégories selon la phase de la conception et de l'utilisation du modèle sous-jacent :

1. les méthodes explicatives pré-modélisation visent à comprendre et décrire les données utilisées dans la production des modèles ;
2. les méthodes explicatives conjointes à la modélisation contribuent à la conception de modèles intrinsèquement plus explicables ;
3. enfin, les méthodes explicatives post-modélisation tentent de produire des explications satisfaisantes à partir de modèles préalablement conçus et entraînés.

11.1. Méthodes explicatives pré-modélisation

En amont de la phase d'apprentissage d'un modèle de ML, une forme d'explicabilité – certes limitée - peut être fournie : elle vise à comprendre et décrire les données utilisées par l'algorithme. Les principales méthodes utilisées dans ce but sont les suivantes.

Analyse exploratoire des données

Il s'agit souvent de méthodes basées sur des visualisations de données, permettant de mettre au jour des caractéristiques potentiellement masquées par les statistiques descriptives.

Ces méthodes sont agnostiques tant vis-à-vis du type de modèle de ML que du domaine considéré. Dans le domaine financier, elles sont particulièrement utiles à la détection et à la remédiation des biais non souhaités (cf. section « Éthique et équité »). Les sources potentielles de tels biais problématiques sont multiples (Kamishima, 2012) : dépendance directe, indirecte ou latente vis-à-vis de variables sensibles, biais d'échantillonnage (le cas le plus difficile à détecter) ou d'annotation dans les données d'apprentissage, ou encore convergence imparfaite de la phase d'apprentissage.

Documentation des jeux de données

Plusieurs standards de documentation ont été suggérés, soit pour les modèles d'IA (Mitchell, 2019) soit pour les services associés (Hind, 2018).

Ce type d'approche, basée sur une documentation claire, rigoureuse et formalisée des jeux de données et des services associés, est approprié au niveau 1 d'explication décrit dans ce document de réflexion (« Niveaux d'explication »).

Toutefois le focus technique des standards proposés jusqu'à présent les rend peu adaptés aux clients et utilisateurs finaux d'algorithmes d'IA : ils sont davantage destinés aux concepteurs de ces outils, voire aux responsables de leur fonctionnement sous un angle opérationnel. Cet effort de standardisation est néanmoins récent et appelé à évoluer dans un futur proche.

Méthodes de résumé de jeux de données

Afin de faciliter la représentation mentale et l'interprétation de jeux de données, notamment les plus volumineux et les plus hétérogènes, des méthodes de résumé de ces données peuvent être employées, notamment en complément des méthodes d'analyse exploratoire et de documentation citées précédemment.

On peut citer par exemple :

- pour des données textuelles, le résumé et la classification automatiques de documents ;
- pour des images, la synthèse de scènes visuelles ;
- pour tout type de données, l'extraction d'exemples représentatifs (ou typiques) d'un jeu de données, et d'exemples particulièrement atypiques en complément (appelés respectivement *prototypes* et *criticisms* en anglais : Kim, 2016).

Méthodes de « feature engineering » explicable

Ce dernier type de méthode explicative pré-modélisation part du constat qu'une explication d'un algorithme prédictif n'est jamais meilleur que les variables prédictives (*features*) qu'il utilise. Partant, un soin particulier doit être accordé dans la conception d'un système de ML à la phase de *feature engineering*, c'est-à-dire la construction de variables prédictives à partir des variables d'origine afin de restructurer adéquatement les données d'apprentissage d'un algorithme de ML.

On distingue notamment deux types de méthodes (Murdoch, 2019) :

- l'intervention d'experts métiers, qui connaissent suffisamment le domaine dont sont issues les données sources pour en extraire des variables (combinaisons d'autres variables, résultats de calculs préliminaires, etc.) qui augmentent la performance prédictive du modèle obtenu tout en maintenant l'interprétabilité de ses résultats : autrement dit, l'expertise humaine permet ici, dans certains cas, de s'affranchir de l'arbitrage généralement inéluctable entre efficacité et explicabilité du modèle de ML (cf. « Arbitrage simplicité/efficacité ») ;
- une approche automatique basée sur la modélisation : des méthodes très courantes en analyse de données sont alors utilisées, par exemple la réduction du nombre de dimensions et le clustering, afin d'extraire des variables prédictives aussi concises et représentatives que possible.

