Machine Learning et Modèles IRB : Avantages, Risques et Préconisations

Christophe Hurlin, Christophe Pérignon

Machine Learning et Modèles IRB : Avantages, Risques et Préconisations *

Christophe Hurlin† Christophe Pérignon ‡

23 mars 2024

Résumé

L'objectif de cette étude est de proposer une réflexion théorique et pratique sur les enjeux de l'utilisation des méthodes d'apprentissage automatique (Machine Learning) dans le cadre spécifique des modèles de risque de crédit basés sur les notations internes (IRB), permettant le calcul des fonds propres réglementaires des banques. Si le ML est aujourd'hui encore peu utilisé dans le domaine réglementaire (IRB, IFRS9, stress tests), les récentes discussions initiées par l'Autorité Bancaire Européenne (EBA) laissent penser que cet usage pourrait se développer dans le futur. Bien que techniquement complexe, ce sujet est crucial compte tenu des craintes pour la stabilité financière que soulèvent l'utilisation de modèles internes sophistiqués et opaques pour la gestion des fonds propres. A l'inverse, pour leurs défenseurs, ces modèles offrent la perspective de mieux mesurer le risque de crédit et d'ouvrir de nouvelles perspectives en termes d'inclusion financière. De cette étude, ressortent plusieurs conclusions et recommandations concernant (i) les enjeux de concurrence bancaire internationale liés aux données utilisées par ces modèles, (ii) l'amélioration de la précision dans l'estimation des paramètres de risque, (iii) la réduction des fonds propres réglementaires pour les banques, (iv) la nécessité de remettre en cause l'arbitrage entre performance et interprétabilité, et de développer dans ce contexte des modèles de ML nativement interprétables et (v) le défi de la gouvernance, des risques opérationnels et de la formation.

Mots clés : Machine Learning ; réglementation prudentielle bancaire ; modèles internes ; capital réglementaire.

Classification JEL : G21, G29, C10, C38, C55.

*Ce rapport a été réalisé le cadre de la collection Opinion et Débats de l'Institut Louis Bachelier (ILB) et financé par le Labex FCD (ANR-11-LABX-00019-016). Nous remercions Henri Fraisse, Olivier Gassner, et Guillaume Vuillemey pour leurs commentaires et suggestions. Nous remercions également l'Institut Universitaire de France (IUF), la Chaire ACPR Régulation et Risque Systémique, et l'Agence Nationale de la Recherche (MLEforRisk ANR-21-CE26-0007) pour le soutien apporté à nos recherches.

†Université d'Orléans et Institut Universitaire de France, Rue de Blois, 45067 Orléans, France. E-mail : christophe.hurlin@univ-orleans.fr

‡HEC Paris, 1 Rue de la Libération, 78350 Jouy-en-Josas, France. E-mail : perignon@hec.fr

Table des matières

1 Introduction **3**

2 Quels sont les usages possibles du ML dans le contexte IRB ? **12**

2.1 Définition du ML dans le contexte IRB 12

2.2 Les applications possibles du ML dans le contexte IRB 15

2.3 Le ML est-il utilisé dans le contexte IRB ? 19

3 Quelles données pour les modèles de ML en IRB ? **23**

3.1 ML et « nouvelles » données 24

3.2 Des enjeux de concurrence bancaire 28

4 Quels sont les bénéfices attendus du ML pour la modélisation IRB ? **29**

4.1 Des gains potentiels en termes de précision et de productivité 30

4.2 Des gains potentiels en termes de capital réglementaire 36

5 Le défi de l'explicabilité / interprétabilité des modèles IRB **41**

5.1 Quelle explicabilité / interprétabilité pour les modèles IRB ? 42

5.2 Les méthodes d'interprétabilité et d'explicabilité ex-post 43

5.3 Existe-il un arbitrage interprétabilité / performance ? 47

6 Les autres défis posés par le ML dans le contexte IRB **50**

6.1 Le challenge de la gouvernance du ML dans le contexte IRB 50

6.2 Les risques opérationnels associés au ML 52

6.3 L'évaluation de l'équité et les biais systématiques 56

7 Conclusion **59**

1 Introduction

En novembre 2021, l'Autorité Bancaire Européenne (EBA) publie un document de réflexion (EBA, 2021) portant sur l'usage des algorithmes d'apprentissage automatique (*Machine Learning* ou ML) dans le cadre des modèles de notation interne (*Internal Rating Based* ou IRB). Selon l'approche IRB de la réglementation prudentielle de Bâle, ces modèles internes sont utilisés par les banques afin de déterminer le montant des fonds propres réglementaires qu'elles doivent posséder pour couvrir leurs pertes de crédit non anticipées et garantir leur solvabilité financière. Pour l'EBA, l'objectif de cette consultation est double. [1] Il s'agit d'une part de poser les bases d'une discussion sur le sujet avec les banques, et d'autre part de définir les attentes du régulateur concernant les modèles de ML les plus sophistiqués, et d'évaluer leur conformité aux règlements et directives CRR/CRD sur les exigences de fonds propres des banques. [2] Si le sujet peut paraître technique, il est éminemment important pour les superviseurs bancaires, tant la crainte est grande de voir se reproduire les déconvenues de la crise de 2008 en autorisant les banques à utiliser des modèles très sophistiqués dans le domaine de la gestion du risque de crédit. Immédiatement après cette publication, sont apparus dans la presse spécialisée des dizaines de titres (cf. Figure 1) sur l'émergence du ML et de l'intelligence artificielle (IA) dans le calcul des fonds propres réglementaires des banques, avec l'idée que l'EBA aurait « ouvert la porte » à cet usage. Ces réactions furent parfois très tranchées, mettant en avant les dangers du ML dans ce contexte avec le spectre de modélisations mal maîtrisées au cœur même du dispositif réglementaire mis en place au milieu des années 2000, précisément pour limiter l'occurrence des crises bancaires et en limiter les effets (Allen, 2021). A l'inverse, pour leurs défenseurs, ces modèles offrent la perspective de mieux mesurer le risque de crédit et donc de permettre aux banques, soit de réduire leurs fonds propres tout en garantissant leur stabilité financière, soit de faciliter l'accès au crédit de segments de particuliers ou d'entreprises pour lesquels les risques sont surestimés par les modélisations traditionnelles. Gageons enfin que certains de ces avis

1. La consultation s'est achevée en février 2022 et l'EBA a recueilli 11 retours émanant principalement d'associations professionnelles bancaires, i.e., Asociación Española de Banca (AEB), Assilea- Associazione Italiana Leasing (ASSILEA), Association Européenne des Banques Coopératives (EACB), Groupement Européen des Caisses d'Epargne et des Banques de Détail (ESBG), Fédération Bancaire Française (FBF), et l'association de l'industrie bancaire allemande (GBIC).

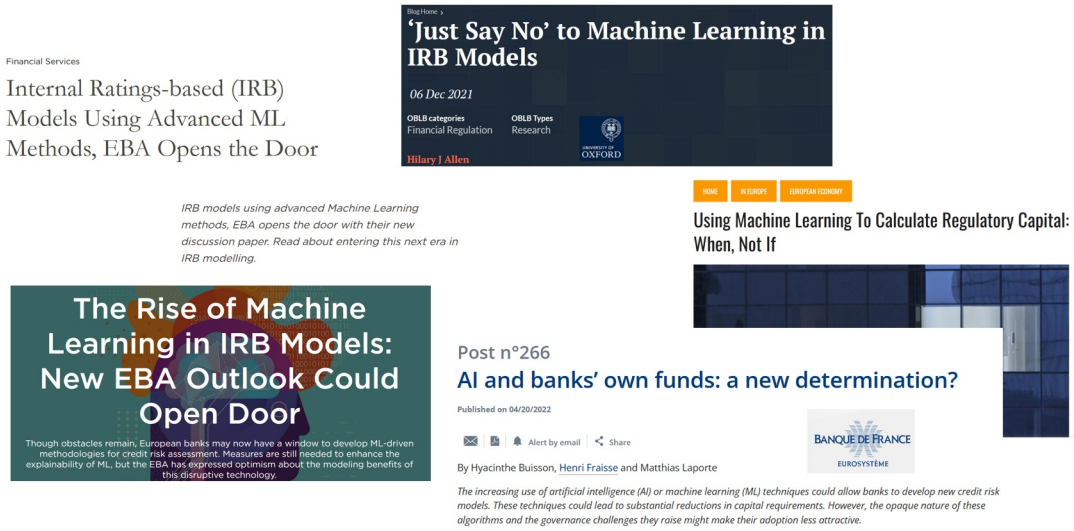
2. Les dispositions des accords de Bâle sont transposées dans le droit de l'Union européenne sous la forme d'un règlement (*Capital Requirements Regulation* ou CRR) et d'une directive (*Capital Requirements Directive* ou CRD) A ce jour, les dernières versions de ces textes sont les CRR3/CRD6 visant à transposer les réformes de Bâle III contenu dans l'accord du Comité de Bâle du 7 décembre 2017, portant sur le plancher en fonds propres, le risque de crédit, et le risque opérationnel, ainsi que l'accord de janvier 2019 portant sur le risque de marché.

reflètent des intérêts financiers, la perspective d'un développement du ML dans le contexte IRB permettant potentiellement de générer de nouvelles perspectives de croissance pour les différents acteurs chargés de développer et de maintenir ces modèles, tant en interne qu'en externe (cabinets de conseil, prestataires de services informatiques, etc.).

Mais qu'en est-il réellement de l'usage du ML dans le contexte très spécifique des modèles IRB ? Quels sont les limites, les risques et les avantages que l'on peut attendre de l'application de ces techniques dans ce domaine ? Avant de répondre à ces questions, commençons par mettre en évidence deux paradoxes. Le premier est que, contrairement à ce qui a souvent été écrit, en 2021 l'EBA n'a pas « ouvert la porte » à l'utilisation du ML, puisque cette porte était déjà largement ouverte. En effet, dans les textes réglementaires CRD/-CRR qui régissent actuellement le développement et la gouvernance des modèles internes, rien n'empêche a priori l'utilisation de modèles issus d'algorithmes de ML pour l'estimation des paramètres de risque de crédit, que ce soit pour la probabilité de défaut (PD), la perte en cas de défaut (LGD), ou le facteur de conversion de crédit (CCF). Aucun article des CRR ne proscrit, par exemple, l'utilisation d'un réseau de neurones, d'une méthode de Bagging ou de Boosting, pour la modélisation des PD. Pourtant dans la pratique, force est de constater que peu de banques utilisent le ML dans le contexte IRB, et quand elles le font, ces techniques sont souvent cantonnées en amont (prétraitement des données et des variables) ou en aval (construction de modèles challengers pour la validation des modèles internes) de la phase centrale de différenciation et de calibration des risques. Cette observation est d'autant plus surprenante que plusieurs rapports montrent que le ML est par ailleurs largement utilisé dans l'industrie bancaire (ACPR, 2018, 2020; EBA, 2020) que ce soit dans le cadre des activités marketing (connaissance des besoins des clients), des activités commerciales et de l'amélioration de la relation client (assistant bancaire, robot-conseiller, etc.), de la prévention de la fraude, de la lutte contre le blanchiment et le financement du terrorisme, de la gestion d'actifs, de la surveillance des risques de marchés, etc. Le ML est également largement utilisé dans la gestion du risque de crédit que ce soit dans le cadre des processus d'octroi et de tarification, dans le suivi des crédits, le recouvrement ou la restructuration (IIF, 2019, 2022). Paradoxalement, le ML reste très peu utilisé dans le domaine réglementaire pour le calcul des exigences en capital réglementaire (IRB), tout comme pour la construction des stress tests ou le provisionnement (IFRS 9).

Le second paradoxe est que le document de réflexion de l'EBA a été publié moins d'un mois

FIGURE 1 – Titres d'articles sur l'usage du ML dans le contexte IRB



(a) Note : Copies d'écran reprenant les titres de différents articles réagissant au document de réflexion de l'EBA sur l'utilisation du ML dans le contexte IRB.

après la publication du paquet bancaire CCR3/CRD6 qui met en œuvre les réformes dites de Bâle III finalisé ou Bâle IV. [3] Or, ce dernier avait précisément pour objectif de réduire le degré de liberté laissé aux modélisateurs dans le cadre des approches IRB. Le régulateur déplorait en effet une trop forte variabilité entre les banques, des exigences de fonds propres calculées selon les modèles internes. Pour réduire cette variabilité, le Comité de Bâle a adopté une série de réformes parachevant les accords de Bâle III, visant à introduire des mécanismes pour limiter le gain en fonds propres associé à l'utilisation de ces modèles. [4] Ainsi, le CRR3 introduit par exemple (i) des planchers en capital (*output floor*) de telle sorte que le niveau de fonds propres calculé par les modèles internes ne soit pas inférieur à 72.5% des exigences calculées en approche standard, et (ii) des valeurs planchers (*input floor*) sur les paramètres prudentiels de PD, LGD et CCF qui entrent dans la formule réglementaire pour déterminer les exigences en fonds propres. A première vue, il peut dès lors paraître paradoxal que l'EBA engage une discussion avec les banques sur l'utilisation de techniques de ML, connues pour offrir des modélisations plus flexibles que les modélisations paramétriques usuelles. En effet, cette plus grande flexibilité dans les formes fonctionnelles des modèles pourrait tout à fait

3. Le document de réflexion de l'EBA sur l'utilisation du ML dans le contexte IRB a été publié en novembre 2021, tandis que le paquet bancaire CCR3/CRD6 a été publié en octobre 2021.

4. La réforme CCR3/CRD6 vise également à restreindre l'utilisation des approches fondées sur les modèles internes, avec par exemple la suppression de l'approche basée sur des modèles internes (*Advanced Measurement Approach*) pour le risque opérationnel.

accentuer la variabilité des estimations des paramètres de risque, et in fine des exigences en fonds propres réglementaires, ce qui serait contraire aux objectifs de l'EBA. Dit autrement, comment interpréter les deux messages concomitants et potentiellement contradictoires du superviseur européen prônant « moins » de degrés de liberté dans les modélisations internes, mais qui dans le même temps laisserait la porte ouverte à « plus » de ML ?

L'objectif de cet article est de proposer une réflexion à la fois académique et pratique sur les enjeux de l'utilisation des méthodes de ML dans le cadre spécifique des modèles internes IRB. Avant toute chose, il convient de définir la notion même d'algorithme ou de modèle de ML, et de discuter de l'usage qu'il pourrait en être fait dans le cadre des modèles internes de risque.[5] La définition du ML retenue par le régulateur renvoie à l'interprétabilité des modèles : est entendu comme modèle de ML, un modèle peu ou pas interprétable (EBA, 2021). La notion d'interprétabilité fait référence ici à la capacité d'un modèle à être compris par un être humain, c'est-à-dire la capacité à comprendre comment le modèle fonctionne et comment il prend ses décisions. Dans le cadre de la modélisation IRB, un modèle de ML peut être utilisé (i) pour différents paramètres de risque (PD, LGD ou CCF) et (ii) à différentes étapes de leur modélisation, i.e., dans le cadre de la préparation des données, du prétraitement des variables (*feature engineering*), de la sélection des variables retenues par le modèle, ou de la validation du modèle. Toutefois, c'est principalement lorsque les méthodes de ML sont utilisées en tant que modèle primaire de différenciation et de calibration des risques que se pose l'essentiel des enjeux que nous allons discuter par la suite.

Nos résultats et recommandations peuvent être résumés en trois points. Premièrement, la question de l'usage du ML dans le cadre spécifique IRB ne peut pas être dissociée de la question des données à partir desquelles ces algorithmes sont entraînés et utilisés. Dans son document de réflexion, l'EBA évoque l'utilisation de nouvelles données, y compris des données non structurées. Le caractère de nouveauté s'entend ici par opposition aux données usuelles des modèles de scoring bancaire portant sur les données socio-économiques et les historiques de paiement de l'emprunteur, ou les caractéristiques du contrat. Hors périmètre réglementaire, il a été montré que de « nouvelles » sources d'information pouvaient améliorer la différenciation des risques, que ce soit pour les emprunteurs individuels

5. Un algorithme de ML est une procédure qui permet d'apprendre à partir de données, tandis qu'un modèle de ML est le résultat de l'application de cet algorithme à un ensemble de données. L'algorithme décrit la manière dont les données sont traitées et transformées afin d'obtenir un modèle prédictif qui sera utilisé en production sur de nouvelles données.

ou les entreprises, en particulier dans le cadre des processus d'octroi (Hurlin et Pérignon, 2019). La combinaison d'algorithmes de ML et de ces nouvelles données permet en effet de sélectionner de nouveaux prédicteurs du défaut et d'identifier des risques qui auraient été ignorés par les approches traditionnelles. L'usage de ces nouvelles données peut également améliorer l'inclusion financière en permettant, par exemple, de mieux évaluer les risques de sous-populations d'emprunteurs n'ayant pas ou peu d'historique de crédit, e.g., les jeunes emprunteurs ou les entreprises nouvellement créées. Cependant, la transposition de ces résultats, obtenus dans le cadre des procédures d'octroi, au contexte réglementé des modèles IRB, ne va pas de soi, et pose de nombreux problèmes de gouvernance et de mise en œuvre. Un premier défi réside dans la collecte et la gestion des données requises pour entraîner ces modèles de ML, tout en respectant les contraintes imposées par les CRR. Par exemple, comment garantir de pouvoir disposer d'un historique sur 5 ans conformément aux CRR, alors que ces nouvelles données n'ont généralement pas de profondeur historique ? De plus, les modèles internes constituant le « cœur du réacteur » de l'activité bancaire, les banques sont fortement réticentes à mobiliser des données ou des scores produits par un prestataire externe dans le cadre de leurs modèles de PD, LGD ou CCF. Ainsi, ce débat sur l'usage de nouvelles sources de données couplées au ML, pourrait n'être que théorique, si la question de l'IA n'apparaissait pas comme une priorité stratégique pour les régulateurs bancaires. L'objectif ici est de permettre aux banques de rester au plus près de la frontière technologique et de ne pas laisser un oligopole technologique se constituer au profit de nouveaux acteurs. La question de l'utilisation du ML dans le contexte IRB touche en effet des enjeux de concurrence bancaire internationale et de barrières à l'entrée pour les nouveaux acteurs que pourraient être les Fintechs ou les Bigtechs. Réguler l'utilisation du ML dans le cadre IRB est alors crucial si l'on souhaite éviter qu'un acteur puisse tirer avantage d'un accès privilégié à des données pour distordre la concurrence bancaire, soit de façon directe en offrant du crédit moins coûteux en fonds propres, soit au travers de l'organisation d'une plateforme de services bancaires (Clerc *et al.*, 2020).

Notre deuxième résultat porte sur les potentiels bénéfices du ML à ensemble d'information inchangé, i.e., sans utilisation de nouvelles données. Que pouvons-nous attendre des modèles de ML dans le contexte IRB lorsque ces modèles sont utilisés en tant que modèles primaires ? Pour la plupart des praticiens, il est admis que les modèles de ML fournissent des estimations/prévisions plus précises des paramètres de risque que les approches paramétriques

usuelles, e.g., la régression logistique. [6] Pour autant, il convient d'être prudent quant à cette conclusion. Concernant la modélisation de la PD, la littérature académique montre clairement (i) que seules les méthodes d'ensemble (Bagging et Boosting notamment) permettent d'améliorer significativement les estimations des paramètres de risque par rapport aux approches standards (Lessmann *et al.*, 2015), et (ii) que ces gains prédictifs ont tendance à plafonner très rapidement avec la complexité des modèles. Par exemple, des travaux récents (Gunnarsson *et al.*, 2021) montrent que dans le cadre du scoring bancaire, l'apprentissage profond (*Deep Learning*) n'apporte pas de gain significatif en comparaison des méthodes d'ensemble jusqu'à présent utilisées. Ce résultat s'explique par le fait que, contrairement à d'autres applications (par exemple dans le domaine de la santé), l'évaluation de la solvabilité d'un emprunteur est un problème relativement « simple » dans lequel les non-linéarités et les interactions entre les facteurs explicatifs sont peu marquées. C'est pourquoi les banques privilégient les méthodes d'ensemble qui tirent profit d'une réduction de la variance des estimations par un mécanisme d'agrégation de prédicteurs individuels (méthodes de Bagging, e.g., les forêts aléatoires) ou selon une démarche itérative de surpondération des erreurs de prévision (méthodes de Boosting, e.g., XGBoost). En ce qui concerne la modélisation de la LGD, la littérature académique montre que le ML peut également améliorer la précision des estimations (Loterman *et al.*, 2012), à condition toutefois d'être utilisé dans le cadre de modélisations emboîtées tenant compte des particularités de la distribution des pertes en cas de défaut. Toutefois, les gains prédictifs issus du ML pour l'estimation de la LGD sont généralement plus faibles que ceux obtenus pour la PD.