11.2. Méthodes explicatives conjointes à la modélisation

Certaines méthodes permettent de bâtir un modèle explicatif simultanément à l'apprentissage et à la construction du modèle prédictif.

Elles sont bien plus rarement mises en œuvre que les méthodes pré- et surtout post-modélisation, pour plusieurs raisons :

- elles nécessitent d'avoir accès au code produisant ce modèle (et pas seulement au modèle lui-même, comme c'est le cas des méthodes post-modélisation qui sont donc plus généralement applicables) et de pouvoir modifier l'algorithme ;
- elles ne sont mises en œuvre que lorsque l'explicabilité est recherchée dès la phase de conception de l'algorithme de ML, ce qui réclame un niveau de maturité et de planification adéquat lors de l'introduction d'IA au sein de processus métier ;
- enfin, elles sont par nature peu appropriées à une situation d'audit – *a fortiori* lorsque le modèle prédictif est fourni en « boîte noire » sans description de l'algorithme lui-même.

Leur défi principal, ambitieux, est de s'affranchir au moins partiellement, pour un modèle de ML considéré, de l'arbitrage entre efficacité et explicabilité déjà cité : elles fournissent un supplément d'explicabilité sans (forcément) sacrifier la performance.

On peut toutefois citer les méthodes suivantes.

Choix d'un modèle intrinsèquement explicable

Il s'agit de choisir un type de modèle intrinsèquement explicable, par exemple des modèles linéaires ou basés sur des arbres de décision. C'est la solution la plus triviale, à condition d'une part d'être conscient de l'arbitrage généralement présent entre explicabilité et simplicité, d'autre part de s'assurer que le modèle particulier produit par l'algorithme sera explicable : en effet dans certains cas, l'adoption d'une famille explicable ne suffit pas car elle conduit à un modèle de dimension trop élevée pour être compréhensible.

Modèles explicables hybrides

La construction d'un modèle explicable hybride est applicable à un type de modèle spécifique, à savoir les réseaux neuronaux. On peut mentionner dans cette catégorie :

- le *Deep k-NN* (Papernot et McDaniel, 2018) qui fait ressortir la représentation interne d'un réseau neuronal dans chacune de ses couches, afin d'illustrer l'obtention du résultat final (donc dans la dernière couche). Une variante est le *Deep Weighted Averaging Classifier* ;
- le SENN (*Self-Explaining Neural Networks* : Alvarez-Melis, 2018) qui utilise des réseaux neuronaux pour entraîner simultanément les variables prédictives, les coefficients, et la méthode d'agrégation d'un modèle linéaire. Une variante est le *Contextual Explanation Network* (Al-Shedivat, 2018).

Modèles conjoints de prédiction/explication

Cette approche, qui consiste à entraîner le modèle à fournir conjointement une prédiction et son explication, a récemment reçu une attention considérable. Elle a toutefois deux limitations majeures : d'une part, outre que l'algorithme de ML doit être modifié, elle nécessite de fournir des explications à l'ensemble des données d'apprentissage, ce qui dans la pratique est souvent illusoire ; d'autre part, les explications qu'ils produisent correspondent aux motivations indiquées par des humains au moment de l'apprentissage hybride, et ne constituent pas nécessairement des justifications du processus prédictif de l'algorithme.

On peut toutefois citer parmi ces modèles conjoints :

- la méthodologie TED (*Teaching Explanations for Decisions* : Hind, 2019) associe à chaque donnée d'apprentissage la motivation de la prédiction associée. Une variation est la production d'explications multimodales (Park, 2018) ;
- des méthodes spécifiques à un type de données : explications visuelles (*Visual Explanations* : Hendricks, 2016) pour la reconnaissance d'objets dans des images, ou génération d'explications concises en langage naturel (français, anglais) à un modèle prédictif à partir de sources de données textuelles (Lei, 2016) ;

Méthodes d'ajustement architectural

Ces méthodes sont pour la plupart spécifiques aux réseaux neuronaux profonds (*Deep Learning*, relativement inusité en finance).