Si la grande flexibilité des modèles de ML leur permet potentiellement d'améliorer la différenciation des risques et de fournir des estimations plus précises des paramètres Bâlois, rien ne dit que ces gains prédictifs se transformeront nécessairement en des gains en capital réglementaire pour les banques. Supposons que, suite à la substitution d'une régression logistique par un modèle XGBoost en tant que modèle primaire de différenciation des PD, on observe une amélioration de l'AUC ou de tout autre critère statistique d'évaluation du modèle. Rien ne garantit en effet que cette amélioration se traduise par une baisse des PD estimées pour l'ensemble, ou pour une partie, des crédits du portefeuille. L'amélioration de la capacité prédictive du modèle ne signifie en rien que la régression logistique conduisait à une surestimation des PD. De plus, on peut imaginer que les PD estimées diminuent

6. Au sens strict, les modèles sont utilisés pour fournir des *prévisions* des paramètres de risque Bâlois, mais la pratique veut que l'on parle plutôt d'*estimation* des paramètres de risque.

pour des crédits ayant de faibles expositions (*Exposure At Default* ou EAD) ou de faibles LGD, et augmentent pour les autres. C'est pourquoi, une méthode de ML peut tout à fait réduire l'AUC d'un modèle de PD ou le R^2 d'un modèle de LGD, sans pour autant nécessairement conduire à une diminution des niveaux d'actifs pondérés par le risque (*Risk Weighted Assets* ou RWA). Cette divergence potentielle des critères d'évaluation statistique et économique est illustrée dans le cas de la LGD par l'étude de Hurlin *et al.* (2018). Cette étude montre que les classements de différents modèles alternatifs de LGD qui peuvent être obtenus sur la base de critères statistiques (MSE, MAE, etc.) sont parfois très différents de ceux obtenus à partir de fonctions de perte définies en termes de capital réglementaire. De façon générale, la question des gains en fonds propres est donc beaucoup plus ouverte que celle des gains en termes de précision d'estimation, car non seulement les approches de ML doivent être comparées en termes de gain/pertes en capital réglementaire, mais de plus ces modèles doivent également passer les tests de validation internes et les tests de validation externes qu'imposent le régulateur. Or, c'est la question centrale qui pourrait potentiellement convaincre les banques de changer leurs modèles internes pour adopter le ML. Peu d'études académiques sont disponibles sur le sujet, mais nous montrons que les premières évaluations empiriques (Alonso et Carbó, 2020; Fraisse et Laporte, 2022) tendent à confirmer la perspective d'une réduction sensible des fonds propres que pourraient offrir certains modèles de ML, tout en garantissant de satisfaire les tests de validation usuels.

La troisième recommandation porte sur la question de l'interprétabilité. [7] Il est évident que la principale raison de la non-utilisation des modèles de ML dans le contexte réglementaire, tient à leur complexité qui entraîne des difficultés de compréhension et d'interprétation des résultats. Cette difficulté pose la question centrale de l'exigence d'interprétabilité des modèles IRB que ce soit vis-à-vis de la gouvernance interne des banques ou du régulateur, et de la conformité de ces modèles au regard des CRR. Les modèles de ML peuvent aider à estimer les paramètres de risque avec plus de précision mais cette amélioration de la précision se fait souvent au prix d'une plus grande complexité, puisque la relation entre les variables d'entrée et de sortie du modèle est plus difficile à évaluer et à comprendre. Le compromis entre pouvoir prédictif et complexité / interprétabilité est donc au cœur de la décision de l'établissement d'utiliser ou non des modèles ML à des fins prudentielles. Le défi consiste ici à articuler la relation entre les modèles de ML et le jugement humain

7. Dans la section 5.1, nous discuterons la différence entre les notions d'interprétabilité et d'explicabilité, mais à ce stade de l'exposé, nous n'évoquerons que la première.

concernant la différenciation des risques. Cela pose clairement la question des exigences en matière d'interprétabilité du régulateur. Quelle interprétabilité, pour quels objectifs et quels interlocuteurs ? Ici deux visions s'opposent. La première consiste à utiliser des modèles de ML nativement non-interprétables, e.g., une forêt aléatoire ou un modèle XGBoost, et à mobiliser ex-post des techniques permettant d'interpréter et d'expliquer leurs prévisions. Ces méthodes peuvent être axées sur une explication globale du modèle (PDP, ICE, ACE, etc.) ou sur des explications individuelles construites à l'échelle de chaque crédit (LIME, SHAP, etc.). [8] Ces explications peuvent être fondées sur une analyse graphique (PDP), une évaluation numérique (LIME), ou une approche théorique (SHAP). C'est cette démarche en deux temps, i.e, ML non-interprétable et méthode d'interprétabilité ex-post, qui est habituellement retenue par les banques et qui est discutée dans le document de réflexion de l'EBA. La difficulté de cette approche est qu'il n'existe pas de méthode universelle d'interprétabilité, chacune de ces méthodes renvoyant à une explication particulière du fonctionnement du modèle (Krishna *et al.*, 2022). Dès lors, chaque explication prise individuellement peut ne pas répondre pleinement aux attentes d'un modélisateur, d'un valideur interne ou d'un superviseur (Bracke *et al.*, 2019). Par ailleurs, ces différentes méthodes peuvent fournir des explications divergentes (Rudin, 2019) et/ou instables, par exemple, dans le cas de portefeuilles de crédits avec peu de défauts observés (Chen *et al.*, 2022).

C'est pourquoi, nous recommandons une approche alternative qui consiste à mobiliser des approches de ML qui soient performantes, tout en étant nativement interprétables. Cette approche suppose de remettre en cause l'existence, par ailleurs largement admise dans l'industrie, d'un arbitrage entre performance prédictive et interprétabilité. Sur le plan théorique, des travaux récents fondés sur la notion d'espace de Rashomon (Semenova *et al.*, 2022), montrent en effet que pour tout problème de classification ou de régression, il existe très probablement un modèle nativement interprétable tout aussi performant que le meilleur modèle de type boîte noire. Toutefois, même si ce résultat théorique est avéré, d'aucuns pourraient penser qu'il est toujours plus facile en pratique de mettre en œuvre un modèle de type boîte noire que de chercher un hypothétique modèle, à la fois performant et interprétable. Cependant, plusieurs exemples montrent que tel n'est pas le cas. Dumitrescu *et al.* (2022) proposent ainsi une approche hybride d'évaluation du crédit, dite PLTR (*Penalised Logistic Tree Regression*), qui vise à améliorer les performances prédictives du modèle de régression

8. Les acronymes et le principe de ces méthodes seront discutés dans la section 5.1.

logistique par le biais d'une transformation préalable des variables explicatives fondée sur des arbres de décision peu profonds. Ce modèle de régression logistique augmenté, tout en étant parfaitement interprétable, permet d'obtenir des performances prédictives quasiment équivalentes à une forêt aléatoire. Des travaux récents de Flachaire *et al.* (2022) proposent une approche similaire, fondée cette fois-ci sur des modèles additifs généralisés couplés à une procédure de sélection automatique des variables. Les résultats montrent que ces modèles nativement interprétables, dits GAM(L)A, offrent des performances prédictives similaires aux principaux modèles de type boîte noire utilisés par les banques. Dans le même esprit, un modèle additif généralisé nativement interprétable, proposé par Chen *et al.* (2018), a remporté le challenge du ML interprétable organisé par FICO (*Fair Isaac Corporation*), le leader américain des scores de crédit, face à des propositions alternatives toutes fondées sur des modèles non-interprétables. Cette démarche de ML nativement interprétable est actuellement utilisée par plusieurs banques américaines. [9] Cette démarche pourrait clairement constituer une voie de développement du ML dans le contexte IRB. Une des conséquences de cette évolution vers un ML nativement interprétable, est qu'il convient dès lors de déplacer la discussion sur l'interprétabilité des prévisions du modèle (approche classique du ML interprétable) vers une discussion portant davantage sur l'interprétabilité de la performance prédictive des modèles de ML (Hué *et al.*, 2022).

Le simple fait qu'un régulateur adopte un point de vue prudemment positif sur le ML est une étape très importante qui ouvre de nouveaux défis, au-delà de celui de l'interprétabilité. Un de ces défis réside dans les risques opérationnels liés à la mise en œuvre des modèles de ML, comme par exemple ceux liés à la fixation des hyperparamètres. On peut également évoquer la question des biais de discrimination qui pourraient survenir dans ce type de modèles. Les modèles de scores de crédit ont ainsi été classifiés parmi les systèmes « à haut risque » dans le cadre de la proposition européenne de règlement sur l'IA (EC, 2021), ou *Artificial Intelligence Act*, et ils seront sans doute soumis à des exigences accrues en termes de transparence, de supervision et de responsabilité pour les entreprises qui développent ou utilisent ces systèmes. Dans ce cadre, la question de l'équité algorithmique concerne principalement les processus d'octroi de crédit, mais les modèles internes étant nécessairement liés aux modèles d'octroi, la question des biais systématiques des modèles IRB doit également

9. Dans une publication récente (Sudjianto et Zhang, 2021), Agus Sudjianto, directeur des risques pour les crédits aux entreprises de Wells Fargo, présente les principes de conception pour le développement de modèles de ML interprétables à haute performance, tels qu'ils sont mis en œuvre par sa banque dans le cadre des processus d'octroi et de suivi du risque de crédit.

être abordée. Enfin, l'utilisation des algorithmes de ML soulève la question d'une gouvernance spécifique permettant d'évaluer l'ensemble des risques, de la phase de construction à la phase de mise en œuvre opérationnelle. Ainsi, le questionnement sur l'utilité du ML dans le contexte IRB s'inscrit dans le débat plus général sur l'IA de qualité, les liens entre performance et fiabilité, ou entre transparence et interprétabilité, l'équité et l'absence de parti pris, la responsabilité et sur le respect de la vie privée.

2 Quels sont les usages possibles du ML dans le contexte IRB ?

Dans cette section, nous définissons la notion de modèle de ML et nous discutons les usages qui pourraient être faits de ces outils dans le contexte IRB. Nous montrons que ces modèles ne sont actuellement que peu utilisés par les banques pour les aspects réglementaires (IRB, IFRS9, stress-tests réglementaires, etc.), alors même qu'ils sont très largement utilisés dans d'autres contextes non-réglementaires de l'analyse du risque de crédit (octroi, suivi des risques, etc.).

2.1 Définition du ML dans le contexte IRB

Le ML se définit de façon générale comme un domaine de l'informatique qui traite du développement de modèles dont les paramètres sont estimés automatiquement à partir de données, avec une intervention humaine limitée ou nulle (EBA, 2020). Le ML regroupe donc l'ensemble des algorithmes qui permettent d'apprendre en identifiant des relations entre des données et de produire des modèles prédictifs de manière autonome (ACPR, 2018). Selon ces définitions générales, adossées aux standards internationaux, la plupart des méthodes statistiques et des modèles économétriques utilisés par les banques relèvent du ML.[10] Par exemple, la régression logistique peut être assimilée au ML, tout comme les arbres de classification ou de régression, qui a priori ne posent pas de problème en termes d'interprétabilité.

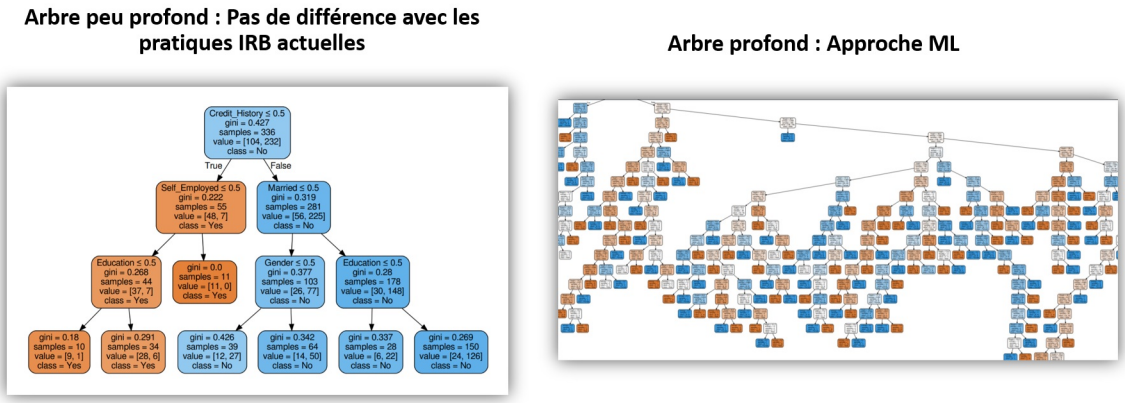
Toutefois, dans le contexte spécifique des modèles IRB, il est convenu de restreindre le terme de ML aux seuls modèles les plus complexes, qui sont les plus difficiles à interpréter et à utiliser à des fins de détermination du capital réglementaire. Ainsi, dans son document de réflexion, l'EBA limite la définition du ML « *aux modèles qui se caractérisent par un*

10. La norme ISO relative à la gouvernance des données définit le ML comme « un processus utilisant des algorithmes plutôt qu'un codage procédural qui permet d'apprendre à partir de données existantes afin de prédire les résultats futurs » (Standard on IT governance ISO/IEC 38505-1/2017).

grand nombre de paramètres et qui, par conséquent, nécessitent un grand volume de données (potentiellement non structurées) pour leur estimation et qui sont capables de refléter des relations non linéaires entre les variables » (EBA, 2021). Cette définition recoupe l'ensemble des modèles non interprétables de type boîte noire, e.g., une méthode de Bagging de type forêt aléatoire ou un réseau de neurones multicouche, ainsi que les modèles qui seraient intrinsèquement interprétables, mais trop complexes pour être intelligibles.

La notion de complexité d'un modèle en ML n'est pas univoque et il existe une abondante littérature consacrée à sa mesure. La complexité peut être appréciée selon différentes approches théoriques telles que la théorie de Vapnik-Chervonenkis (Vapnik, 2013) ou de celle du rasoir d'Ockham. Mais si l'on se limite au cas simple d'un arbre de décision, il est reconnu que la complexité peut être mesurée par sa profondeur et le nombre de nœuds terminaux. Ainsi, comme l'illustre la Figure 2, un arbre de décision de faible profondeur ne relèverait pas de la définition du ML dans le contexte IRB au sens de l'EBA, tandis qu'au contraire un arbre très profond avec de nombreuses feuilles terminales relèverait du ML, puisqu'il nécessite l'estimation de nombreux paramètres et qu'il est difficilement compréhensible du fait de la complexité des règles de décision induites. Ainsi, c'est la complexité algorithmique et le manque d'interprétabilité qui définissent la notion de ML dans le contexte spécifique des modèles IRB.

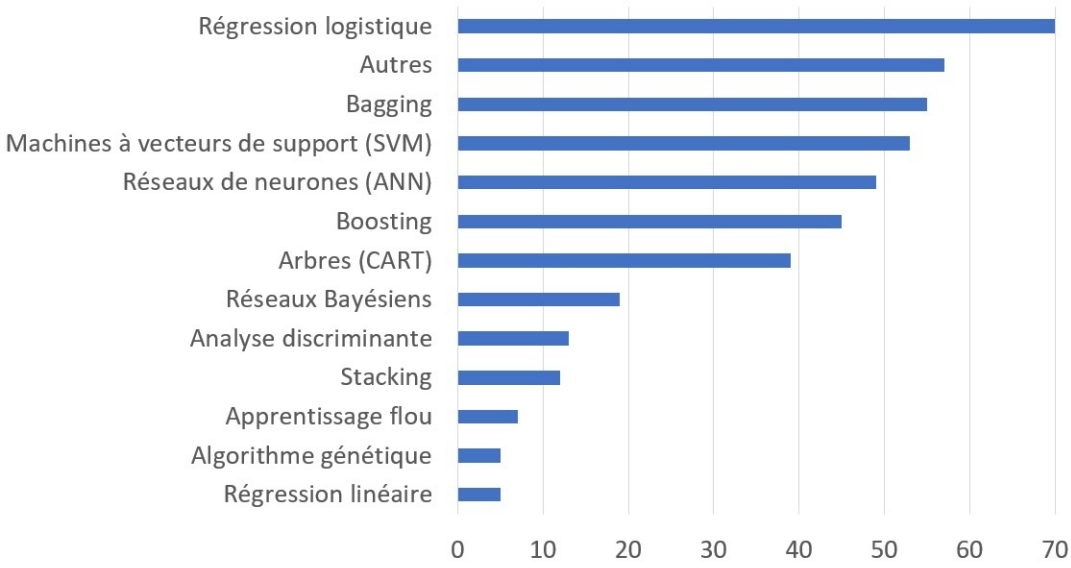
FIGURE 2 – Illustration de la notion de complexité : le cas des arbres de décision



De très nombreux algorithmes de ML répondent à ces définitions et il est difficile d'établir une liste exhaustive de ceux effectivement utilisés pour la modélisation du risque de crédit, ne serait-ce que dans les applications hors périmètre réglementaire. Par exemple, Lessmann *et al.* (2015) recensent pas moins de 41 algorithmes de classification différents utilisés pour

la modélisation de la PD, dans 48 articles académiques publiés entre 2003 et 2015. Dans une étude plus récente, Markov *et al.* (2022) analysent 110 articles publiés entre 2016 et 2020 pour identifier les algorithmes de ML, les critères d'évaluation, et les méthodes des prétraitement les plus couramment utilisés dans la littérature académique sur le scoring de crédit. Comme le montre la Figure 3, ils montrent que, hormis la régression logistique considérée dans la plupart des publications comme un benchmark, les méthodes de ML les plus utilisées sont les méthodes de Bagging, les machines à vecteurs de support (SVM), les réseaux de neurones (ANN) et les méthodes de Boosting.[11] Dans le cas de la modélisation de la LGD, les algorithmes les plus utilisés sont les arbres de régression, les forêts aléatoires, les méthodes de Gradient Boosting, les ANN, et les SVR (Loterman *et al.*, 2012; Yao *et al.*, 2017; Hurlin *et al.*, 2018).

FIGURE 3 – Les méthodes de ML les plus utilisées dans le scoring de crédit



(a) Note : cette figure indique le nombre de fois où un algorithme de ML a été utilisé dans les 110 articles considérés dans l'étude de Markov *et al.* (2022), chaque article pouvant utiliser plusieurs algorithmes. Source : Markov *et al.* (2022).

Au-delà des seules études académiques, il n'existe pas de recension des méthodes de ML effectivement utilisées par les banques pour la modélisation du risque de crédit que ce soit dans un contexte non réglementaire (octroi) ou réglementaire (IRB, IFRS9). Sans prétendre

11. Une autre méta-analyse des méthodes de scoring de crédit, portant sur de 74 articles publiés entre 2010 et 2018, a été proposée par Dastile *et al.* (2020). Cette étude donne des résultats légèrement différents de ceux Markov *et al.* (2022), puisque les algorithmes les plus utilisés dans l'échantillon sont dans l'ordre, la régression logistique, les SVM, les ANN, les arbres de décision, le Boosting, les forêts aléatoires, puis les autres méthodes de Bagging (AdaBoost, etc.).

à l'exhaustivité, nous pouvons toutefois lister quelques algorithmes de ML utilisés par les banques à partir des réponses qu'ont apportées les principales fédérations bancaires européennes au document de réflexion de l'EBA (cf. Encadré 1).

Encadré 1 : Les principaux algorithmes de ML utilisés dans le contexte IRB

A la suite de la publication de son document de réflexion sur l'usage du ML dans le contexte IRB (EBA, 2021), l'EBA a lancé une consultation publique sur le sujet. Une des questions posées portait sur la nature des méthodes de ML qui étaient actuellement utilisées, ou que les banques envisageaient d'utiliser, dans le contexte IRB. [a] En réponse à cette consultation, le comité du secteur bancaire allemand (GBIC) a mentionné l'utilisation de méthodes de clusterisation de type K-means, d'algorithmes de type ANN ou de forêt aléatoire pour la modélisation de la PD, de méthodes des plus proches voisins (k-NN) et de différents modèles de régression pour la LGD, et des arbres de régression pour le CCF. L'association bancaire espagnole (AEB) revient sur la difficulté qu'il y a de lister les algorithmes utilisés par les banques, mais évoque les forêts aléatoires et les méthodes de type *Gradient Boosting Trees* (GBT). L'association italienne des entreprises de leasing (Assilea) évoque l'utilisation de réseaux de neurones pour les modèles d'évaluation sur les risques immobiliers et de méthodes de *Gradient Boosting* pour les autres applications de crédit. Enfin, la fédération bancaire française (FBF) n'évoque succinctement que les seules forêts aléatoires, sans préciser le contexte d'application.

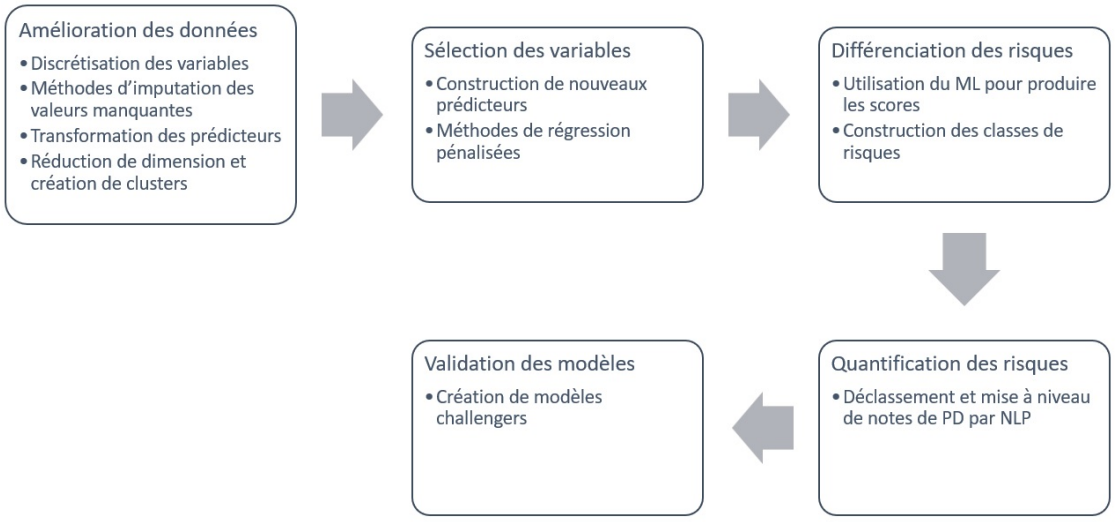
a. La question 1.3. était libellée de la façon suivante : « *Veuillez préciser le type de modèles et d'algorithmes de ML (par ex. forêt aléatoire, etc.) que vous utilisez actuellement ou que vous prévoyez d'utiliser dans le contexte de l'IRB ?* », (EBA, 2021).

2.2 Les applications possibles du ML dans le contexte IRB

Dans le contexte IRB, les algorithmes de ML peuvent être mobilisés tant pour la modélisation de la PD, de la LGD ou du CCF. Si l'on se limite dans un premier temps aux seules applications concernant la PD, le ML peut être utilisé à différents stades de la modélisation, cette dernière pouvant être résumée en trois étapes. La première étape, dite de *différenciation des risques*, consiste à utiliser un modèle pour estimer un score continu exprimé en fonction de facteurs de risque, puis d'en déduire une échelle de risque discrète. Les crédits sont alors regroupés en classes homogènes de risque, garantissant l'homogénéité du risque de défaut à l'intérieur d'une classe et l'hétérogénéité de ce risque entre les classes. La deuxième étape, dite de *quantification des risques*, consiste à affecter à chaque classe de risque homogène, une probabilité de défaut. Généralement, ces probabilités sont estimées par

la fréquence empirique des défauts observée sur longue période au sein de chaque classe. Ce sont ces probabilités de défaut qui constituent les données clés pour le calcul des exigences de fonds propres. La troisième étape, dite de *validation*, vise à s'assurer de la cohérence et de la stabilité des prévisions de défaut issues du modèle interne. L'EBA distingue différentes applications possibles pour les modèles de ML dans ces différentes phases de la modélisation IRB, comme le résume la Figure 4.

FIGURE 4 – Exemples d'utilisations possibles du ML dans les phases de la modélisation IRB



La préparation des données. Les algorithmes de ML peuvent être utilisés en amont de la modélisation, dans les phases de prétraitement des données et de *feature engineering*. Ces phases de traitement préalable à la modélisation ont pour objet d'augmenter la robustesse des modèles (e.g., traitement des valeurs manquantes et des valeurs aberrantes), de faciliter l'interprétation des résultats du modèle (e.g., discrétisation des variables continues), de supprimer les informations redondantes (suppression ou réduction des ensembles de variables fortement corrélées), de présélectionner les variables candidates à la modélisation (analyse des dépendances à la variable cible), et/ou d'augmenter les capacités prédictives des modèles (e.g., regroupement de classes ou optimisation de la construction des classes sur les variables explicatives du modèle). Il est reconnu que ces phases de prétraitement constituent une étape essentielle pour augmenter les capacités prédictives des modèles de scoring bancaire (Verdonck *et al.*, 2021). Concernant la phase de *feature engineering*, les méthodes les plus couramment utilisées sont généralement des techniques statistiques de transformation des variables univariées (transformation de Box-Cox pour les variables continues, codage par

dummy, codage par percentile, transformation des variables catégorielles polytomiques en variables binaires, etc.) ou multivariées (analyse en composante principale, analyse discriminante linéaire, etc.). [12] Mais de plus en plus de techniques de ML sont également mobilisées dans cette phase de prétraitement des données (Dastile *et al.*, 2020). On peut citer ici par exemple les auto-encodeurs, i.e., un type de réseau neuronal à trois couches permettant de réduire la dimension de l'ensemble de représentation des données. On peut également mentionner les techniques de ML utilisées pour l'imputation des valeurs manquantes, comme les méthodes d'imputation itérative par réseau Bayésien (BNII), ou les méthodes de Bagging de type XGBoost et CatBoost (Markov *et al.*, 2022). Dans de nombreuses banques, des arbres de régression ou de classification sont également utilisés pour discrétiser ou regrouper les variables explicatives des modèles internes (Dumitrescu *et al.*, 2022).

La sélection des variables. Un des principaux atouts des algorithmes de ML résident dans leur capacité à sélectionner les variables explicatives pertinentes et leurs combinaisons dans un grand ensemble de données. C'est l'une des principales différences avec les approches économétriques classiques pour lesquelles la spécification du modèle repose sur un raisonnement économique préalable et des tests statistiques (Mullainathan et Spiess, 2017; Charpentier *et al.*, 2018). A l'inverse, les algorithmes de ML permettent de construire automatiquement un modèle prédictif à partir des données, sans procédure d'inférence a priori, en sélectionnant et combinant les variables explicatives. A titre d'exemple, l'algorithme CART (*Classification And Regression Tree*) de Breiman *et al.* (1984) permet de construire un arbre de classification permettant de partitionner les individus de l'échantillon d'apprentissage en groupes d'individus homogènes du point de vue de la variable cible au sens d'une mesure d'impureté (entropie de Shannon, indice d'impureté de Gini, etc.). Pour ce faire, l'algorithme sélectionne automatiquement les prédicteurs et les seuils associés à chaque noeud de décision. Au final, c'est donc l'algorithme qui détermine, sans intervention humaine, quelles seront les variables qui seront incluses dans le modèle prédictif et qui estime les paramètres de seuil associés à chaque noeud. En cela, les algorithmes de ML s'apparentent aux méthodes de sélection automatique des variables explicatives (stepwise, backward, autometrics, etc.) qui existent pour les modèles économétriques. Mais les algorithmes de ML permettent de gérer un très grand nombre de variables et surtout, permettent également de créer automatiquement de nouvelles variables par transformation des variables initiales. De façon générale, l'utilisation

12. Voir Dastile *et al.* (2020) pour une recension des différentes méthodes de *feature engineering* utilisées dans la littérature académique consacrée aux modèles de score de crédit.

du ML dans la phase de sélection des variables peut se faire de façon concomitante à la modélisation (e.g., lorsqu'un algorithme de ML est utilisé pour la différenciation des risques) ou dans une démarche de présélection de variables candidates visant à réduire la dimension du problème de classification ou de régression. Dans ce dernier cas, la sélection se fait souvent au travers de méthodes de régressions pénalisées, par exemple, de type Lasso (Tibshirani, 1996) ou Adaptative Lasso (Zou, 2006).

La différenciation des risques. L'EBA distingue clairement le degré d'attention qui doit être consacré aux méthodes de ML suivant les types d'applications qui en sont faits dans le contexte IRB. Il est évident qu'une moindre attention est requise lorsque les modèles ML sont utilisés en amont ou en aval de la différenciation des risques. Les enjeux les plus importants apparaissent lorsque le ML est employé en tant que modèle primaire pour (i) construire un score continu en fonction de facteurs de risque et/ou (ii) regrouper les crédits suivant ces scores afin d'identifier les classes de risque homogène sur lesquels seront calibrées les PD. C'est sur ce champ d'application crucial que constitue la différenciation des risques, que nous focaliserons la suite de notre analyse.

La quantification des risques. Les méthodes de ML ne sont généralement pas utilisées pour calibrer les paramètres de risque au sein des classes de risque, puisque cette calibration se fait sur la base de fréquences de défaut observées sur longue période pour le cas de la PD, ou de moyenne de long-terme des taux de perte pour la LGD. La quantification des risques ne nécessite donc généralement pas de modèle prédictif. Toutefois, l'EBA évoque ici des applications potentielles du ML qui pourrait être utilisé pour réaliser des mises à niveau ou des déclassements de note de PD sur la base, par exemple, de données textuelles préalablement traitées par des techniques de traitement automatique du langage (NLP).

La validation de modèle. Une des utilisations les plus courantes du ML dans le contexte IRB (avec la phase de préparation des données) consiste à l'utiliser dans la phase de validation du modèle de différenciation des risques. Concrètement, cela revient souvent à développer des modèles de ML dits « challengers », qui servent à évaluer les estimations des paramètres de risque, ou plus directement les RWA et les exigences en fonds propres réglementaires, fournies par les modèles internes en production. Puisque ces modèles challengers ne sont pas utilisés directement pour le calcul des exigences en fonds propres réglementaires, ils échappent aux contraintes imposées par les CRR. Il s'agit ainsi de l'application du ML la plus facile à mettre en œuvre dans le cadre IRB, notamment en termes de gouvernance

et de conformité réglementaire.

Encadré 2 : Les usages du ML dans les différentes phases de la modélisation IRB

Les usages potentiels du ML dans le contexte IRB reportés par les associations bancaires européennes sont assez variés. [a] L'association bancaire espagnole (AEB) mentionne qu'il existe des possibilités d'utiliser le ML dans la sélection des facteurs de risque, l'imputation des variables manquantes, la combinaison et la transformation des variables, le contrôle de la qualité des données, la construction de modèles challengers. La fédération bancaire française (FBF) indique que les modèles ML sont principalement utilisés pour la sélection des variables, la différenciation des risques et la construction de modèles challenger. L'association bancaire allemande (GBIC) distingue les utilisations du ML pour la modélisation de la PD (buckteting) de celles pour la LGD, où ces techniques sont principalement utilisées dans la phase de différenciation des risques. L'association évoque également l'imputation des valeurs manquantes et la détection d'anomalies dans les données, e.g., la détection de biais systématiques. L'association italienne des entreprises de leasing (ASSILEA) mentionne l'utilisation du ML dans le calcul des paramètres de LGD, au travers notamment de l'évaluation du collatéral.

a. La question 1.2 est libellée de la façon suivante : « *Pouvez-vous préciser à quelles fins spécifiques les modèles ML sont utilisés ou envisagés ? Veuillez préciser à quel stade du processus d'estimation ils sont utilisés, c'est-à-dire préparation des données, différenciation des risques, quantification des risques, validation* » (EBA, 2021).

On retrouve ces différents usages potentiels du ML dans les réponses faites par les principales associations professionnelles bancaires européennes à l'EBA (cf. Encadré 2), sans pour autant qu'il se dégage de norme en la matière. Cette absence de consensus sur les usages tient sans doute au fait que, pour la plupart des banques, l'usage du ML dans le contexte IRB reste encore à l'état de projet.

2.3 Le ML est-il utilisé dans le contexte IRB ?

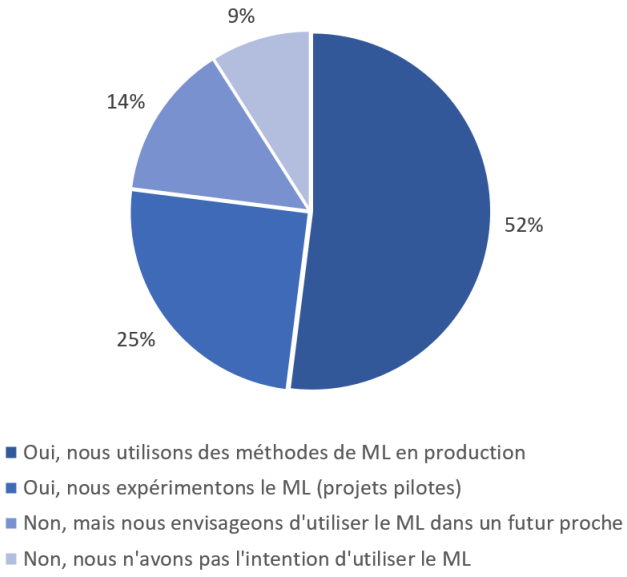
La question est de savoir si les techniques de ML sont effectivement utilisées par les banques dans l'une ou l'autre des phases de la modélisation IRB. Malheureusement, il existe peu d'études à ce sujet. Toutefois, l'*Institute of International Finance* (IIF) a récemment mené deux enquêtes sur l'utilisation du ML dans l'industrie du risque de crédit (IIF, 2019, 2022), qui permettent de dresser un premier constat de ces usages. [13] De ces deux études, il ressort

13. La première enquête (IIF, 2019) interroge les représentants de 60 banques internationales sur leur utilisation du ML pour l'analyse du risque de crédit entre 2018 et 2019. L'échantillon regroupe 10 banques américaines, 4 du Canada, 14 de la zone euro, 7 du reste de l'Europe, et 9 dont les sièges sont en Asie. La seconde enquête (IIF, 2022), menée sur un échantillon de 43 banques internationales entre janvier et septembre 2022, porte sur les usages du ML à la fois pour le risque de crédit et pour la lutte contre

un paradoxe apparent : le ML est aujourd'hui largement utilisé par les banques internationales dans la gestion de leur risque de crédit hors périmètre réglementaire (octroi, suivi des risques, recouvrement), mais le ML reste, au contraire, très peu utilisé dans le contexte réglementaire (IRB, IFRS9, stress tests).

Utilisation du ML dans la gestion du risque de crédit. L'enquête IIF (2022) montre en effet que 52% des banques de l'échantillon utilisent en *production* des modèles de ML pour la gestion du risque de crédit et que 25% des institutions expérimentent actuellement l'usage de ces techniques, comme le montre la Figure 5. Seules 9% des banques interrogées n'utilisent pas le ML et n'envisagent pas de le faire. Le taux de pénétration est encore plus fort dans la zone euro et aux États-Unis, puisque toutes les banques de l'échantillon utilisent actuellement en production des modèles de ce type pour la gestion du risque de crédit.

FIGURE 5 – Utilisation du ML pour la gestion du risque de crédit

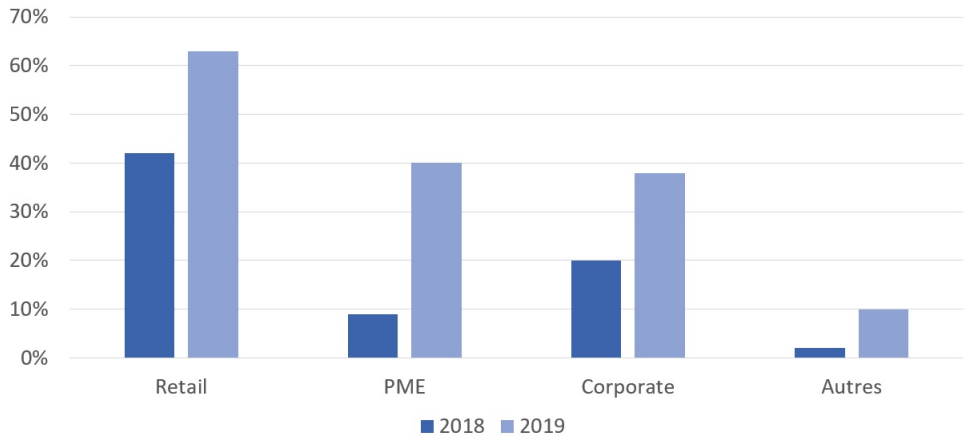


(a) Note : Dans l'enquête de l'IIF, la question posée était libellée de la façon suivante « Utilisez-vous des techniques de ML dans vos analyses liées au risque de crédit ? », Source : IIF (2022).

le blanchiment des capitaux (AML). L'échantillon regroupe 8 banques de la zone euro, 7 d'autres pays d'Europe, 3 d'Amérique latine, 4 des États-Unis, 2 du Canada, 4 du Japon, 3 de Chine et 5 dont les sièges se situent au Moyen orient ou en Afrique.

Les enquêtes de l'IIF montrent également que la plupart des banques qui utilisent le ML le font déjà depuis longtemps et dans plusieurs fonctions du processus de risque de crédit, i.e., le nettoyage des données, la sélection des variables, l'exploration et la segmentation des données, le développement de modèles et la validation des modèles. Enfin, le ML est utilisé pour tous les types de portefeuilles comme le montre la Figure 6, même si c'est principalement dans le cadre de l'activité de crédit aux particuliers (*retail*) que l'on trouve le plus fort taux d'utilisation.