On mentionnera tout de même les mécanismes d'attention (*Attention-Based Models*) qui visent à identifier les composantes des données d'entrée les plus importantes pour l'algorithme, que les données soient des images, du texte ou – de façon plus pertinente dans le secteur financier – des séries temporelles. Certaines études (Jain, 2019) ont toutefois montré les limites de cette approche en termes de performance du modèle résultant.

Méthodes de régularisation

Elles sont en principe utilisées pour améliorer la performance d'un modèle de ML, toutefois certains types de régularisation parviennent aussi à augmenter le niveau d'explicabilité.

Par exemple, la zone de décision d'un modèle peut être au cours de l'apprentissage contrainte à être approximable par un arbre de décision, rendant ainsi les futures prédictions aisément compréhensibles par un humain (Wu, 2017). Ou encore, les décisions du modèle peuvent être orientées pour prendre en compte les variables prédictives annotées comme importantes par un expert du domaine (Ross, 2017).

Découplage entraînement/modélisation

Une mention particulière doit être faite des approches bien spécifiques consistant à découpler la conception de l'algorithme de ML de la structure du modèle résultant.

C'est le cas de l'approche hybride décrite dans le Sous-atelier « Probabilité de défaut » : un modèle avancé, peu explicable par nature, est entraîné afin de fournir une performance élevée, suite à quoi des experts métiers extraient « manuellement » un ensemble de règles qui alimentent un modèle intrinsèquement explicable (de type arbre de décision). Le système obtenu bénéficie ainsi à la fois de la performance d'un algorithme complexe et de l'explicabilité du modèle prédictif.

11.3. Méthodes explicatives post-modélisation

On notera que les méthodes opérant sur les modèles préalablement entraînés sont en fait l'acceptation la plus couramment utilisée pour les méthodes explicatives du ML en général. Il s'agit ainsi de fournir une explication « post-hoc », visant à motiver ou comprendre un résultat (ou un ensemble de résultats) produit par un modèle de ML. Le modèle est alors un objet d'étude sur lequel on ne peut opérer (contrairement aux méthodes conjointes à la modélisation), pas plus qu'on ne peut modifier les données qu'il manipule (contrairement aux méthodes pré-modélisation).

Deux critères principaux permettent de distinguer les méthodes post-modélisation. D'une part, leur caractère local ou global :

- les méthodes explicatives locales fournissent une explication à une décision relative à un point de données particulier en entrée de l'algorithme (par exemple, pourquoi telle demande de crédit a été octroyée à un individu donné) ;
- les méthodes explicatives globales tentent d'expliquer simultanément l'ensemble des décisions possibles (quelles sont les caractéristiques générales des décisions d'octroi ou de refus des demandes de crédit par l'algorithme).

D'autre part, leur caractère applicable à tout type de modèle de ML (*model-agnostic*) ou à un type de modèle ou d'algorithme en particulier (*model-specific*).

11.3.1. Méthodes explicatives locales

Méthodes de type « boîte noire »

Ces méthodes, également qualifiées de « *model-agnostic* », sont applicables à tout type de modèle. Elles peuvent consister en un simple classificateur (par exemple un simple classificateur bayésien entraîné sur des fenêtres de Parzen), ou s'avérer être plus sophistiquées (nombre d'entre elles consistent à perturber le modèle et à observer l'influence des variables prédictives).

Parmi les méthodes les plus courantes d'interprétation locale « *model-agnostic* », on peut citer :

- les méthodes d'analyse bayésienne (*Naive Bayes Models*), souvent frustes par rapport aux suivantes ;
- la méthode LIME (*Locally Interpretable Model-Agnostic Explanations*), qui repose sur la création d'un domaine de représentation intermédiaire (entre le modèle de ML et le modèle de la « réalité du monde ») afin de trouver le compromis optimal entre la fidélité de l'explication au modèle et la simplicité de l'explication (laquelle vise à être comprise par un expert métier non nécessairement doté d'une expertise technique) ;
- la méthode SHAP, qui combine théorie des jeux (les valeurs de Shapley) et l'optimisation d'allocation de crédits pour expliquer l'influence de chaque variable prédictive sur les valeurs prédites, de nouveau de façon agnostique vis-à-vis de la nature du modèle (Lundberg, 2017) ;
- des variantes de la méthode SHAP, par exemple adaptées à des données structurées en réseau (Chen, 2019) ;
- des méthodes d'interprétation individuelle causale, qui calculent l'influence marginale de chaque variable prédictive et les influences jointes de paires de variables prédictives (Datta, 2016) ;
- la méthode SLIM (*Supersparse Linear Integer Models*) sélectionne des règles de décision qui optimisent la précision d'un classificateur binaire selon des contraintes sur le nombre de variables et leurs poids relatifs.

Il convient de noter que même les méthodes d'interprétation locale les plus couramment utilisées, telles que LIME et SHAP toutes deux basées sur des perturbations de modèles, ont des limites pratiques en termes de sécurité (Dylan, 2020). Elles sont notamment vulnérables à des attaques « adversarielles » (ou *adversarial attacks* : voir annexe « Recension des attaques contre un modèle de ML ») produisant des modèles comportant des biais à caractère discriminatoire sur lesquels ces méthodes explicatives produisent cependant des explication de nature rassurante, voire indiscernables de celles produites sur un modèle non biaisé.

Il existe aussi des méthodes explicatives spécifiques au domaine du NLP, qui fournissent généralement des explications sous forme numérique ou d'un exemple textuel :

- une adaptation de la méthode LIME au NLP (Ribeiro, 2016), fournissant des explications sous forme de degré d'importance de chaque variable prédictive ;
- une méthode générative (Liu, 2018), fournissant des explications sous forme d'un exemple textuel.

Méthodes spécifiques à un type de modèle

Il existe quantité de méthodes dépendantes du type de modèle, ou « *model-specific* ».

Tout d'abord, certains modèles sont directement interprétables :

- régressions logistiques ;

- régressions linéaires et leurs variantes (GLM ou *Generalized Linear Models*) du moins lorsqu'elles sont peu denses ;
- modèles additifs (GAM ou *Generalized Additive Models*);
- arbres de décision et forêts aléatoires, du moins lorsque celles-ci sont de taille modérée.

Ensuite, de nombreuses méthodes explicatives pour les modèles de *Deep Learning* :

- explications sous forme de modèles de substitution (*Surrogate Models*) de type arbre de décision pour approximer le réseau de neurones (Craven, 1995) ;
- explications basées sur les mécanismes d'attention (Choi, 2016) ;
- explications sous forme d'attribution d'une décision à des variables prédictives (Shrikumar, 2017).

Enfin, des méthodes appliquées à des domaines particuliers :

- pour les algorithmes de NLP (traitement automatique du langage) basés sur des réseaux neuronaux récurrents (Strobelt, 2018) ;
- pour les algorithmes de CV (reconnaissance d'images), par exemple : unités interprétables (Bau, 2017), cartographie des zones d'incertitude (Kendall, 2017) ou des zones saillantes (*Saliency Maps* : Adebayo, 2018).

Explications contrefactuelles

Les explications contrefactuelles ont une place à part dans les méthodes cherchant à expliquer un algorithme de ML, en tant qu'elles sont les seules à faire intervenir des éléments de causalité[21] (et pas simplement des explications de nature statistique ou des inférences résultant de généralisations à partir d'un volume important de données).

Plus précisément, une explication contrefactuelle à la prédiction Y fournie par un modèle à partir de données d'entrée X, correspond aux données d'entrée X' les plus proches de X qui auraient abouti à une prédiction Y'. En général, Y est une prédiction ou décision défavorable, par exemple un score faible conduisant à refuser un octroi de crédit en fonction de la demande X. Dans ce cas, une explication pertinente (pour un concepteur ou un auditeur du système, mais surtout pour l'individu impacté par la prédiction, en l'occurrence celui ayant déposé la demande d'emprunt) répond à la question : quel est le changement aussi minimal que possible dans la demande de crédit qui aurait conduit à une acceptation ? Ainsi, plutôt qu'une explication locale quantifiant l'influence de plusieurs variables prédictives (l'âge, le revenu, l'historique de crédit, etc.) sur la décision négative, on obtient une explication bien plus utile (ou pragmatique) et plus simple, telle que « si le revenu du foyer avait été de tant au lieu de tant, le prêt aurait été octroyé ».