FIGURE 6 – Les types de portefeuilles concernés par l'usage du ML

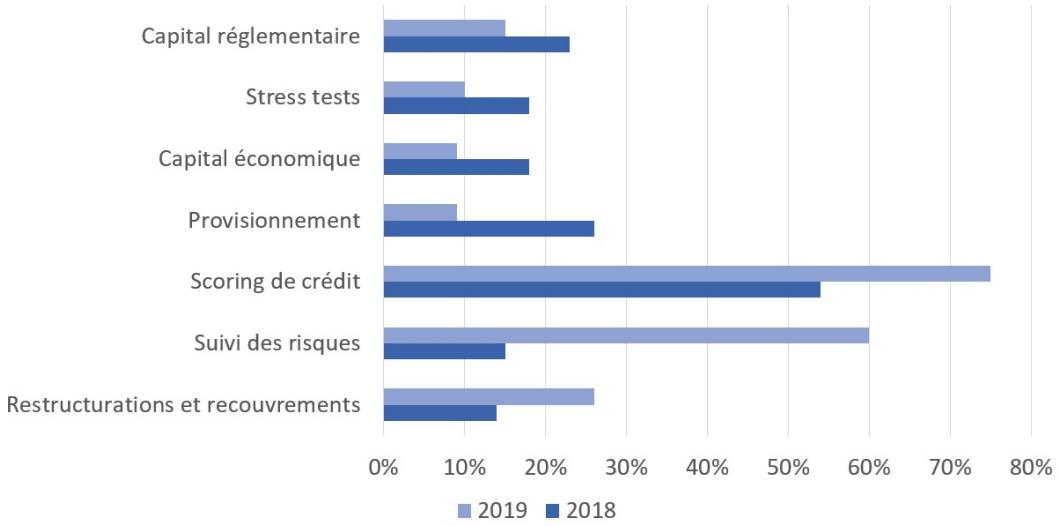


(a) Source : IIF (2019)

Utilisation du ML dans le contexte IRB. Alors que les banques utilisent largement le ML pour la gestion du risque avec une expertise avérée, ces techniques restent encore très peu utilisées dans le domaine IRB. Comme le montre la Figure 7, seulement 10% des banques interrogées déclarent utiliser le ML en production, à un stade ou un autre de leur modélisation IRB (IIF, 2019). Pire, l'utilisation du ML est même en recul, puisque ce taux atteignait 15% en 2018. Si l'on ajoute les banques déclarant envisager d'utiliser le ML pour le calcul des RWA, les taux passent respectivement à 15% en 2019 contre 23% en 2018. Ces taux d'utilisation du ML sont à rapprocher de ceux déclarés par les banques en 2019 pour leurs modèles d'octroi (75%), leurs modèles de suivi des risques (60%), et pour les modèles de restructuration et de recouvrement (26%). Par ailleurs, pour ces trois types de modèles non-réglementaires, on observe une tendance à l'augmentation de l'utilisation du ML par les banques entre 2018 et 2019. Il est intéressant de noter que la faible utilisation du ML pour les modèles internes IRB, se retrouve plus généralement dans toutes les applications qui touchent aux aspects réglementaires, comme par exemple les stress tests (10% de taux

d'utilisation effectif ou envisagé) ou le provisionnement (8% de taux d'utilisation effectif ou envisagé).

FIGURE 7 – Les usages du ML dans la gestion du risque de crédit



(a) Source : IIF (2019)

Il est évident que les banques ont aujourd'hui suffisamment de connaissances et d'expérience pour développer des modèles de ML dans le contexte IRB, et plus généralement dans toutes les applications réglementaires. Elles utilisent déjà largement cette technologie dans d'autres domaines qui ne nécessitent pas l'approbation du régulateur (ACPR, 2018). Dès lors, comme l'indiquent les réponses des associations professionnelles bancaires à la consultation de l'EBA (cf. Encadré 3), c'est principalement la question de l'acceptabilité par le régulateur de ces modèles, notamment en lien avec la question de l'interprétabilité, qui semble limiter leur usage. Sachant que le développement des modèles IRB est un processus qui demande beaucoup de temps et de ressources, les banques semblent attendre de disposer de critères clairs sur le processus d'approbation de ces modèles en lien avec les exigences réglementaires.

Encadré 3 : L'utilisation des modèles de ML dans le contexte IRB

Dans le cadre la consultation menée par l'EBA, plusieurs questions portaient sur l'utilisation effective des méthodes de ML dans le contexte IRB. [a] Les réponses apportées par les associations professionnelles bancaires européennes confirment les résultats des enquêtes de l'IIF. La fédération bancaire française (FBF) et l'association bancaire espagnole (AEB) confirment qu'à l'heure actuelle, leurs membres n'utilisent pas le ML dans la phase de différenciation des risques en IRB. Les raisons avancées tiennent principalement à l'incertitude quant à l'acceptabilité de la part du régulateur, ainsi qu'aux difficultés d'interprétation et de mise en œuvre de ces techniques. La FBF mentionne que le ML est toutefois utilisé par certaines banques pour l'estimation de la LGD, typiquement au travers de forêts aléatoires, et pour la validation des modèles. L'AEB mentionne également le fait que certaines entités financières ont présenté à l'autorité de surveillance des modèles de ML dans des applications d'IRB, et que ces applications sont actuellement en cours de validation. L'association bancaire allemande (GBIC) constate, elle aussi, que les modèles de ML sont très peu utilisés dans le contexte IRB. Pour elle, une des raisons tient au fait que les portefeuilles sont très homogènes et que le pouvoir prédictif de la probabilité de défaut basé sur des méthodes de régression usuelles (e.g., analyse discriminante, régression logistique) est déjà excellent.

a. Les intitulés des questions étaient les suivants : (1) Utilisez-vous actuellement ou prévoyez-vous d'utiliser des modèles de ML dans le contexte de l'IRB dans votre institution ?, (1.1) Pour l'estimation de quels paramètres votre établissement utilise-t-il actuellement ou prévoit-il d'utiliser des modèles ML, c'est-à-dire PD, LGD, ELBE, EAD, CCF ?

3 Quelles données pour les modèles de ML en IRB ?

Par essence, le ML ne peut pas être dissocié des données sur lesquelles les modèles sont entraînés (échantillon d'apprentissage) et évalués (échantillon de test), puisque ce sont ces données à partir desquelles les algorithmes détermineront in fine, la forme fonctionnelle du modèle. Ainsi, lorsque l'on discute de l'apport du ML dans le cadre des modèles internes, il est important de dissocier ce qui relève des techniques de ML en tant que telles, de l'apport de « nouvelles » données qui pourraient être rendues exploitables par ces techniques. Dans son document de réflexion, l'EBA évoque ainsi l'utilisation de données non structurées. Dans cette section, nous discutons quelles pourraient être ces données, leur apport potentiel, ainsi que les enjeux en termes de gouvernance et de concurrence bancaire qui leur sont associés.

3.1 ML et « nouvelles » données

L'idée selon laquelle le ML permettrait d'exploiter de nouvelles données améliorant la différenciation des risques n'est pas spécifique au contexte IRB, puisqu'on la retrouve de façon générale dans la réflexion sur le renouveau des modèles de scoring bancaire (Hurlin et Pérignon, 2019). [14] Pour certains, la recherche sur le développement de nouvelles méthodes de ML visant à améliorer les modèles de scoring de crédit, est plus ou moins au point mort, et la seule façon d'améliorer ces modèles consiste à exploiter de nouvelles sources de données (Oskarsdottir *et al.*, 2019). Mais quelles sont ces nouvelles sources de données ? Le caractère de nouveauté doit s'entendre ici par référence aux données traditionnelles d'un modèle de score portant typiquement sur la nature du crédit, les caractéristiques de l'emprunteur (âge, revenus, situation matrimoniale, etc) ou son historique bancaire. Ces nouvelles données sont d'origines très variées, mais elles proviennent généralement de l'open banking mis en œuvre dans le cadre de la Directive européenne sur les systèmes de paiement (DSP2), de la digitalisation de la relation clientèle, des réseaux sociaux, etc. Hurlin et Pérignon (2019) utilisent à ce propos le terme de « data-diversité » pour rendre compte de l'hétérogénéité de ces nouvelles sources d'information. Or, c'est justement cette hétérogénéité qui rend difficile l'appréciation générale de leur apport, que ce soit dans le contexte de l'octroi de crédit, ou dans le contexte réglementaire.

La question centrale est de savoir si ces nouvelles données permettent de capter des signaux faibles du défaut, et favorisent in fine l'accès au crédit d'individus ou d'entreprises jusque-là considérés comme trop risqués par les modèles traditionnels. Cette amélioration potentielle de l'inclusion financière passe de façon évidente par le processus d'octroi, mais elle pourrait également se faire via une réduction des exigences en fonds propres, autorisant une plus grande prise de risque de la part des établissements de crédit. Dans le cadre des modèles d'octroi de crédit, plusieurs études montrent que ces nouvelles données permettent en effet de révéler des signaux faibles qui améliorent sensiblement la qualité de l'évaluation de la solvabilité des emprunteurs. Citons ici quelques exemples. Berg *et al.* (2020) analysent le score de crédit d'une société de E-commerce qui utilise des données d'empreinte numérique laissée par les clients lors de leur achat en ligne telles que le type d'appareil utilisé, le système d'exploitation, le type d'adresse mail, l'heure de connexion, etc. Ils montrent que

14. Dans cette section nous limiterons notre analyse aux modèles de défaut, i.e., aux modèles de scoring bancaire.

ces informations permettent d'améliorer les performances du modèle de score en comparaison d'un modèle basé uniquement sur les données socio-économiques des clients.

Oskarsdottir *et al.* (2019) montrent, quant à eux, que les données de téléphonie mobile constituent une source d'information très riche pour évaluer les probabilités de défaut sur des crédits à la consommation, notamment pour les emprunteurs les plus fragiles. Les données de leur étude, fournies par une banque et un opérateur de télécommunications, comprennent les journaux d'appels des clients de la banque, ceux des personnes avec lesquelles ils sont en contact, l'historique bancaire et les données socio-démographiques de ces clients. L'étude couvre un an et demi d'historique bancaire pour plus de deux millions de clients de la banque et recense l'activité d'appels de près de 90 millions de numéros de téléphone uniques sur une période de cinq mois. Ces journaux d'appels sont utilisés pour construire des réseaux d'appels. De ces réseaux sont déduits des indicateurs statistiques (indicateurs de liens, mesures de type PageRank, etc.) qui sont ensuite utilisés comme variables explicatives pour améliorer la performance des modèles de prévision de la solvabilité des demandeurs de cartes de crédit. Les modèles de score sont de simples régressions logistiques, des arbres de classification ou des forêts aléatoires. Plutôt que de chercher un raffinement méthodologique dans les modèles de score, l'idée générale consiste ici à exploiter l'homophilie des réseaux de communications vis-à-vis de la caractéristique de défaut total ou partiel. Les résultats de l'étude montrent clairement que les modèles construits avec les caractéristiques qui représentent le comportement d'appel sont plus performants, à la fois en termes d'AUC et de profit, que les modèles basés uniquement sur les historiques bancaires et les données socio-démographiques. Sur la base des mesures d'importance (e.g., réduction moyenne de l'impureté dans les arbres de décision), les auteurs montrent que les caractéristiques de réseaux sont les variables plus importantes du modèle. Ce résultat montre que la façon dont les gens utilisent leur smartphone pourrait être utilisée comme seule source de données pour décider si un prêt doit leur être accordé ou non. Bien évidemment, une telle approche pose des problèmes éthiques puisqu'elle pourrait, par exemple, conduire à refuser un prêt à emprunteur sur la simple observation d'un lien préalable avec d'autres personnes ayant connu des incidents de paiement, sans que l'emprunteur en soit responsable. C'est pourquoi, les auteurs mentionnent que ce type d'information ne doit être utilisé uniquement que dans un cadre strictement positif afin d'améliorer l'accès au financement des individus dont le crédit aurait été potentiellement refusé sur la base d'informations traditionnelles, e.g., les

jeunes emprunteurs sans historique bancaire. Enfin, une autre façon de montrer la qualité de ces nouvelles sources d'information consiste à comparer les scores entre eux. Jagtiani et Lemieux (2019) montrent ainsi que la corrélation entre les scores propriétaires de la plateforme de crédit LendingClub et les scores FICO obtenus pour un échantillon de crédits comparables est passée de $0,80$ pour les prêts contractés en 2007 à moins de $0,35$ pour les prêts contractés en 2015.

Ces nouvelles données peuvent être collectées et traitées directement par les banques, mais elles peuvent l'être également par l'intermédiaire de Fintechs qui construisent et revendent des scores aux banques. Ces derniers étant ensuite utilisés comme input complémentaire dans leurs propres modèles de score. On peut citer ici l'exemple de la Fintech Big Data Scoring qui permet aux banques d'intégrer dans leurs modèles d'octroi des données de médias sociaux relatives à l'entreprise et/ou à son gérant. Ces scores sont parfois construits de façon à offrir aux banques de nouveaux outils pour développer l'inclusion financière. Par exemple, la Fintech Zest AI, fondée par d'anciens cadres de Google, propose la solution ZAML (*Zest Automated Machine Learning*) qui permet de construire un score à partir d'informations très disparates pour évaluer la solvabilité des jeunes clients ayant des antécédents de crédit très limités. L'utilisation de ces nouvelles sources de données et du ML permet alors d'augmenter les taux d'acceptation, tout en contrôlant les risques et en limitant également les éventuels biais de discrimination. [15]

Toutefois, la transposition de ces résultats obtenus dans le contexte de l'octroi, au contexte réglementé des modèles IRB, ne va pas de soi, même si bien évidemment les deux types de modèles sont liés. La principale raison tient aux contraintes réglementaires imposées par les CRR sur les données mobilisées pour la modélisation IRB, qui supposent notamment d'en garantir l'exactitude, l'exhaustivité et l'adéquation. Ces contraintes sont souvent incompatibles avec la collecte, la transformation et l'utilisation des nouvelles sources de données. Par exemple, les CRR imposent aux établissements de crédit d'estimer les PD par classe de risque à partir des moyennes de long terme des taux de défaut à un an et d'utiliser une période d'observation historique d'au moins cinq ans. De la même façon, les estimations de la LGD doivent être fondées sur des données couvrant une période minimale de cinq

15. Que ces nouvelles données soient opérées directement par les banques ou par des prestataires, elles ne nécessitent pas toujours l'utilisation de techniques de ML, ce qui justifie la distinction entre nouvelles données (*New Data*) et méga-données (*Big Data*). Par exemple, Berg *et al.* (2020) résument les caractéristiques de l'empreinte digitale d'un client à une dizaine de variables seulement, qu'ils intègrent ensuite dans un modèle de régression logistique classique pour évaluer leur solvabilité.

ans. Or, pour de nombreuses données alternatives, de tels historiques ne sont tout simplement pas disponibles. Par ailleurs, en cas d'externalisation de la collecte et du traitement de ces données, comment garantir la qualité des traitements et la pérennité dans le temps des sources de données externes ? En cas d'utilisation d'un score alternatif produit par un tiers, la banque cliente devrait veiller à la conformité des nouvelles sources de données et des traitements réalisés par le prestataire, ce qui complexifierait d'autant l'évaluation de la conformité de ses modèles IRB. C'est sans doute pourquoi, dans les réponses à l'EBA, aucune association professionnelle bancaire européenne ne fait état d'une externalisation, voire d'un projet d'externalisation, par un de leurs membres de leur modélisation IRB. [16]

Encadré 4 : Les « nouvelles » données utilisées par les banques dans le contexte IRB

Dans le cadre de la consultation lancée par l'EBA, une question portait sur l'utilisation de données alternatives non structurées dans la modélisation IRB. [a] En réponse, l'association bancaire espagnole (AEB) insiste sur l'intérêt que peut représenter ce type de données, notamment pour exploiter les informations contenues dans des rapports financiers des entreprises, au travers de techniques de NLP et de ML. L'AEB avance que des contrôles suffisants peuvent être mis en place pour garantir que les données résultantes répondent aux exigences réglementaires de qualité et d'exhaustivité requises dans le contexte IRB. L'AEB souhaite voir se développer à l'échelle européenne des initiatives visant à certifier la qualité de ces données externes. A l'inverse, les associations bancaires françaises (FBF) et allemandes (GBIC) ne font état d'aucune initiative allant dans le sens de l'utilisation de ce type de données. De même, l'ESBG (*European Savings and Retail Banking Group*) ne mentionne une telle initiative, même si l'association insiste sur les perspectives que pourraient revêtir l'utilisation de ces données et encourage la recherche sur ce sujet.

a. La question 1.4 était libellée de la façon suivante : « Utilisez-vous ou prévoyez-vous d'utiliser des données non structurées pour ces modèles de ML ? Dans l'affirmative, veuillez préciser le type de données ou le type de sources de données que vous utilisez ou prévoyez d'utiliser. Comment garantissez-vous une qualité de données adéquate ? », (EBA, 2021).

Ainsi, les seules sources de nouvelles données qui semblent exploitables dans le contexte IRB sont, soit des données collectées dans le cadre des nouvelles réglementations environnementales, soit des données non structurées, le plus souvent des données textuelles, actuellement en possession des banques, mais qui ne sont pas encore intégrées dans les modèles internes. [17]

16. La question posée par l'EBA était « Avez-vous externalisé ou envisagez-vous d'externaliser le développement et la mise en œuvre des modèles de ML et, dans l'affirmative, pour quelle phase de modélisation ? Quels sont les principaux défis auxquels vous êtes confrontés à cet égard ? ».

17. Cette observation rejoint une des conclusions de l'étude de Guégan et Hassani (2018) qui tendait à promouvoir l'utilisation des données non structurées pour élargir l'ensemble d'information des modèles réglementaires.

On peut évoquer par exemple les données liées aux émissions de gaz à effet de serre, telles que les mesures scopes 1, 2 et 3, pour les crédits aux entreprises, ou les diagnostics de performance énergétique (DPE) pour les crédits immobiliers. Dans le cas des crédits aux entreprises, on peut également penser à l'exploitation par NLP des rapports extra-financiers, et ce d'autant plus que la réglementation européenne est en passe d'imposer un audit extra-financier à un grand nombre d'entreprises. Certaines associations bancaires professionnelles évoquent ces perspectives (cf. Encadré 4), mais sans pour autant que ces modèles soient effectivement mis en œuvre, voire en cours de développement.

3.2 Des enjeux de concurrence bancaire

Dès lors, si aucune banque n'utilise, ou n'envisage d'utiliser, ces nouvelles sources de données pour leurs modèles internes, comment comprendre que le régulateur s'intéresse à cette question ? Une explication possible tient aux enjeux de concurrence liés à l'exploitation de ces données par de *nouveaux acteurs* qui pourraient venir bouleverser, à terme, le marché du crédit. Un nouvel acteur pourrait-il profiter d'un avantage informationnel et distordre la concurrence bancaire, que ce soit au travers d'un processus d'octroi augmenté plus performant, ou grâce à des modèles internes permettant d'offrir du crédit moins coûteux en fonds propres, tout en respectant les exigences réglementaires ?

Les premiers acteurs auxquels l'on peut penser dans ce contexte sont les néobanques et les Fintechs qui opèrent directement ou indirectement du crédit. Dans leur vaste étude, Clerc *et al.* (2020) distinguent quatre générations de néobanques depuis l'apparition de ces entités à la fin des années 90. Cependant, hormis pour celles appartenant à la dernière génération et qui proposent des offres nativement mobiles (N26, Revolut, etc.), la plupart des néobanques en France dépendent directement du système bancaire traditionnel, soit suite à des rachats, soit parce que ce dernier les a directement créées. Par ailleurs, ces entreprises ne disposent généralement pas d'un accès privilégié à des données qui pourraient leur permettre de tirer un avantage informationnel bouleversant la concurrence bancaire. Ainsi, en France jusqu'à présent, les néobanques (y compris celles de dernière génération) n'ont pas radicalement chamboulé le marché du crédit et elles pâtissent encore d'une faible rentabilité.