Certaines méthodes d'explication contrefactuelle vont au-delà de cette description (McGrath, 2018) :

- d'une part en proposant des explications contrefactuelles positives, applicables dans le cas où la décision Y d'origine est favorable à l'individu concerné. Dans l'exemple ci-dessus, Y' correspond à un refus de la demande de crédit, et l'explication contrefactuelle indique donc

[21] Cette capacité d'appréhender la causalité est prometteuse pour le déploiement de l'IA en général, et pour le secteur financier en particulier. Par exemple, l'explicabilité des modèles internes mis en place dans les établissements bancaires serait renforcée dès lors que ces modèles permettraient de mesurer des liens causaux. L'identification causale est de fait au cœur des préoccupations de l'économie empirique depuis un quart de siècle. Or elle est absente des modèles d'IA grand public, tout comme des modèles plus classiques utilisés actuellement dans les établissements bancaires.

la marge de sécurité de la décision favorable. Ce type d'explication peut être utile pour décider en toute connaissance de cause de faire à une occasion future une nouvelle demande de crédit suite à une première demande acceptée ;

- d'autre part en pondérant les facteurs explicatifs en fonction de leur variabilité : dans l'exemple ci-dessus, si un individu a démontré plus de facilité à réduire ses dépenses qu'à accroître ses revenus, alors l'explication « si les dépenses mensuelles avaient été réduites de 50%, le prêt aurait été octroyé » est privilégiée car elle est plus utile que celle concernant le revenu.

Idéalement, une méthode d'explication contrefactuelle est applicable à un algorithme étudié comme une boîte noire, et certaines méthodes satisfont à cette condition dans des situations bien déterminées (Wachter, 2018).

11.3.2. Méthodes explicatives globales

Les méthodes explicatives globale fournissent une explication à l'ensemble des décisions prises par l'algorithme : par exemple, quelle est la contribution de la variable « âge » aux décisions d'octroi ou refus de crédit sur l'ensemble des demandeurs.

Les méthodes explicatives globales peuvent être utiles à un contrôleur interne ou un auditeur afin d'obtenir une compréhension du fonctionnement général de l'algorithme, mais elles montrent généralement leurs limites par rapport à l'étude d'un cas concret (via une explication locale qui justifie une décision individuelle) ou de plusieurs cas concrets (par exemple pour comparer le traitement réservé par l'algorithme à deux individus et détecter une éventuelle inégalité de traitement).

Les méthodes explicatives globales sont par ailleurs très difficiles à matérialiser en pratique. De telles méthodes existent pour un type de modèle spécifique, par exemple pour les réseaux neuronaux profonds il est possible d'extraire des règles de décision aisément interprétables et, selon les situations, relativement fidèles au modèle de *Deep Learning* (*DeepRED* : Zilke, 2016).

En revanche, peu de méthodes sont capables de fournir une explication globale indépendamment du type de modèle étudié. C'est le cas des *Partial Dependence Plots* (PDP), qui montrent l'effet marginal d'une variable donnée sur les prédictions du modèle (Friedman, 2001).

12. Recension des attaques contre un modèle de ML

La sécurité du ML est un domaine d'étude certes récent, mais assez important pour avoir conduit à une taxonomie – provisoire étant donné le caractère évolutif du sujet (Papernot, 2018). On distingue ainsi parmi les attaques les plus notoires contre des modèles ML, en indiquant un exemple de scénario à chaque fois :

- Les attaques causatives (*Data Poisoning*) : les données d'apprentissage sont altérées (valeurs d'attribut modifiées, nouveaux attributs créés)
    - Attaques contre l'intégrité : par exemple pour faire octroyer des emprunts généreux ou de faibles primes d'assurance aux malfaiteurs
    - Attaques contre la disponibilité : par exemple pour discriminer contre un groupe de population en leur refusant les mêmes avantages
- Les « attaques par filigrane » (*Watermark Attacks*) : le code est alors modifié par le malfaiteur
    - Attaques contre l'intégrité : par exemple conditions sur certains attributs pour déclencher des résultats bénéfiques
    - Attaques contre la disponibilité : par exemple injection de règles pour éliminer ces résultats pour la population visée
- Les attaques par substitution de modèle (*Surrogate Models*)
    - Attaques d'inversion : l'équivalent du *reverse engineering* pour les modèles de ML
    - Tests d'appartenance
- Les attaques « adversarielles » : il s'agit de la construction d'exemples synthétiques permettant d'échapper à un résultat négatif ou d'en obtenir un positif
- Les attaques d'usurpation *(Impersonation Attacks*) : il s'agit de l'injection de données correspondant à une identités réelle (ou composite d'identités réelles) afin d'usurper cette identité.

Un aspect particulièrement intéressant de la sécurité du ML est que les parades ont un « effet de bord » bénéfique (Hall, 2019), à savoir qu'elles répondent à l'ensemble des critères d'évaluation des algorithmes : explicabilité, efficacité, stabilité et qualité des données.

Pour ne citer qu'un exemple, une parade contre les attaques causatives par les données (*Data Poisoning Attacks*) est la méthode RONI (*Reject On Negative Impact*) qui consiste à refuser les données d'apprentissage engendrant une baisse de performance (Barreno, 2010) ; elle permet donc aussi de se prémunir contre les dégradations du modèle liées à une dérive des données d'apprentissage. En guise d'illustration, un algorithme de reconnaissance faciale sécurisé par RONI va exclure de son jeu d'apprentissage une série de photos associées chacune à un document d'identité qui ferait drastiquement baisser le taux de précision : cela contribue à garantir l'intégrité mais aussi la performance du modèle – qui peut être par exemple utilisé pour l'identification de clients dans une relation bancaire à distance.

Bibliographie

Louis Abraham. *In Algorithms We Trust*. ACPR (21 mars 2019).
Peter Addo, Dominique Guégan, Bertrand Hassani. *Credit Risk Analysis using Machine and Deep Learning models*. ffhalshs-01719983f (2018).
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: *Sanity Checks for Saliency Maps*. NeurIPS 2018: 9525-9536 (2018).
AEAPP. *Final Report on public consultation No. 19/270 on Guidelines on outsourcing to cloud service providers*. EIOPA-BoS-20-002 (2020).
Daniel Felix Ahelegbey, Paolo Giudici, Branka Hadji-Misheva. *Latent Factor Models for Credit Scoring in P2P Systems*. Physica A: Statistical Mechanics and its Applications No. 522 (10 February 2019): pp. 112-121 (2018).
Maruan Al-Shedivat, Avinava Dubey, Eric P. Xing. *Contextual Explanation Networks.* arXiv:1705.10301v3 [cs.LG] (2018).
David Alvarez-Melis, Tommi S. Jaakkola. *Towards Robust Interpretability with Self-Explaining Neural Networks.* arXiv:1806.07538v2 [cs.LG] (2018).
0 David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, Klaus-Robert Muller. *How to Explain Individual Classification Decisions.* J. Mach. Learn. Res. 11: 1803-1831 (2009).
Marco Barreno, Blaine Nelson, Anthony D. Joseph, J.D. Tygar. *The security of machine learning.* Mach Learn (2010) 81: 121–148 DOI 10.1007/s10994-010-5188-5 (2010).
Robert P. Bartlett, Adair Morse, Richard Stanton, Nancy Wallace. *Consumer-lending discrimination in the FinTech era* (No. w25943). National Bureau of Economic Research (2019).
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba. *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. CVPR 2017: 3319-3327 (2017).
Roland Berger. *The road to AI Investment dynamics in the European ecosystem*. AI Global Index (2019).
Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh Ruchir Puri, José M. F. Moura, Peter Eckersley. *Explainable Machine Learning in Deployment.* arXiv:1909.06342 [cs.LG] (2019)
Or Biran, Courtenay V. Cotton. *Explanation and Justification in Machine Learning: A Survey* (2017).