La question de l'avantage informationnel se pose donc beaucoup plus dans la perspective de l'arrivée sur le marché du crédit des « Bigtechs », i.e., des géants de l'internet GAFAM (Google Amazon, Facebook (Meta), Apple, Microsoft) et BAXT (Baidu, Alibaba, Xiaomi et

Tencent). Comme le notent Clerc *et al.* (2020), la coopération entamée par les Bigtechs avec de nombreux acteurs financiers, leur capacité à mobiliser un ensemble très vaste de données sur leurs abonnés et leur recours au ML, conduisent à anticiper une emprise croissante de ces plateformes sur le secteur financier. Cette emprise peut passer par une offre de crédit développée en propre. On peut citer ici l'exemple d'Alibaba qui a été un pionnier sur ce marché avec son application de paiement Alipay, qui propose à ses utilisateurs des micro-crédits à la consommation depuis 2015. On peut également citer le cas de la carte de crédit *Apple Card* développée par Apple, lancée en 2019. Mais l'arrivée des Bigtechs pourrait également passer par le développement de plateformes s'analysant comme des marchés « bifaces » (Rochet et Tirole, 2003) et mettant en contact des consommateurs et des offreurs de services financiers. Dans ce schéma, la plateforme en situation de quasi-monopole, se rémunère en faisant payer des droits d'entrée aux deux parties du marché et en percevant des commissions sur toutes les transactions réalisées. En outre, la plateforme pourrait monétiser les informations dont elle dispose sur les consommateurs, en les revendant aux prestataires financiers afin que ces derniers puissent améliorer leurs processus d'octroi, de suivi des risques, voire leurs modèles internes réglementaires.

4 Quels sont les bénéfices attendus du ML pour la modélisation IRB ?

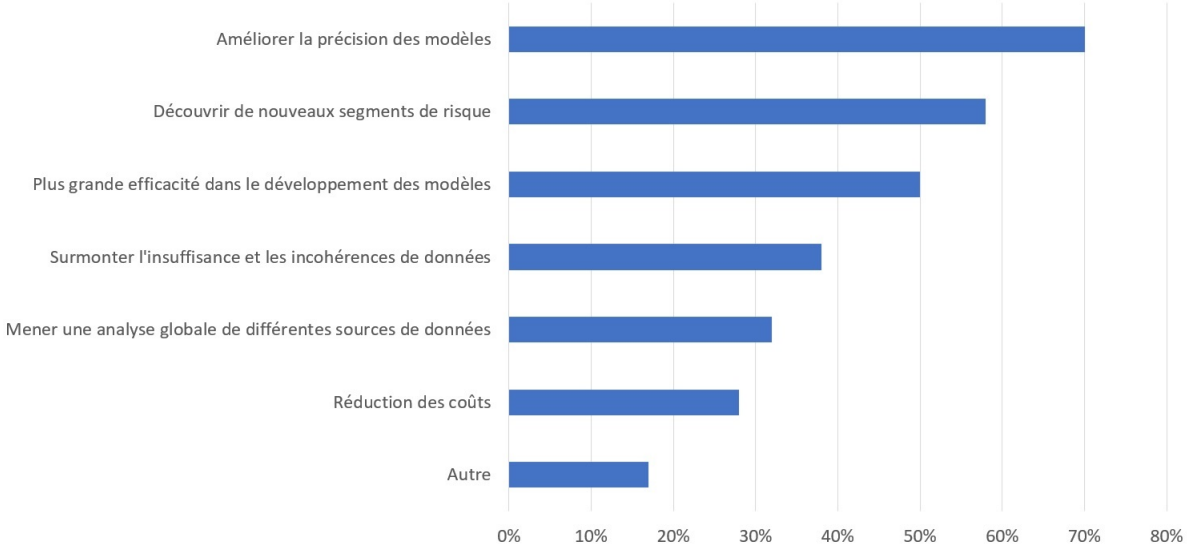
La principale idée qui sous-tend l'utilisation du ML dans le contexte IRB consiste à penser que ces techniques pourraient rendre le crédit moins « coûteux » en termes de capital réglementaire, tout en garantissant le même niveau de couverture aux banques face aux pertes non anticipées. Ceci serait rendu possible grâce à une meilleure anticipation du risque de solvabilité des emprunteurs. Cette idée repose sur deux présupposés. Premièrement, les modélisations internes de PD ou de LGD fondées sur une utilisation directe ou indirecte des techniques de ML permettraient d'obtenir des gains prédictifs par rapport aux approches classiques. Typiquement, le ML permettrait de mieux prévoir le défaut, au minimum pour certaines sous-populations du portefeuille de crédits, en captant des signaux faibles que ne captent pas les modèles standards. Deuxièmement, ces gains en précision se traduiraient nécessairement par des gains en termes de capital réglementaire, tout en permettant de respecter les exigences du régulateur en matière de validation des modèles. Nous allons voir que si la première assertion peut être globalement validée sous certaines conditions, il n'en

va pas nécessairement de même pour la seconde. Dans la suite de l'analyse, nous raisonnerons à ensemble d'information constant (sans nouvelles données) et nous nous focaliserons principalement sur l'adoption du ML en tant que modèle primaire de différenciation des risques.

4.1 Des gains potentiels en termes de précision et de productivité

Comme le montrent les résultats de l'enquête IIF (2019), lorsque l'on interroge les banques sur les bénéfices attendus du ML dans la gestion du risque de crédit (cf. Figure 8), les deux réponses qui arrivent en tête sont d'une part l'amélioration de la précision des estimations des paramètres de risque et la mise en perspective de nouveaux segments de risques, et d'autre part les gains de productivité dans le déploiement des modèles et le prétraitement des données.

FIGURE 8 – Quels sont les gains attendus du ML pour la gestion du risque de crédit ?



(a) Source : IIF (2019)

Une amélioration de la précision des estimations des paramètres de risque. Les avantages du ML par rapport aux méthodes paramétriques classiques sont bien connus (Mullainathan et Spiess, 2017; Charpentier *et al.*, 2018). Ces algorithmes ont la capacité de sélectionner de façon flexible la forme fonctionnelle du lien entre la variable cible (typiquement une variable binaire associée au défaut) et les caractéristiques disponibles dans la base d'apprentissage. En cela, le ML s'apparente à l'économétrie semi-paramétrique. La principale différence étant que les algorithmes de ML peuvent être appliqués en présence d'un

grand nombre de prédicteurs. De même, comme nous l'avons mentionné, les algorithmes de ML sélectionnent de façon autonome les variables qui entrent dans la spécification du modèle et s'apparentent ainsi aux méthodes automatiques de sélection de variables et de spécification des modèles économétriques. Enfin et surtout, ces algorithmes peuvent créer de nouveaux prédicteurs en combinant et/ou en transformant les prédicteurs initiaux. Ainsi, par exemple, un algorithme de SVM permet de projeter un problème de classification dans un espace de plus grande dimension par le biais d'une fonction kernel. Cette transformation kernel revient implicitement à créer de nouvelles variables explicatives par transformation des variables initiales. [18]

Ces propriétés rendent les modèles de ML particulièrement efficaces pour détecter de façon autonome des non-linéarités et des interactions entre les prédicteurs. La question est de savoir si la prise en compte de ces non-linéarités et de ces interactions permet d'obtenir des gains prédictifs. Afin de répondre à cette question, il convient de distinguer deux générations d'algorithmes de ML utilisés dans le cadre du scoring de crédit. La première génération correspond aux méthodes de classification supervisée dites « individuelles » qui visent à partitionner l'espace des prédicteurs afin de prévoir l'événement. Les plus utilisées dans le domaine du risque de crédit sont les arbres de classification, les SVM, et les ANN. La seconde génération correspond aux méthodes d'ensemble qui utilisent une agrégation ou une combinaison d'un grand nombre de méthodes individuelles d'apprentissage dans le but de réduire la variance du modèle et d'éviter le surapprentissage.

Dans la littérature académique, le ML a très tôt été appliqué à la modélisation du risque de crédit. Ainsi, l'algorithme CART (*Classification and Regression Tree*) a été utilisé pour une application de scoring de crédit moins d'un an après la publication de l'article de Breiman *et al.* (1984), par Makowski (1985), puis par Carter et Catlett (1987) et Srinivasan et Kim (1987). De la même façon, les premières applications des réseaux de neurones dans ce contexte datent du début des années 90 (Tam et Kiang, 1992; Altman *et al.*, 1994). Mais très rapidement, il est apparu que ces classifieurs individuels ne permettaient pas d'améliorer significativement les performances prédictives de la régression logistique, comme le montre la première synthèse de la littérature académique sur le scoring de crédit de Thomas (2000).

18. L'avantage de cette transformation est qu'il est plus facile d'identifier un hyperplan séparateur optimal dans un espace de représentation de plus grande dimension. La difficulté réside dans le fait qu'on ne sait pas définir explicitement les nouvelles variables transformées qui ont été introduites par l'algorithme dans le modèle en plus des variables initiales, d'où un problème d'interprétabilité.

Dans une étude comparative basée sur 17 algorithmes et 8 bases de données, Baesens *et al.* (2003) confirment que les SVM ou les ANN offrent de très bonnes performances prédictives. Cependant, pour la plupart des bases, les différences entre les AUC de la meilleure méthode de ML et celle de la régression logistique sont inférieures à 2%, ce qui est très faible. La principale conclusion que l'on peut tirer de ces études, est que le scoring de crédit est un champ d'application dans lequel il y a trop peu de non-linéarités dans les données pour que les gains de performances prédictives du ML soient significatifs. Ce sont sans doute ces premiers résultats qui expliquent la faible adoption des techniques de ML par les banques avant le début des années 2010 pour leurs modèles de risque de crédit.

Le changement intervient avec l'utilisation des premières méthodes d'ensemble qui vont permettre d'obtenir des gains prédictifs significatifs. Les méthodes d'ensemble les plus couramment utilisées dans le domaine du risque de crédit sont le Bagging (contraction des termes *Bootstrap Aggregation*) et le Boosting, introduits respectivement par Breiman (1996) et Freund et Schapire (1997). Les méthodes de Bagging (e.g. Random Forests, Rotation Forest) reposent sur des techniques d'échantillonnage par bootstrap et d'agrégation qui visent à améliorer la qualité prédictive de modèles à faible pouvoir de généralisation, comme les arbres de décisions. [19] Sous certaines conditions, il est démontré théoriquement que la combinaison de ces classifieurs individuels obtenus sur des échantillons ré-échantillonnés permet de réduire la variance du classifieur agrégé, et donc d'en améliorer la qualité prédictive sur un échantillon test. Les méthodes de Boosting (e.g., XGBoost, AdaBoost) reposent quant à elles sur une stratégie adaptative pour « booster » les performances prédictives d'un prédicteur faible. [20] L'idée générale consiste à entraîner de façon itérative un modèle de sorte à réduire les erreurs de prévision obtenues à chaque étape. Certaines méthodes de Boosting, comme la méthode XGBoost (*Extreme Gradient Boosting*) développée par Chen et Guestrin (2016), sont aujourd'hui très populaires dans les compétitions de classification (Kaggle, etc.).

19. Le Bagging consiste à entraîner un ensemble d'arbres de classification sur des sous-échantillons ré-échantillonnés par bootstrap, i.e., comportant des individus tirés au hasard parmi l'échantillon initial. La réduction de la variance obtenue est alors une fonction décroissante de la corrélation entre les arbres individuels entraînés sur ces échantillons ré-échantillonnés. La méthode des forêts aléatoires (Random Forest) a pour objectif de réduire encore plus cette corrélation en ré-échantillonnant à la fois les individus de la base d'apprentissage, mais aussi les variables candidates à la division binaire.

20. Un classifieur AdaBoost est un méta-modèle qui ajuste un prédicteur faible, typiquement un arbre de faible profondeur, sur l'ensemble de données original, puis entraîne des copies supplémentaires de ce prédicteur sur le même ensemble de données, mais en surpondérant les instances mal classées de sorte que les classifieurs suivants se concentrent davantage sur ces erreurs de prévision.

Ces méthodes d'ensemble figurent parmi les techniques de ML les plus utilisées par les banques, notamment pour la modélisation de la PD. Plusieurs dizaines d'articles scientifiques et d'études comparatives montrent en effet que ces méthodes permettent de mieux identifier les risques de défaut en comparaison des modèles paramétriques usuels. L'étude comparative la plus complète sur la question est celle de Lessmann *et al.* (2015) qui évaluent 41 algorithmes de classification, dont de nombreux algorithmes avancés de ML, sur 8 bases de données de crédits aux particuliers, avec de nombreux critères d'évaluation statistiques et économiques. Ils montrent clairement que les méthodes d'ensemble prévoient le risque significativement mieux que la régression logistique. Les forêts aléatoires dominent ainsi systématiquement les classifieurs individuels, que ces derniers soient paramétriques (régression logistique) ou de type ML (réseaux de neurones, SVM, etc.). Toutefois, les auteurs montrent également que les gains de performance prédictive permis par le ML ont tendance à plafonner avec la complexité des modèles. Les raffinements méthodologiques des algorithmes de ML n'améliorent pas nécessairement les performances. Par exemple, les AUC obtenues avec des algorithmes de type Rotation Forests ne diffèrent pas de celles obtenues avec de simples forêts aléatoires.

On retrouve un résultat similaire chez Alonso et Carbó (2020) qui opposent les gains prédictifs du ML aux coûts supplémentaires engendrés en termes de supervision. Leur analyse de la littérature montre qu'au-delà des méthodes d'ensemble homogènes de type Bagging et Boosting, les évolutions méthodologiques de type apprentissage profond (Deep Learning), apprentissage par renforcement, ou méthodes d'ensemble hétérogènes, conduisent à des résultats mitigés et ne permettent pas d'améliorer significativement la prévision du défaut. Gunnarsson *et al.* (2021) montrent également que les algorithmes de Deep Learning ne semblent pas être des modèles appropriés pour le scoring de crédit.[21] Leur étude comparative menée sur 10 bases de données, conclut que les méthodes de Deep Learning ne sont généralement pas les plus performantes, alors que leur mise en place est considérablement plus coûteuse en termes de temps de calcul que les méthodes d'ensemble. Dans leur étude, il ressort que la méthode XGBoost est la méthode la plus performante.

Il convient de noter que ces évaluations académiques des modèles de ML ne tiennent pas compte des prétraitements qui sont généralement appliqués par les banques dans le cas des

21. Les auteurs utilisent des algorithmes de type Deep Belief Network (DBN) et un réseau de neurones multicouche profond.

modélisations paramétriques standards.[22] En effet, la construction d'un modèle de PD dans le contexte IRB repose généralement sur une chaîne de prétraitements des données et des variables qui entrent dans la régression logistique, utilisée in fine pour la différenciation des risques. Ainsi, par exemple les variables continues sont généralement discrétisées afin de limiter les effets des valeurs extrêmes et de faciliter l'interprétation de la grille de score. Les seuils de discrétisation sont parfois choisis de sorte à maximiser le pouvoir discriminant du modèle. De la même façon, les modalités des variables catégorielles sont regroupées pour aider le modèle à mieux identifier le risque de défaut dans le portefeuille. Enfin, la présélection des variables du modèle est basée sur une suite de tests permettant d'identifier les variables les plus liées à la cible (tests de Kruskall-Walis, tests du chi-deux, mesure du V de Cramer, etc.) et de retirer les variables les plus corrélées entre elles. La sélection finale parmi les variables candidates repose alors sur une méthode de sélection automatique (e.g. stepwise) et sur une analyse fine des résultats prédictifs du modèle sur un échantillon test. Or, dans la littérature académique, ces prétraitements ne sont généralement pas appliqués, et la comparaison de la régression logistique aux méthodes de ML se fait sur la base de données non transformées. Il y a donc un risque que les travaux académiques surestiment les gains en termes de précision des techniques de ML, dès lors que ces travaux n'appliquent pas ces différents prétraitements.

Une amélioration de la productivité. Un autre argument qui peut être avancé pour justifier l'utilisation du ML réside dans les gains de productivité que peuvent générer ces techniques (Grennepois *et al.*, 2018), notamment dans la phase de prétraitement des données et des variables que nous venons de mentionner. Par exemple, il est reconnu que les performances prédictives des forêts aléatoires sont généralement robustes à la non-imputation des valeurs manquantes, à la présence de fortes corrélations entre certaines variables explicatives, au non-regroupement des modalités des variables discrètes, et à la non-discrétisation des variables continues. De la même façon, on observe que les méthodes de forêts aléatoires, de Boosting ou de réseaux de neurones sont plus robustes aux déséquilibres entre les classes, que les arbres de décision, l'analyse discriminante linéaire ou la régression logistique (Chen *et al.*, 2022). Dès lors, l'utilisation de ces méthodes permet d'envisager de ne pas effectuer certains des prétraitements requis dans le cadre de la modélisation logistique. Si cet argu-

22. Une exception notable est celle de l'étude de Guégan et Hassani (2018) qui compare les résultats obtenus par des approches de ML aux estimations de PD produites par les modèles internes d'une banque, qui ont été validées par le régulateur.

ment de gains de productivité est sans doute plus pertinent pour la modélisation des scores d'octroi ou des scores de suivi non réglementaires, du fait du grand nombre de modèles de ce type en production et de leur fréquence de révision, il n'est pas pour autant à exclure dans le contexte IRB. Mais au-delà des gains de productivité, l'utilisation du ML dans ce contexte peut être vue comme un moyen de réduire les éventuels biais de modélisation puisque, au final, ces algorithmes laissent parler les données brutes. Par définition, le recours au ML est un moyen de réduire le risque de modèle inhérent aux choix humains de modélisation.

Autres utilisations en matière de risque. Le ML peut également être utilisé pour la modélisation des autres paramètres de risque, i.e., la LGD, le CCF ou l'EAD. Toutefois, dans la littérature académique, on trouve encore peu d'applications pour le CCF ou la modélisation directe de l'EAD, l'essentiel des techniques utilisées restant des spécifications paramétriques de type modèles à réponse fractionnaire à la Papke et Wooldridge (1996) ou de type modèle à loi Gamma ajustée à zéro (ZAGA), comme par exemple dans l'étude de Tong *et al.* (2016). On trouve en revanche de nombreuses applications du ML pour la modélisation de la LGD. L'intérêt du ML réside ici dans sa flexibilité et sa capacité à capter le caractère multimodal de la distribution des pertes en cas de défaut. Mais cet apport est beaucoup plus limité que dans le cas de la PD, puisque les $\mathbf{R^2}$ des modèles de LGD sont souvent très faibles, y compris pour les modèles de ML (Hurlin *et al.*, 2018). C'est pourquoi les LGD sont rarement calibrées à partir de modèles prédictifs. Les banques utilisent généralement de simples techniques de segmentation permettant de regrouper les crédits en groupes homogènes, et leur affectent un taux de recouvrement moyen observé sur longue période.

L'étude comparative de Loterman *et al.* (2012) évalue 24 techniques de modélisation de LGD à partir de 6 bases de pertes en cas de défauts observés pour des crédits à la consommation. La plupart sont des techniques statistiques de base (tableau de contingence) ou des modélisations paramétriques (régression linéaire, modèle de survie, régression à réponse fractionnée, régression Bêta, modèle Tobit, etc.). Les seuls algorithmes de ML considérés sont des arbres de régression, des forêts aléatoires, des méthodes de Gradient Boosting, des ANN, et des SVR. Quel que soit le modèle considéré, les auteurs montrent qu'il est difficile d'identifier les facteurs de la perte en cas de défaut. Tous les modèles de LGD présentent une faible performance prédictive : la performance moyenne de prédiction des modèles sur les différentes bases, mesurée en termes de R^2, varie de 4% à 43%. Toutefois, il apparaît que

les modèles de ML, et en particulier les SVM et les ANN, sont légèrement plus performants que les techniques paramétriques traditionnelles. Plusieurs travaux montrent également que des approches en deux étapes combinant des méthodes de ML (arbres de décision et LS-SVM) et des méthodes paramétriques de classification et de régression, permettent de mieux capturer la dimension multimodale de la distribution des pertes (Bellotti et Crook, 2012; Yao *et al.*, 2017).

4.2 Des gains potentiels en termes de capital réglementaire

Si l'on admet que l'usage du ML accroît la précision des estimations des paramètres de risque, encore faut-il que ces gains en précision se traduisent in fine par un allégement du capital réglementaire pour que les banques soient enclines à adopter ces techniques dans leurs modèles internes. Or, dans le cas de la modélisation de la PD, rien n'indique a priori que les gains prédictifs associés aux forêts aléatoires ou aux techniques de Gradient Boosting ne soient dus à une surestimation du risque de défaut par les approches paramétriques traditionnelles. Par ailleurs, si tant est que les méthodes de ML permettent des réductions de capital réglementaire, encore faut-il que ces modèles passent les tests de validation internes et les tests de validation des régulateurs. Sur le plan académique, on dispose encore de très peu de recul sur ces deux points. Seules quelques études ont été consacrées à l'évaluation économique des modèles de ML dans le contexte IRB, deux d'entre elles concernant les modèles de PD, et une portant sur les modèles de LGD.

Modélisation de la PD. L'étude la plus complète sur l'impact du ML sur les exigences en fonds propres réglementaires est celle de Fraisse et Laporte (2022). La démarche générale des auteurs consiste (i) à construire des pseudo modèles internes de PD fondés sur des algorithmes de ML usuels (forêt aléatoire, Boosting, régression Ridge, ANN), (ii) à vérifier que ces modèles seraient validés par un régulateur, puis (iii) à comparer les RWA déduits de ces modèles à ceux obtenus à partir des modèles usuels généralement mis en place dans les banques, ces derniers reposant essentiellement sur une combinaison de régression logistique et de jugements d'experts. Cette analyse mobilise des données granulaires qui portent sur 6 portefeuilles de crédits aux entreprises de grands groupes bancaires français, ces derniers représentant 80% des crédits distribués aux entreprises sur la période. [23] Conformément à la

23. Les données proviennent de la centrale des risques (registre des crédits) et de la base FIBEN (FIchier Bancaire des ENtreprises). Elles couvrent des crédits aux entreprises à l'exclusion des entreprises appartenant au secteur financier, au secteur immobilier, au secteur public et au secteur non lucratif. La base de données comporte des observations trimestrielles sur 229 657 entreprises de mars 2009 à juin 2015, soit au total

réglementation sur la validation des modèles internes, les auteurs considèrent des données trimestrielles historiques sur un horizon de 5 ans, de 2009 à 2014, pour construire les modèles internes, puis valident ces modèles sur les observations de l'année 2015.

Le grand avantage de l'étude de Fraisse et Laporte (2022) est qu'elle reproduit exactement la méthodologie IRB d'une banque. Dans une première étape, les auteurs construisent un score de risque continu exprimé en fonction d'une dizaine de facteurs de risque en utilisant soit une régression logistique, soit un algorithme de ML. Il est à noter que le ML n'intervient que dans cette phase de différenciation des risques. Dans une seconde étape, les entreprises sont regroupées en classes de risque homogène à partir du score de risque continu, transformé au préalable en une échelle de risque discrète, qui correspondrait à un système de notation. Enfin, dans une troisième étape, les auteurs affectent une probabilité de défaut à chaque classe de risque à partir d'un historique de défauts observés sur 1 an, puis calculent les exigences de fonds propres selon la formule Bâloise associée à l'approche IRB-F. [24]

La comparaison avec le modèle traditionnel utilisé par les banques est rendue possible par le fait que là encore, les auteurs cherchent à reproduire de façon fidèle toute la chaîne de prétraitements généralement mise en place par celles-ci avant d'appliquer la régression logistique comme modèle de différenciation des risques. Ces prétraitements incluent notamment l'élimination des variables présentant un trop grand nombre de valeurs manquantes, le contrôle des corrélations entre les variables, la discrétisation des variables continues, la création d'une classe associée aux valeurs manquantes, etc. La sélection des variables du modèle repose à la fois sur une approche quantitative et une approche qualitative (i.e., entretiens avec les agents de la Banque de France en charge de la notation des entreprises). [25]

La comparaison des modèles traditionnels et des modèles de ML se fait selon trois critères :

5 846 627 observations.

24. Dans le cadre d'une approche IRB Foundation ou IRB-F, la banque est autorisée à développer son propre modèle interne pour estimer la PD, mais elle est tenue d'utiliser la LGD prescrite par le régulateur et d'autres paramètres requis pour calculer les RWA. A l'inverse, dans une approche IRB avancée ou IRB-A, la banque peut utiliser ses propres modèles internes pour l'ensemble des paramètres de risque, i.e., PD, LGD et CCF.

25. L'analyse quantitative consiste à estimer un ensemble de régressions logistiques univariées sur l'indicateur de défaut incluant chaque variable discrétisée et des effets fixes de l'industrie, du statut judiciaire et de la taille. Les variables sont alors classées selon les mesures d'AUC obtenues dans ces régressions univariées et sont introduites dans le modèle final selon une procédure itérative en partant des variables associées aux mesures d'AUC les plus élevés. La sélection s'arrête lorsque l'ajout d'une nouvelle variable n'accroît pas de façon significative la performance prédictive du modèle. Cette méthode garantit la validité externe des conclusions de l'étude, qui sont ainsi obtenues dans un cadre méthodologique très proche des pratiques bancaires actuelles.

(i) les gains prédictifs mesurés en termes d'AUC et de F-score[26], (ii) la capacité à passer les tests de conformité utilisés par les régulateurs lors des missions sur site de validation des modèles internes et (iii) les évolutions induites du capital requis mesuré par le ratio des RWA sur le total des actifs de la banque. La capacité des modèles à passer les tests de conformité est mesurée notamment par un indicateur de différenciation des risques correspondant au nombre de fois où le z-test conduit à rejeter, pour chaque trimestre de la période de test, l'hypothèse nulle d'égalité des taux de défaut moyens entre deux classes de risque adjacentes. Plus cet indicateur est élevé, meilleure est l'échelle de notation puisqu'elle permet de mieux discriminer les risques.

En ce qui concerne les gains prédictifs, les résultats de l'étude confirment globalement les résultats de la littérature académique. Les forêts aléatoires permettent de mieux prévoir les défauts d'entreprise que le modèle paramétrique traditionnel, mais aussi que les autres techniques de ML (XGBoost, régression Ridge et ANN) qui donnent approximativement les mêmes AUC et F-scores que la régression logistique, une fois pris en compte les prétraitements. Ainsi, la moyenne des mesures AUC obtenues sur les bases de tests des 6 portefeuilles de crédits passe de 83% pour la régression logistique à 87% pour les forêts aléatoires, tandis que les écarts avec les autres techniques de ML restent inférieurs à 3%, conformément aux conclusions de Thomas (2000), Baesens *et al.* (2003) et Lessmann *et al.* (2015).

Au delà des gains prédictifs, les résultats de l'étude mettent en évidence des différences notables dans la capacité des modèles de ML à passer les tests réglementaires et à réduire les exigences de fonds propres. Tout d'abord, certains algorithmes de ML conduisent même à une augmentation des RWA, non seulement par rapport au modèle logistique, mais aussi par rapport à l'approche standardisée, ce qui remet en cause l'intérêt même des modèles internes pour la banque. Les forêts aléatoires qui améliorent la précision des estimations de PD, entraînent également de fortes baisses des RWA, mais uniquement pour un sous-échantillon de banques. Par ailleurs, les forêts aléatoires échouent généralement aux tests de validation en dehors de la période de calibration du modèle. Il est en de même pour le Gradient Boosting qui ne passent pas non plus les tests de validation. A l'inverse, tout en affichant une capacité similaire à celle du modèle traditionnel à réussir les tests de conformité, les ANN offrent la plus forte incitation pour les banques à appliquer des modèles de ML, car

26. Le F-score correspond à la moyenne harmonique de la précision et de la sensibilité (*recall*). La précision est la proportion des défauts correctement identifiés parmi l'ensemble des défauts prévus ; la sensibilité est la proportion des défauts correctement identifiés parmi l'ensemble des défauts observés.

ils conduisent dans certains cas à une réduction importante des exigences de capital. Ainsi, les ANN conduisent à une réduction des RWA comprise entre 2% et 27% par rapport aux modèles de régression logistique.

Une étude similaire à celle de Fraisse et Laporte (2022) a été menée sur un portefeuille de crédits à la consommation d'une banque espagnole par Alonso et Carbó (2020) avec des résultats légèrement différents. Les auteurs montrent tout d'abord que les forêts aléatoires et le XGBoost performent mieux que l'approche traditionnelle fondée sur une régression logistique tant en termes de différenciation des risques (mesurée par l'AUC) qu'en termes de calibration (mesurée par un score de Brière). [27] Les auteurs cherchent ensuite à évaluer les gains en termes de capital réglementaire permis par l'utilisation d'un modèle de différenciation des risques de type XGBoost en lieu et place d'une régression logistique pénalisée de type Lasso, qui sert de benchmark. Pour ce faire, ils reproduisent exactement la méthodologie IRB en construisant des classes de risques à partir des probabilités de défaut estimées par le modèle de ML, puis en calibrant les probabilités de défaut au sein de chaque classe sur la fréquence historique des défauts observée sur une base d'apprentissage. Partant d'une classification en 50 classes, les auteurs réduisent le nombre de classes et ajustent les seuils de sorte à obtenir une fréquence observée des défauts strictement croissante entre les classes de risques et la plus proche possible de la moyenne des probabilités estimées par le modèle. Ils obtiennent alors au final 6 classes de risque pour le modèle logistique-Lasso et 8 classes de risque pour le modèle XGBoost. Une fois les probabilités de défaut calibrées pour chaque classe de risque, les auteurs utilisent alors la formule réglementaire de Bâle pour les crédits à la consommation afin de calculer les exigences en fonds propres, en utilisant une LGD fixée à $0,45$ conformément à l'approche IRB-F. Leurs résultats montrent que les gains en capital réglementaire permis par l'utilisation du XGBoost atteignent en moyenne $12,4\%$, voire 17% dans des scénarios basés sur différentes valeurs des corrélations, ce qui est relativement important. Deux remarques doivent être faites ici. Premièrement, contrairement à Fraisse et Laporte (2022), les auteurs n'appliquent pas l'ensemble de la procédure de validation des classes de risque et notamment les tests d'homogénéité des risques au sein de chaque classe. [28] Deuxièmement, puisque les deux modèles ont été calibrés, la PD moyenne

27. Le score de Brière permet de comparer les probabilités estimées par le modèle au sein de chaque classe de risque homogène, à la fréquence des défauts observés sur la période de test au sein de cette classe.

28. Pour rappel, pour que les classes de risque soient approuvées par le régulateur, elles doivent répondre à deux critères critères, i.e., l'hétérogénéité des risques entre les classes et l'homogénéité des risques au sein de chaque classe. Ici seule l'hétérogénéité des risques est évaluée. Par ailleurs, Alonso et Carbó (2020) n'ont pas de dimension temporelle dans leurs données, ce qui expliquent pourquoi ils ne peuvent pas implémenter

globale de chacun d'entre eux ne diffère pas de manière significative. Dès lors, les gains en capital réalisés grâce à l'utilisation de la méthode XGBoost proviennent potentiellement soit d'une différence dans l'affectation des crédits au sein des classes de risque, soit de la différence du nombre de classes de risque obtenues à partir de la procédure de calibration. Les auteurs montrent que les deux effets jouent de façon importante dans la réduction du capital réglementaire requis.

On peut enfin citer ici l'analyse menée par FICO (Ould, 2022) à partir de leur technologie de ML interprétable (intégrée à leur outil *FICO Platform*) et de données provenant d'un modèle IRB d'une banque de la zone euro. Pour chacun des trois segments considérés (crédits hypothécaires, crédits à la consommation et autres prêts), les PD ont été estimées à partir d'un modèle de type Gradient Boosting avec exactement les mêmes variables d'entrée et les mêmes caractéristiques générées que pour les grilles de score initial. Les résultats de l'étude montrent une diminution de $20,8\%$ des RWA pour les crédits à la consommation, de $11,1\%$ pour les prêts et de $8,1\%$ pour les hypothèques. Toutefois, rien n'est dit dans l'étude sur la validation de ces modèles.

Modélisation de la LGD. Dans le cas de la modélisation de la LGD, il n'existe pas d'étude similaire permettant de déterminer quels pourraient être les gains économiques à l'usage des techniques de ML. Toutefois, Hurlin *et al.* (2018) évaluent différentes fonctions de pertes définies en termes de capital réglementaire pour les modèles de LGD. [29] Ces fonctions de perte permettent ainsi d'évaluer et de comparer la performance prédictive de différents modèles de LGD, y compris des modèles de ML, directement en termes de gains ou de pertes en capital requis. L'application empirique sur un portefeuille de crédits et de leasings automobiles permet de comparer différents modèles de ML (ANN, SVM, forêts aléatoires, etc.) et des modèle paramétriques (modèle de régression à réponse fractionnaire, etc.). Les résultats montrent que le classement des modèles obtenus sur la base des fonctions de perte en capital peut être très différent de ceux obtenus sur la base des critères statistiques usuels (MSE, MAE, etc.). Cela illustre, là encore, la divergence qui peut exister entre les gains en précision statistique et les gains économiques du ML.

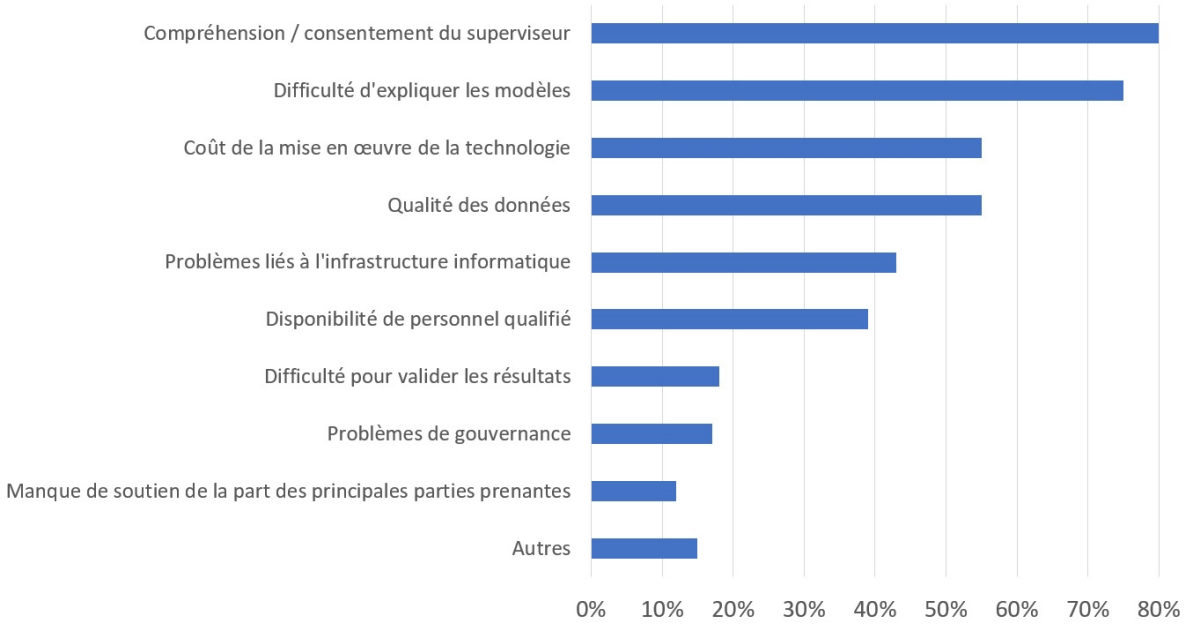
certains tests pour valider leur modèle.

29. Ces fonctions de perte sont fondées sur une erreur quadratique définie par rapport au capital réglementaire optimal, calculé à partir des vraies réalisations des paramètres de risque. Les fonctions de perte peuvent être symétriques ou asymétriques, de sorte à pénaliser toute sous-estimation du capital réglementaire requis.

5 Le défi de l'explicabilité / interprétabilité des modèles IRB

Le ML peut être un outil puissant pour la modélisation du risque de crédit, mais il soulève également de nombreux défis. Comme le montre la Figure 9, lorsque l'on interroge les banques à ce propos (IIF, 2019), les principaux défis mentionnés concernent (i) l'interprétabilité ou l'explicabilité des modèles, que ce soit dans une perspective de contrôle du risque de modèle en interne, ou vis-à-vis du régulateur, (ii) les coûts d'implémentation, ainsi que la problématique de la qualité et de la disponibilité des données, (iii) la gouvernance de ces modèles et les besoins accrus de formation du personnel, et (iv) les risques opérationnels liés notamment à la difficulté d'évaluer leur capacité de généralisation.

FIGURE 9 – Les principaux défis du ML dans la gestion du risque de crédit



(a) Source : IIF (2019)

Dans le contexte spécifique des modèles IRB, on retrouve les mêmes préoccupations qui découlent de l'analyse des exigences du CRR (EBA, 2021). De façon générale, ces préoccupations sont liées à la complexité et à la fiabilité des modèles de ML. Cependant, la principale préoccupation des banques porte sans conteste sur la question centrale de l'interprétabilité de ces modèles, et surtout sur les exigences que peut avoir le régulateur en la matière. C'est cette question de l'interprétabilité et de sa définition par le régulateur qui, sans aucun doute, conditionnera l'adoption du ML dans le contexte IRB à l'avenir.

5.1 Quelle explicabilité / interprétabilité pour les modèles IRB ?

La question de l'interprétabilité des modèles est le sujet central de préoccupation de toutes les applications à haut risque du ML et de l'IA (EC, 2021). Le problème tient au fait que certains modèles de ML peuvent être très complexes, ce qui rend difficile, voire impossible, l'interprétation de leurs prédictions et la compréhension des facteurs qui ont contribué à ces prédictions. Or, il est évident que dans le contexte IRB, il n'est pas possible de confier la détermination des fonds propres réglementaires d'une banque à un modèle de type boîte noire. Le superviseur, les services de validation interne et la gouvernance de la banque doivent pouvoir s'assurer que les paramètres utilisés pour le calcul des exigences de fonds propres reflètent fidèlement le niveau de risque pris par l'institution. Par conséquent, les modèles doivent être facilement interprétables et compréhensibles, à la fois par la direction de la banque afin qu'elle puisse prendre des décisions stratégiques en matière d'allocation de portefeuille de façon éclairée, et par le superviseur en charge de l'application des réglementations prudentielles. Or, les modèles de ML les plus performants pour la gestion du risque de crédit, e.g., les forêts aléatoires et les méthodes de Gradient Boosting, sont précisément non interprétables et sont typiquement assimilés à des boîtes noires, ce qui limite par conséquent leur usage dans le contexte IRB.

Dans le détail, on doit distinguer ici les notions d'explicabilité et d'interprétabilité (ACPR, 2020). L'interprétabilité fait référence à la capacité d'un modèle à être compris par un être humain, c'est-à-dire la capacité à comprendre comment le modèle fonctionne et comment il prend ses décisions. L'explicabilité, quant à elle, fait référence à la capacité d'un modèle à être expliqué, c'est-à-dire la capacité à fournir une explication compréhensible pour les décisions qu'il prend. En d'autres termes, l'interprétabilité se concentre sur la compréhension globale du modèle, tandis que l'explicabilité se concentre sur la compréhension locale des décisions individuelles prises par le modèle. Dans le contexte IRB, ces deux niveaux d'analyse ne s'adressent pas aux mêmes parties prenantes. Ainsi, selon Bracke *et al.* (2019), les modélisateurs et les valideurs de premier niveau doivent s'intéresser tant à l'intepretabilité qu'à l'explicabilité afin de comprendre notamment quels sont les facteurs essentiels qui déterminent les paramètres de risque, mais aussi ce qui détermine les prévisions de PD et de LGD obtenues pour chaque contrat. A l'inverse, les superviseurs et les valideurs de second niveau doivent s'intéresser davantage à l'interprétabilité, i.e., au fonctionnement général du modèle, et surtout à sa capacité à correctement mesurer les risques dans un nouvel

environnement économique.