Alexis Bogroff, Dominique Guégan. *Artificial Intelligence, Data, Ethics: An holistic Approach for Risks and Regulation*, HAL (2019)
Jenna Burrell. *How the machine 'thinks': Unnderstanding opacity in machine learning algorithms*. Big Data & Society (2016).
Cambridge Judge Business School. *The Global RegTech Industry Benchmark Report* (2019).
Cambridge Judge Business School, World Economic Forum. *Transforming Paradigms A Global AI in Financial Services Survey* (2019).
Jianbo Chen, Le Song, Martin J. Wainwright, Michael I. Jordan. *L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data.* ICLR (2019).
Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F. Stewart. *RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism.* NIPS 2016: 3504-3512 (2016).
Mark Craven, Jude W. Shavlik. *Extracting Tree-Structured Representations of Trained Networks*. NIPS 1995: 24-30 (1995).
Wei Dai, Isaac Wardlaw. *Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking.* Information Technology, New Generations. Advances in Intelligent Systems and Computing. 448. pp. 439–450. ISBN 978-3-319-32466-1 (2016).
Anupam Datta, Shayak Sen, Yair Zick. *Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems*. In Security and Privacy (SP), 2016 IEEE Symposium on, pp. 598–617. IEEE (2016).
Doshi-Velez, Been Kim. *Towards a Rigorous Science of Interpretable Machine Learning* (2017).
EBA. *Draft recommendations on outsourcing to cloud service providers under Article 16 of Regulation (EU) No 1093/20101*. EBA/CP/2017/06 (2017).
European Commission High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy Artificial Intelligence* (2019).
Jerome H. Friedman. *Greedy function approximation: A gradient boosting machine.* Annals of statistics (2001).
Donna Fuscaldo. *ZestFinance Using AI To Bring Fairness To Mortgage Lending* (2019).
Aurélien Garivier. *Vers une intelligence artificielle responsable,* Institut mathématique de Toulouse (26 mars 2018).
Bryce Goodman, Seth Flaxman. *European Union regulations on algorithmic decision-making and a ''right to explanation''* (2017).

Dominique Guégan, Bertrand Hassani. *Regulatory Learning: how to supervise machine learning models? An application to credit scoring*. ffhalshs-01592168v2f (2017).
Patrick Hall. *Proposals for model vulnerability and security.* O'Reilly (2019).
Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell. *Generating Visual Explanations.* arXiv:1603.08507v1 [cs.CV] (2016).
Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, Kush R. Varshney. *Increasing Trust in AI Services through Supplier's Declarations of Conformity.* CoRR abs/1808.07261 (2018).
Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, Kush R. Varshney. *TED: Teaching AI to Explain its Decisions.* arXiv:1811.04896v2 [cs.AI] (2019).
Sarthak Jain, Byron C. Wallace. *Attention is not Explanation.* arXiv:1902.10186v3 [cs.CL] (2019).
Konstantinos Koutroumbas, Sergios Theodoridis. *Pattern Recognition (4th ed.).* Burlington. ISBN 978-1-59749-272-0 (2008).
C. Jung, H. Mueller, S. Pedemonte, S. Plances, O. Thew. *Machine learning in UK financial services*, Bank of England & Financial Conduct Authority (2019).
Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, Jun Sakuma. *Fairness-Aware Classifier with Prejudice Remover Regularizer.* P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 35–50 (2012).
Alex Kendall, Yarin Gal. *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?* NIPS 2017: 5580-5590 (2017).
Faye Kilburn. *BlackRock to use machine learning to gauge liquidity risk* (2017).
Been Kim, Rajiv Khanna, Oluwasanmi Koyejo. *Examples are not Enough, Learn to Criticize! Criticism for Interpretability.* 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain (2016).
KPMG. *AI Compliance in Control* (2019).
Tao Lei, Regina Barzilay, Tommi Jaakkola. *Rationalizing Neural Predictions.* EMNLP (2016).
Zachary C. Lipton. *The Mythos of Model Interpretability* (2017).
Hui Liu, Qingyu Yin, William Yang Wang. *Towards Explainable NLP: A Generative Explanation Framework for Text Classification*. CoRR abs/1811.00196 (2018).

Lloyd's, *Taking control - Artificial intelligence and insurance*. Emerging Risk Report (2019).

Scott M. Lundberg, Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. NIPS 2017: 4768-4777 (2017).

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, Su-In Lee. *Explainable AI for Trees: From Local Explanations to Global Understanding.* arXiv:1905.04610v1 [cs.LG] (2019).

MAS (Monetary Authority of Singapore). *Principles to Promote Fairness, Ethics, Accountability and Transparency in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector.* (2019).

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamia, Zhao Shen, Freddy Lécué. *Interpretable Credit Application Predictions With Counterfactual Explanations.* arXiv:1811.05245v2 [cs.AI] (2018).