5.2 Les méthodes d'interprétabilité et d'explicabilité ex-post

Lorsque les modèles issus des algorithmes de ML ne sont pas interprétables a priori, la démarche usuelle consiste à mettre en place des méthodes qui permettent de les rendre interprétables ex-post (Molnar, 2019; Molnar *et al.*, 2022). Ces méthodes sont dites « agnostiques aux modèles » dans le sens où elles peuvent être utilisées pour interpréter le fonctionnement et les résultats de n'importe quel modèle de ML. L'idée générale de ces méthodes consiste à faire de l'ingénierie inversée à partir des seuls inputs (prédicteurs) et outputs (prévisions) afin de révéler la règle de décision de l'algorithme. On distingue ici les méthodes dites globales, destinées à répondre à la question de l'interprétabilité, des méthodes dites locales, destinées à répondre à la question de l'explicabilité.

Méthodes globales. Parmi les méthodes globales, l'approche la plus simple consiste à construire un modèle de substitution qui soit nativement interprétable (e.g., une régression logistique, un arbre de décision) et qui permettent de relier les prévisions du modèle boîte noire aux variables utilisées par celui-ci. L'idée est alors de fournir une approximation de la règle de décision non interprétable d'un algorithme de ML par un modèle de substitution plus simple et interprétable, considéré comme une approximation globale de la vraie règle de décision. Toutefois, la qualité de l'interprétation dépend alors fondamentalement de la qualité d'ajustement du modèle de substitution.

D'autres approches fondées sur des analyses graphiques permettent de ne pas avoir à construire un modèle de substitution. Une première approche est celle des graphiques de dépendances partielles ou PDP (Friedman, 2001). Le principe de construction des PDP consiste à balayer sur un ensemble de valeurs possibles pour une variable explicative d'intérêt, d'appliquer le modèle prédictif de ML en utilisant pour tous les individus de la base ces valeurs successives (tout en laissant inchangées les valeurs des autres variables explicatives), et de calculer, pour chaque valeur testée, la moyenne des scores obtenus. Le graphique de PDP permet alors de représenter l'effet marginal d'une caractéristique sur la prédiction du modèle de ML. On peut ainsi vérifier si la relation entre la cible et un prédicteur est linéaire, monotone ou plus complexe. Cependant, la principale limite des PDP est que cette méthode suppose que la caractéristique pour laquelle la dépendance partielle est calculée, soit indépendante des autres caractéristiques. Plusieurs extensions récentes des PDP per-

mettent toutefois de lever cette hypothèse d'indépendance entre les variables du modèle, comme par exemple les graphiques des effets locaux cumulés ou ALE (Apley et Zhu, 2020).

Méthodes locales. Les méthodes locales fournissent une explication à une décision prise par le modèle boîte noire au niveau individuel, e.g., au niveau d'un emprunteur. Ces méthodes permettent par exemple de déterminer pourquoi la PD d'une entreprise est estimée à 4% par le modèle. Il existe de très nombreuses méthodes d'explication locale, mais les deux plus utilisées par les banques sont les méthodes LIME (Ribeiro *et al.*, 2016) et SHAP (Lundberg et Lee, 2017).

L'intuition de la méthode LIME (*Local interpretable model-agnostic explanations*) est de créer des modèles de prédiction au niveau local pour chaque individu, qui soient plus simples et plus facilement interprétables que le modèle original. Pour un individu de référence, on génère aléatoirement des variables explicatives d'individus proches de l'individu de référence (perturbations locales), et l'on applique alors l'algorithme initial sur ces données simulées. A partir des caractéristiques simulées et des prévisions qui en découlent (pondérées par rapport à leur distance à l'individu de référence), on ajuste alors un modèle interprétable. Il s'agit typiquement d'un arbre de décision ou d'une régression pénalisée de type Lasso. Grâce à ce modèle local, on peut alors mesurer de façon approchée l'influence de chacun des prédicteurs sur la décision finale du modèle initial pour l'individu de référence. Ainsi, par exemple, la banque est en mesure d'expliquer pourquoi tel emprunteur a été classé en défaut et quelles sont les caractéristiques qui ont conduit à ce choix.

La seconde méthode d'interprétation locale largement utilisée par les banques est la méthode SHAP (*SHapley Additive exPlanations*). Cette méthode est basée sur la notion de valeur de Shapley (Shapley, 1953), tirée de la théorie des jeux et qui permet de répartir équitablement un paiement entre des joueurs. Dans le cas du ML interprétable, la prévision du modèle obtenue pour un individu est assimilée au paiement que l'on cherche à répartir entre les différentes variables explicatives du modèle, ces dernières étant assimilées aux joueurs. Le principe des valeurs de Shapley repose sur l'idée que cette répartition doit être déterminée par l'excédent de gain généré par chaque joueur. Dans le contexte du ML, ce gain sera défini comme la différence entre la prévision pour l'individu de référence et la moyenne des prévisions pour les autres individus. Ainsi, la contribution de Shapley d'une variable correspond à sa contribution marginale et pondérée à la prévision, calculée à partir de toutes les combinaisons possibles dans lesquelles cette variable aurait pu être ajoutée à

l'ensemble des autres variables explicatives du modèle. Cette décomposition a pour avantage de reposer sur un fondement théorique et de vérifier plusieurs propriétés intéressantes. Toutefois, le calcul des valeurs de Shapley devient très rapidement impossible dès lors que le nombre de variables du modèle excède une dizaine en raison du nombre trop important de combinaisons possibles de variables à considérer. C'est dans ce contexte que la méthode SHAP de Lundberg et Lee (2017) a été développée. Cette méthode permet d'obtenir des approximations des valeurs de Shapley dans des problèmes de grande dimension. Mais en dépit de ces méthodes de calcul efficaces, le calcul des valeurs de Shapley peut rester coûteux en temps de calcul en fonction du nombre de prédicteurs, d'observations, et de la complexité du modèle considéré.

Plusieurs extensions de ces méthodes locales ont été proposées, notamment dans le cadre des modèles de risque de crédit. Ainsi, Bracke *et al.* (2019) proposent une approche englobante, dite QII (*Quantitative Input Influence*), basée sur les valeurs de Shapley, mais à partir de laquelle les auteurs tirent une interprétation globale de l'influence des variables. Cette représentation graphique, similaire dans l'esprit au PDP, permet de prendre en compte à la fois les non-linéarités et les interactions entre les variables.

Les limites des méthodes d'interprétabilité ex-post. La démarche qui consiste à utiliser un modèle de ML de type boîte noire comme modèle de différenciation des risques, puis à mettre en œuvre ex-post une méthode d'interprétabilité / explicabilité locale ou globale est aujourd'hui très discutée en raison des limites inhérentes à ces méthodes. Ainsi, Rudin (2019) préconise l'arrêt total de l'utilisation des modèles boîte noire pour les applications à haut risque, y compris lorsque ces modèles sont accompagnés de méthode de ML interprétable, et cela pour plusieurs raisons. La première est que les méthodes d'interprétabilité ex-post fournissent des explications qui ne sont pas fidèles à ce que le modèle original calcule. L'argument de bon sens étant que si l'explication était parfaitement fidèle au modèle original, l'explication serait identique à ce dernier, et donc le modèle original serait par définition lui-même interprétable. Il y a donc un risque que les méthodes d'interprétabilité ex-post fournissent une représentation inexacte du modèle original dans certaines parties de l'espace des caractéristiques. La seconde raison est que les explications n'ont parfois pas de sens ou ne fournissent pas suffisamment de détails pour comprendre comment fonctionne le modèle. Même dans le cas où le modèle de ML original produit une prévision correcte et la méthode d'interprétabilité fournit une bonne approximation de cette prévision, il est

possible que cette explication omette tellement d'informations qu'elle n'ait aucun sens. La troisième raison c'est que les explications fournies par différentes méthodes d'interprétabilité pour un même modèle, sont parfois incohérentes les unes par rapport aux autres (Krishna *et al.*, 2022). Or, lorsque plusieurs méthodes donnent des explications différentes, il n'est pas évident de savoir à laquelle se fier. Ainsi, pour Rudin, les méthodes d'interprétabilité ne créent pas nécessairement la confiance vis-à-vis des modèles de ML, elles permettent simplement aux utilisateurs de décider s'ils veulent faire confiance à ces modèles. En d'autres termes, ces méthodes permettent de prendre une décision de confiance, sans créer la confiance elle-même.

Un article récent de Chen *et al.* (2022) met également en évidence l'instabilité des interprétations fournies par les LIME et les SHAP dans le cas d'un modèle de scoring bancaire non interprétable (XGBoost) appliqué à un portefeuille de crédit présentant très peu de défauts. En ML, le problème des données déséquilibrées est bien connu puisqu'il influe à la fois sur la capacité d'apprentissage des modèles (la prévision de la variable dépendante tend à être biaisée en faveur de la classe majoritaire) et rend caduque la validité de certains critères d'évaluation (e.g., le pourcentage de classifications correctes). [30] Or, c'est une configuration que l'on retrouve fréquemment dans le cadre des modélisations de PD, notamment dans le contexte IRB (puisque les emprunteurs ont déjà été sélectionnés lors de la phase d'octroi), du fait que les taux de défaut observés sont souvent très faibles. Plusieurs méthodes basées sur des techniques de ré-échantillonage (SMOTE, ROSE, etc.), ont été proposées pour améliorer les performances de classification des modèles de ML en présence de données déséquilibrées. Mais l'étude de Chen *et al.* (2022) montre que le déséquilibre des données influe également sur les méthodes d'interprétabilité. Cela se traduit notamment par une très grande variabilité des contributions des variables mesurées par les LIME et les SHAP lorsque les taux de défaillance sont très faibles, i.e., de l'ordre de 1% à $2,5\%$. Ainsi, la présence de données déséquilibrées peut nuire à la fiabilité des interprétations économiques des modèles internes de ML.

30. Dans le cas d'un apprentissage supervisé, un échantillon est dit déséquilibré lorsque le nombre d'observations dans la classe minoritaire est nettement inférieur à celui de la classe majoritaire.

5.3 Existe-il un arbitrage interprétabilité / performance ?

La question sous-jacente posée par l'utilisation des modèles de ML de type boîte noire dans le contexte IRB est celle de l'arbitrage entre l'interprétabilité et la performance des modèles de différenciation des risques. L'idée communément admise, serait que les modèles complexes de ML permettraient des gains de précision dans les estimations des paramètres de risque. Dès lors, en raisons des ces gains, il conviendrait d'accepter de mettre en place des techniques permettant de rendre leurs prédictions interprétables ex-post, tout en étant conscient des limites des explications fournies par ces méthodes.

Mais cette vision d'un arbitrage entre performance et interprétabilité qu'il conviendrait d'accepter avec fatalisme, est aujourd'hui pour partie remise en cause dans la littérature et dans la pratique de certaines banques. Ainsi, pour Semenova *et al.* (2022), il est faux de croire qu'il existe nécessairement un compromis entre la précision et l'interprétabilité. Leur raisonnement théorique est basé sur la notion d'ensemble de Rashomon. Pour un jeu de données, l'ensemble de Rashomon est défini comme l'ensemble des modèles prédictifs qui fournissent une précision très proche de celle fournie par le modèle optimal, que l'on admet être de type boîte noire. Les données étant finies, elles admettent de nombreux modèles proches de l'optimum qui prédisent différemment les uns des autres, i.e., un grand ensemble de Rashomon. [31] Si l'ensemble de Rashomon contient un ensemble suffisamment grand de modèles produisant des prédictions proches, il contient probablement des modèles interprétables. Si cette théorie est vérifiée, il existe donc des modèles qui sont à la fois précis et interprétables. [32]

Si l'on admet qu'il puisse ne pas y avoir d'arbitrage entre performance et interprétabilité en théorie, reste encore à trouver en pratique des modèles de scoring bancaire qui soient à la fois performants et interprétables nativement. Plusieurs exemples montrent que ceci est réalisable en pratique. Dumitrescu *et al.* (2022) proposent ainsi une approche hybride, dite PLTR, qui repose sur un modèle de régression logistique incluant des prédicteurs extraits d'arbres de décision. Ces prédicteurs correspondent à des règles binaires (feuilles) produites

31. Dans la pratique, on observe en effet souvent que de nombreux algorithmes de ML de classification supervisée différents, donnent des résultats relativement proches en termes d'AUC, etc. sur le même ensemble de données, bien qu'ils aient des spécifications très différentes.

32. Semenova *et al.* (2022) introduisent un ratio de Rashomon qui est une nouvelle mesure de la simplicité pour un problème d'apprentissage, en fonction d'une classe de fonctions et d'un ensemble de données. Le ratio de Rashomon est le rapport entre le volume de l'ensemble des modèles exacts et le volume de l'espace des hypothèses. Cette mesure est différente des mesures de complexité standard. L'étude du ratio de Rashomon permet de vérifier s'il existe un modèle plus simple pour un problème avant de le trouver.

par des arbres de décision de très faible profondeur construits avec les variables prédictives originales. Pour traiter un nombre éventuellement élevé de règles d'arbres de décision, le modèle de régression logistique est estimé par le biais d'une régression pénalisée de type Lasso adaptatif permettant de sélectionner automatiquement les prédicteurs les plus importants. Cette phase de prétraitement des variables permet ainsi d'améliorer les performances prédictives d'une simple régression logistique en captant notamment les potentiels effets non linéaires, e.g., les effets de seuils, et les interactions entre les variables explicatives, tout en conservant son caractère parfaitement interprétable. L'application sur un portefeuille de crédits montre que le modèle PLTR permet de prévoir les défauts aussi bien qu'une forêt aléatoire. Dans le même esprit, Flachaire *et al.* (2022) proposent l'approche GAM(L)A fondée sur des modèles additifs généralisés (*Generalized Additive Model*). Cette approche repose sur la création de fonctions paramétriques et non paramétriques des variables initiales permettant de capter avec précision les linéarités et les non-linéarités existant entre les variables dépendantes et explicatives, ainsi qu'une procédure de sélection des variables (de type Lasso ou Autometrics) pour contrôler les problèmes de surajustement. Les résultats obtenus pour un portefeuille de crédits à la consommation, montrent également que ces modèles nativement interprétables offrent des performances prédictives similaires aux modèles de forêts aléatoires ou XGBoost largement utilisés par les banques. On peut également citer les modèles récents de type Explainable Boosting Machine (EBM), qui sont des modèles similaires aux modèles GAM, mais qui contrairement à ceux-ci permettent automatiquement de détecter et d'inclure des termes d'interaction par paire, ou les modèle GAM avec des interactions structurées (GAMI-Net). [33]

Afin d'illustrer à quel point la croyance dans le mythe du compromis entre précision et interprétabilité est ancrée dans l'esprit des professionnels du scoring de crédit, Semenova *et al.* (2022) évoquent l'exemple du challenge sur le ML interprétable, qui fut organisé en 2018 par FICO. Dans ce challenge, les participants avaient pour objectif de créer un modèle de ML de type boîte noire pour prédire les défauts sur un portefeuille de crédits, puis d'expliquer les prévisions du modèle en mobilisant des techniques de ML interprétable. Au final, un modèle généralisé additif à deux niveaux globalement interprétable (Chen *et al.*, 2018) qui a remporté la compétition. La construction du challenge FICO illustre à quel point les organisateurs ne s'attendaient pas à ce qu'un modèle interprétable puisse produire

33. La plupart de ces modèles interprétables sont disponibles dans le package Python PiML (*Python Interpretable Machine Learning*).

des prévisions de même qualité que les modèles boîte noire, puisqu'ils n'avaient même pas demandé aux participants d'essayer de construire un tel modèle. Toutefois, cette nouvelle démarche d'un ML nativement interprétable est sans doute appelée à se développer au sein des banques dans les années à venir, comme le montre par exemple les recherches menées actuellement par Wells Fargo sur la question (Sudjianto et Zhang, 2021). On peut citer ici également les travaux de Sudjianto *et al.* (2021) portant sur une nouvelle méthode d'ensemble basée sur une multitude de réseaux neuronaux étroits à couche cachée unique, appelée LIFE (*Linear Iterative Feature Embedding*), permettant d'atteindre simultanément une grande précision dans l'estimation des paramètres de risque, une interprétation aisée et un temps de calcul raisonnable.

Encadré 5 : Les autres défis du ML dans le contexte IRB

En dehors de la problématique de l'interprétabilité, l'association bancaire espagnole (AEB) évoque d'autres défis liés au ML, à savoir la gouvernance des données, notamment pour la conservation des historiques, la mise en place de nouvelles capacités informatiques, et le développement des compétences humaines. [a] En ce qui concerne l'AI act, l'association considère que les modèles IRB ne doivent pas être considérés comme faisant partie des applications à haut risque. L'argument est que l'utilisation de l'IA dans les modèles IRB n'affecte pas le processus décisionnel du prêteur et n'a donc pas d'incidence sur les droits des personnes. De façon générale, l'AEB souhaite ainsi que les applications de l'IA portant sur toutes les phases qui suivent le décaissement initial du prêt (suivi des risques, provisionnement, etc.) soient clairement exclues du champ d'application de la loi sur l'IA. Sur un tout autre sujet, la fédération bancaire française (FBF) mentionne que l'un des défis essentiels du ML est celui de la reproductibilité des résultats obtenus par ces modèles, notamment entre la plateforme de développement et la plateforme de production. Enfin, l'association allemande (GBIC) met en avant la formation et la nécessité de développer une expertise dans les méthodes de ML afin de s'assurer que les outils utilisés pour développer, maintenir et contrôler les modèles sont adéquats et bien compris.

a. Les trois principales questions qui portaient sur les défis du ML en dehors de l'interprétabilité, étaient libellées de la façon suivante : « Quels sont les défis spécifiques que vous voyez concernant le développement, la maintenance et le contrôle des modèles de ML dans le contexte de l'IRB [.] ? », « Pensez-vous que l'utilisation de la ML dans le contexte des modèles de l'IRB pose des problèmes découlant de l'AI act ? », « Voyez-vous un autre défi ou une autre question à discuter concernant l'utilisation de modèles de ML dans le contexte de l'IRB ? ».

6 Les autres défis posés par le ML dans le contexte IRB

Dans leurs réponses à l'EBA (cf. Encadré 5), les associations professionnelles bancaires européennes évoquent plusieurs défis liés à la mise en place du ML dans le contexte IRB : la question des capacités informatiques et des logiciels, la formation des personnels à ces nouvelles méthodes, la reproductibilité des résultats, le suivi de la performance des modèles dans le temps, etc. La plupart de ces défis relèvent d'une problématique générale de gouvernance du ML. Toutefois, à ces enjeux, nous ajouterons deux défis particuliers, à savoir, d'une part celui de la maîtrise des risques opérationnels, et notamment du risque de modèle, et d'autre part celui de l'équité algorithmique et du contrôle des biais systématiques, en lien avec la future réglementation sur l'IA de la Commission Européenne (EC, 2021).