Tim Miller. *Explanation in Artificial Intelligence: Insights from the Social Sciences* (2018).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, BenHutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. *Model Cards for Model Reporting.* In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19) (2019).

Christoph Molnar, *Interpretable Machine Learning — A Guide for Making Black Box Models Explainable* (2019).

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yua. *Interpretable machine learning: definitions, methods, and applications.* arXiv:1901.04592v1 [stat.ML] (2019).

Nicolas Papernot, Patrick McDaniel. *Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning.* arXiv:1803.04765v1 [cs.LG] (2018).

Nicolas Papernot. *A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private.* Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (2018).

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, Marcus Rohrbach. *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence.* arXiv:1802.08129v1 [cs.AI] (2018).

Keyur Patel, Marshall Lincoln. *It's not magic: Weighing the risks of AI in financial services*, Centre for the Study of Financial Innovation (2019).

James Proudman. *Cyborg supervision-the application of advanced analytics in prudential supervision*, Bank of England (2018).

PwC. *Opportunities await: How InsurTech is reshaping insurance.* Global FinTech Survey (2016).

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. *Model-agnostic interpretability of machine learning.* ICML Workshop on Human Interpretability in Machine Learning (2016).
Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Explainable NLP KDD 2016: 1135-1144 (2016).
Andrew Ross, Michael C. Hughes, Finale Doshi-Velez. *Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations.* arXiv:1703.03717v2 [cs.LG] (2017).
Lukas Ryll, Sebastian Seidens. *Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey*. arXiv:1906.07786 (2019).
Laurent Serrurier. *Un point sur l'explicabilité et l'interprétabilité en (DEEP…) Machine Learning*, IRIT (12 novembre 2018).
Blake Shaw, Tony Jebara. *Structure Preserving Embedding.* Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada (2009).
Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov. *Membership Inference Attacks Against Machine Learning Models.* 2017 IEEE Symposium on Security and Privacy (2017).
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje. *Learning Important Features Through Propagating Activation Differences*. ICML 2017: 3145-3153 (2017).
Justin Sirignano, Apaar Sadwhani, Kay Giesecke. *Deep learning for mortgage risk* (2018).
Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, Himabindu Lakkaraju. *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.* arXiv:1911.02508 [cs.LG] (2020).
Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, Alexander M. Rush. *LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks.* IEEE Trans. Vis. Comput. Graph. 24(1): 667-676 (2018).
Erik Štrumbelj, Igor Kononenko. *An efficient explanation of individual classifications using game theory*. Journal of Machine Learning Research, 11:1–18 (2010).
Tapestry Networks. *Why banks can't delay upgrading core legacy banking platforms* (2019).
Berk Ustun, Cynthia Rudin. *Supersparse Linear Integer Models for Optimized Medical Scoring Systems*. Machine Learning 102.3: 349–391 (2015).
Sandra Wachter, Brent Mittelstadt, Chris Russell. *Counterfactual explanations without opening the black box : automated decisions and the GDPR.* Harvard Journal of Law & Technology Volume 31, Number 2 Spring 2018 (2018).
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio. *Show, Attend and Tell: Neural Image Caption Generation with Visual*

Attention. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057 (2015).

Jan Ruben Zilke, Eneldo Loza Mencía, Frederik Janssen. *DeepRED – Rule Extraction from Deep Neural Networks.* In: Calders T., Ceci M., Malerba D. (Eds) Discovery Science. DS 2016. Lecture Notes in Computer Science, vol. 9956. Springer, Cham (2016).

Remerciements

Les auteurs de ce document remercient tous les participants aux travaux exploratoires décrits dans ce document (experts métiers, Data Scientists, équipes de validation et Conformité, ainsi que leurs responsables hiérarchiques) pour avoir accepté l'appel à projets et contribué activement aux discussions, présentations et ateliers pratiques.

Les auteurs remercient également leurs collègues de l'ACPR ayant contribué à la réflexion et aux analyses dont ce document est le fruit, et tout particulièrement : Jean-Philippe Barjon, Emmanuelle Boucher, Nicolas Carta, Laurent Clerc (Pôle LCB-FT), Richard Diani, Thierry Frigout, Cyril Gruffat, Gauthier Jacquemin, Boris Jaros, Matthias Laporte, Farid Oukaci, Jérôme Schmidt, Pierre Walckenaer.