6.1 Le challenge de la gouvernance du ML dans le contexte IRB

L'utilisation de plus en plus courante par les banques de l'IA, et du ML en particulier, a conduit les régulateurs à proposer des règles de bonnes pratiques concernant la gouvernance de ces nouveaux outils (ACPR, 2020; EBA, 2020; EC, 2020). Dans son rapport sur le Big Data et les approches analytiques avancées (EBA, 2020), l'EBA identifie quatre piliers essentiels pour le développement, la mise en œuvre et l'adoption de la l'IA dans le secteur bancaire. Le premier pilier porte sur la gestion des données et vise à garantir l'utilisation de données de qualité (exactitude et intégrité, actualité, cohérence et exhaustivité), tout en tenant compte des contraintes de protection des données personnelles, et notamment la conformité vis-à-vis du Règlement Général sur la Protection des Données (RGPD). Le second pilier porte sur l'infrastructure technologique. Quel que soit le contexte d'application, il n'y a pas de bonne gouvernance du ML sans une bonne gouvernance des bases de données, des infrastructures de calcul, des logiciels, etc. Dans le contexte réglementaire, on peut par exemple s'interroger sur l'utilisation de logiciels libres tels que Python ou R, et de certains modules et packages de méthodes avancées de ML qui pourraient être utilisés en production sans avoir été testés et validés au préalable, et sur les conséquences en cas d'erreur de programmation de ces packages. De même, la question de la sécurité peut devenir centrale lorsque les solutions de ML sont déployées sur des serveurs de calculs appartenant à des prestataires externes. Le troisième pilier porte sur les structures de gouvernance interne et les mesures organisationnelles destinées à encadrer le développement de ces applications d'IA, ainsi que sur la formation et le développement de compétences et de connaissances

suffisantes. La question de la formation est d'autant plus importante que si les premières méthodes de ML datent de la fin des années 50 et que la plupart des méthodes actuellement utilisées dans les banques ont été élaborées entre le milieu des années 80 et le début des années 2000, la grande majorité des formations supérieures en économie, finance ou économétrie, n'ont intégré ces outils dans leurs programmes qu'à partir de la fin des années 2010.[34] Dès lors, l'utilisation du ML doit s'accompagner d'un effort conséquent de formation à destination des personnes en charge de la construction des modèles et de leur validation, mais aussi à destination des personnels d'encadrement, des organes de direction, et des équipes en charge de la supervision. Le dernier pilier concerne la méthodologie qui doit être mise en place pour faciliter le développement et l'adoption des solutions d'IA au sein des banques. Le développement d'un projet de ML suit un cycle de vie avec des étapes spécifiques, e.g., la préparation des données, la modélisation, la validation, le suivi au cours du temps, etc. La gouvernance mise en place doit ainsi avoir pour objectif de mettre en œuvre une IA de confiance, intégrant notamment les dimensions d'éthique (cf. section 6.3), d'explicabilité, d'interprétabilité, de traçabilité et d'auditabilité.

Ces principes généraux se retrouvent dans la gouvernance du ML qui doit s'opérer dans le contexte spécifique des modèles IRB. Nous ne reviendrons pas ici sur les exigences des CRR/CRD relatives aux modèles internes quels qu'ils soient, mais insisterons davantage sur la spécificité des modèles issus d'algorithmes de ML. Dans son document de réflexion, l'EBA part du postulat que les modèles de ML peuvent apporter une valeur ajoutée dans le contexte IRB, à condition qu'ils soient assortis d'un suivi, d'une validation et d'une explicabilité acceptables. L'EBA décline ces principes généraux en un ensemble de recommandations plus précises qui peuvent être résumées de la façon suivante :

1. **Formation.** Lorsqu'une institution met en place des modèles de ML à des fins de fonds propres réglementaires, toutes les parties prenantes concernées (équipes de développement, équipes de validation et organes de direction) doivent avoir un niveau de connaissance approprié du fonctionnement du modèle.

2. **Interprétation économique.** Les parties prenantes doivent être en mesure d'évaluer la pertinence et l'adéquation des facteurs de risque utilisés, ainsi que la solidité

34. Par exemple, l'algorithme CART pour les arbres de classification a été publié en 1984 (Breiman *et al.*, 1984), les méthodes de bagging en 1996 Breiman (1996), les forêts aléatoires en 2001 (Breiman, 2001). La seule exception parmi les méthodes les plus utilisées par les banques est la méthode XGBoost qui est plus récente et date de la seconde moitié des années 2010 (Chen et Guestrin, 2016), mais la technique du Boosting trouve sa source dans une première publication de 1997 (Freund et Schapire, 1997).

du raisonnement économique sous-jacent dans le modèle global. Il s'agit ici d'évaluer les hypothèses de modélisation et de déterminer si les facteurs de risque sélectionnés contribuent à l'évaluation des risques conformément à leur signification économique. Implicitement, cela revient à préconiser l'utilisation de méthodes d'interprétabilité ou d'explicabilité ex-post, e.g., importance des caractéristiques, PDP, SHAP, LIME.

3. **Arbitrage performance / explicabilité.** L'EBA part du postulat qu'il existe un arbitrage entre la complexité des modèles et leurs performances prédictives, et recommande de trouver un compromis entre ces deux impératifs.[35] On peut noter que les méthodes de ML nativement interprétables constituent ici un compromis particulièrement attractif.

4. **Validation.** Pour les modèles ML complexes, dont l'explicabilité est limitée ou pour les modèles fréquemment mis à jour, une validation fiable est particulièrement importante et peut nécessiter une analyse en profondeur et/ou à une fréquence accrue. La validation doit se concentrer en particulier sur les questions de surajustement et de fixation des hyperparamètres, sur l'analyse de la stabilité du modèle, sur les enjeux de représentativité et de qualité des données, et sur la qualité des documentations qui accompagnent ces modèles.

5. **Mises à jour.** Il est recommandé de ne pas mettre à jour trop fréquemment les modèles de ML sans raison économique, e.g. rupture dans les conditions économiques ou dans les processus de l'établissement. L'objectif est ici de limiter le risque de modèle et les phénomènes dynamiques de surajustement.

6.2 Les risques opérationnels associés au ML

Une source importante de risque opérationnel dans la mise en œuvre du ML, est le risque de surajustement (*overfitting*) du modèle. Typiquement, un modèle ayant des hyperparamètres mal configurés peut être sujet à un surapprentissage, défini comme une situation dans laquelle le modèle surajuste les données d'entraînement (i.e., le biais des prévisions est faible) et ne peut pas généraliser correctement ces prévisions à de nouvelles données (la variance des prévisions est élevée). Le surajustement se traduit par une plus grande erreur de prévision sur l'échantillon de test, que sur l'échantillon d'apprentissage. Le contrôle du

35. L'EBA recommande par exemple le fait de ne pas inclure trop de facteurs explicatifs si cela n'apporte pas d'information prédictive significative, de ne pas utiliser de données non structurées s'il existe des données plus conventionnelles offrant des capacités prédictives similaires, de ne pas choisir des modélisations trop complexes si des approches plus simples donnant des résultats similaires sont disponibles.

risque de surajustement est essentiel dans le contexte IRB car le but des modèles internes, qui sont entraînés sur une base de défauts passés (base d'apprentissage), est précisément de fournir des prévisions fiables des paramètres de risque pour des crédits sains n'ayant pas encore connu le défaut, actuellement dans le portefeuille de la banque (base de test). Dans le contexte IRB, le risque de surajustement est ainsi directement lié au risque de sous estimation des exigences en fonds propres.

Le risque de surajustement peut être contrôlé par le choix des hyperparamètres, i.e., des paramètres de configuration qui déterminent la forme générale du modèle de ML. Les hyperparamètres ne doivent pas être confondus avec les paramètres du modèle, ces derniers étant déterminés par l'algorithme de sorte à permettre au modèle de s'ajuster aux données. A l'inverse, les hyperparamètres ne sont pas déterminés par l'algorithme d'apprentissage lui-même, mais sont définis avant l'entraînement du modèle. Par exemple, dans le cas d'un arbre de décision, les hyperparamètres que l'on fixe sont généralement la profondeur de l'arbre et le nombre minimum d'observations par feuille. Une fois ces hyperparamètres fixés et la forme générale de l'arbre donnée, l'algorithme détermine les variables retenues et estime les paramètres de ce modèle, i.e., les seuils qui déterminent les règles de décision à chaque noeud. Tout changement de la valeur des hyperparamètres peut conduire à modifier radicalement la forme fonctionnelle d'un modèle de ML, sa complexité, son degré d'interprétabilité, ses règles de décision, et in fine ses prévisions. L'EBA prend comme exemple, le cas simple d'un arbre de régression. [36] Si la profondeur de l'arbre est fixée à deux, on obtient un modèle très simple qui peut prédire au maximum quatre valeurs différentes, en exploitant les informations de trois variables différentes au maximum. Mais si la profondeur est fixée à dix, la complexité du modèle devient significative, car le nombre de prédictions différentes possibles grimpe à 1024 et le modèle peut mobiliser jusqu'à 1023 variables différentes. Cet exemple illustre à quel point de petits changements dans la valeur des hyperparamètres, ou dans la manière de les déterminer, peuvent engendrer de forts changements de la structure d'un modèle de ML et de ses capacités prédictives. Ainsi, la détermination des hyperparamètres permet de contrôler la complexité du modèle de ML, cette complexité affectant à son tour l'arbitrage entre le biais et la variance des prévisions : en général, plus un modèle est complexe, plus son biais sera faible et sa variance élevée, impliquant un risque de surajustement.

36. Pour des modèles de ML plus sophistiqués, l'influence des hyperparamètres est tout aussi importante, mais elle est plus difficile à représenter de façon simple.

Il existe différentes méthodes permettant de fixer les hyperparamètres optimaux de façon à contrôler cet arbitrage entre biais et variance, et de limiter le risque de surajustement, la plus connue étant celle de la validation croisée. Il existe plusieurs variantes de cette méthode, e.g., *hold-out*, *K-fold* ou *leave-one-out*.[37] Mais de façon générale, ces méthodes consistent à séparer les données en un échantillon d'apprentissage sur lequel est entraîné le modèle, et un échantillon de validation sur lequel on évalue ses performances prédictives. On cherche alors à déterminer par optimisation (i.e. optimisation sur une grille de valeurs, optimisation aléatoire, ou optimisation bayésienne) la valeur optimale des hyperparamètres, de sorte à obtenir la meilleure performance du modèle sur l'échantillon de validation (Bergstra et Bengio, 2012). Une fois que les hyperparamètres optimaux ont été déterminés de façon à limiter le risque de surajustement, le modèle est entraîné sur l'ensemble des observations des bases d'entraînement et de validation. La phase finale d'évaluation du modèle est réalisée à partir d'un échantillon de test, différent des échantillons d'apprentissage et de validation.

Quelle que soit la méthode retenue, la fixation des hyperparamètres repose toujours sur un ensemble de choix arbitraires laissés au jugement du modélisateur. Ces choix peuvent être une source de risque opérationnel et doivent faire l'objet d'une validation stricte dans le contexte IRB. Le premier de ces choix est tout simplement celui des hyperparamètres qui seront obtenus (« *fine tuned* ») par une méthode de validation croisée, et de ceux qui, au contraire, seront laissés aux valeurs par défaut fixées dans les packages Python ou R. Même si les développeurs expérimentés ont souvent une intuition sur les hyperparamètres appropriés pour un problème donné, il convient de s'assurer de la validité de ces choix. Le deuxième choix arbitraire est souvent celui des grilles de valeurs ou des valeurs limites entre lesquelles s'effectuera l'optimisation aléatoire des hyperparamètres dans la procédure de validation croisée. Enfin, il convient de noter que la taille et la qualité des données peuvent affecter la détermination des hyperparamètres : plus les données sont nombreuses et diversifiées, plus il est probable que les hyperparamètres trouvés soient robustes et généralisables. En revanche,

37. La méthode *hold-out* est la plus simple, et consiste à diviser l'ensemble de données en un ensemble de données d'apprentissage et un ensemble de données de test. La méthode *K-fold* consiste à diviser l'ensemble de données en K parties égales. Le modèle est entraîné K fois, chaque fois en utilisant une des parties comme ensemble de données de validation et les $K-1$ parties restantes comme ensemble de données d'apprentissage. Les performances du modèle sont ensuite calculées en moyenne sur les K évaluations. Cette méthode est plus fiable que la méthode *hold-out*, car elle utilise toutes les données disponibles pour l'apprentissage et l'évaluation du modèle. Cependant, elle est plus coûteuse en temps de calcul. La méthode *leave-one-out* est une méthode *K-fold* particulière où K est égal au nombre d'observations dans l'ensemble de données n. Elle consiste à entraîner le modèle n fois en laissant une seule observation comme échantillon de validation à chaque fois, et à calculer les performances du modèle en moyenne sur les n évaluations.

si les données sont limitées ou biaisées, le choix des hyperparamètres peut conduire à des modèles mal spécifiés. Par conséquent, du fait de l'importance critique de la fixation des hyperparamètres et de la prédominance du jugement humain dans cette étape, l'EBA insiste sur la nécessité d'accorder une grande importance à la validation de ces choix. Les experts en charge de la validation des modèles internes doivent vérifier la logique qui sous-tend le choix des hyperparamètres. Le problème est que cette vérification peut s'avérer particulièrement difficile, notamment pour les modèles complexes, car elle suppose une connaissance approfondie de la méthodologie, qui seule permet d'appréhender toutes les implications des hyperparamètres. Cette validation est en outre essentielle pour limiter tout risque d'arbitrage réglementaire, qui consisterait, par exemple, à ajuster les hyperparamètres selon des approches de validation croisée afin de réduire le niveau des paramètres de risque et, in fine des exigences en fonds propres.

Dans la littérature académique, plusieurs exemples prouvent l'existence de ces risques opérationnels dans le cadre de la gestion du risque de crédit. L'étude de Fraisse et Laporte (2022) montre ainsi que les classes de risque homogènes, établies à partir des scores de certains modèles de ML, échouent aux tests de conformité en raison des comportements de surajustement de ces modèles. C'est notamment le cas pour les forêts aléatoires, qui affichent le niveau le plus élevé de performance prédictive sur les données d'apprentissage. Alors même que dans l'étude, les hyperparamètres de ces modèles ont été fixés selon une approche de validation croisée de type *K-fold* visant à limiter les risques de surajustement, les auteurs observent une baisse de près de 10 points de la moyenne des AUC obtenus sur les 6 bases d'apprentissage et les 6 bases de test, et une baisse de 0.44 points du F-score. [38] Cette instabilité explique l'absence de robustesse du score de risque produit par le modèle.

La fixation des hyperparamètres peut également générer d'autres risques opérationnels que les seuls risques liés au surajustement. Dans le cadre d'un modèle de scoring bancaire, Hurlin *et al.* (2022) montrent ainsi que de petites modifications des hyperparamètres peuvent avoir des conséquences importantes en termes d'équité et d'apparition de biais de discrimination

38. La stratégie de validation croisée de l'étude est basée sur un découpage chronologique en 6 blocs. Pour chaque ensemble de valeurs d'hyperparamètres fixées lors de la recherche sur la grille, le modèle est entraîné 6 fois sur 5 années de données et évalué sur l'année d'observation restante, extraite de l'échantillon d'entraînement et utilisé en tant que base de validation. Les auteurs cherchent alors les hyperparamètres qui minimisent l'AUC moyenne obtenue sur les 6 ensembles de validation. Pour les forêts aléatoires, le nombre d'arbres dans la forêt est fixé, tandis que l'hyperparamètre optimisé est la profondeur de l'arbre. Les paramètres optimaux sont obtenus par une recherche directe sur une grille de valeurs définies sur la profondeur maximale des arbres.

(cf. section 6.3). Plus précisément, l'étude montre qu'un changement dans les grilles de valeurs utilisées pour l'optimisation des hyperparamètres d'un arbre de classification, conduit à obtenir deux modèles de score présentant des performances prédictives globalement comparables (AUC de 0,88 contre 0,83), mais dont l'un génère une forte discrimination de genre dans l'octroi du crédit. [39] Ainsi, même si le genre n'est pas inclus dans l'espace des caractéristiques du modèle, un arbre de classification suivant la valeur des hyperparamètres, peut conduire à une discrimination de genre par proxy, le genre étant approximé par d'autres variables légitimes dans les règles de décision.

6.3 L'évaluation de l'équité et les biais systématiques

La gouvernance du ML dans le contexte IRB s'inscrit à la fois dans le cadre réglementaire des CRD/CRR, mais aussi dans le contexte plus général des règlements qui encadrent l'utilisation et la gouvernance de l'IA. C'est notamment le cas de la proposition européenne de règlement sur l'IA, connue sous le nom *d'Artificial Intelligence Act*. Cette proposition de règlement publiée le 21 avril 2021, vise à réguler l'utilisation de l'IA dans l'Union européenne en établissant des normes de transparence, de supervision et de responsabilité pour les entreprises qui développent, déploient ou utilisent ces systèmes. [40] Dans ce contexte, le choix européen est de réguler en priorité les systèmes d'IA qualifiés d'applications « à haut risque », celles-ci étant définies comme présentant des risques importants pour la santé, la sécurité ou les droits fondamentaux des personnes. L'article 6 du titre III du règlement et l'annexe III énoncent les règles de classification et définissent deux grandes catégories de systèmes d'IA à haut risque : les systèmes d'IA destinés à être utilisés en tant que composants de sécurité de produits et les autres systèmes d'IA autonomes qui soulèvent principalement des questions quant au respect des droits fondamentaux. C'est précisément dans cette seconde catégorie que sont considérés « *les systèmes d'IA utilisés pour évaluer la note de crédit ou la solvabilité des personnes physiques car ils conditionnent l'accès de ces personnes à des ressources financières ou à des services essentiels* ». [41] Pour la Commission Européenne, la

39. Dans cette étude, les hyperparamètres optimisés dans le cadre d'une procédure de type 5-fold sont la profondeur maximale de l'arbre, le nombre minimal d'individus requis pour diviser un nœud, le nombre minimal d'individus par feuille, la valeur seuil pour la diminution de l'impureté.

40. Ce projet de règlement doit encore être adoptée par le Parlement européen et le Conseil de l'Union européenne avant de s'appliquer à tous les pays de l'Union.

41. Le règlement prévoit toutefois une exception pour les systèmes d'IA utilisés à des fins d'évaluation de la solvabilité et de notation de crédit lorsqu'ils sont mis en service par des petits fournisseurs, typiquement des Fintechs, pour leur usage propre. Cette exemption, destinée à ne pas étouffer l'innovation financière, permet de ne pas soumettre aux mêmes exigences réglementaires les grandes banques et les petits acteurs de l'industrie du crédit.

crainte est ici que les systèmes d'IA utilisés dans le domaine du crédit puissent conduire à « une discrimination à l'égard de personnes ou de groupes et perpétuer des schémas historiques de discrimination, par exemple fondés sur les origines raciales ou ethniques, les handicaps, l'âge ou l'orientation sexuelle, ou créer de nouvelles formes d'incidences discriminatoires ». [42] Si le terme de systèmes d'IA « utilisés à des fins d'évaluation de la solvabilité et de notation de crédit » renvoie de façon évidente aux modèles de scores utilisés de façon directe ou indirecte (comme aide à une décision humaine) dans un processus d'octroi, la question se pose de savoir s'il couvre également les modèles IRB et tous les modèles de suivi des risques utilisés en internes par les établissements de crédit. Dans ce cas, les exigences de transparence, de fourniture d'informations aux utilisateurs, de contrôle humain, de robustesse et de sécurité, induites par l'AI act viendront ainsi se rajouter aux exigences définies dans le cadre des CRR. Du fait de la très grande proximité des exigences requises dans les deux cadres réglementaires, la Commission a d'ailleurs prévu que les superviseurs bancaires nationaux et l'EBA soient également en charge de veiller à l'application des recommandations de l'AI act dans les banques.

Au coeur de la préoccupation de l'AI act, figure la question de l'équité algorithmique (*fairness*). Un algorithme est dit non-équitable s'il désavantage de façon systématique un groupe d'individus qui partagent un attribut protégé (e.g., genre, âge, lieu de résidence, origine ethnique). Dans le cas d'un modèle de risque de crédit, ce désavantage peut se traduire en termes d'accès au crédit (taux d'acceptation) ou de coût (taux d'intérêt). La plupart des études académiques sur la question portent sur l'octroi de crédit. Ainsi, Bartlett *et al.* (2022) montrent, qu'aux États-Unis, les emprunteurs hispaniques et afro-américains paient respectivement $7,9$ et $3,6$ points de base de plus en intérêts pour l'achat d'un logement ou le refinancement d'un prêt hypothécaire, en raison de discriminations. Cependant, l'étude montre également que les Fintech dont les décisions de crédit sont fondées sur des algorithmes, réduisent ces disparités de taux de plus d'un tiers entre les emprunteurs et ne font preuve d'aucune discrimination dans les taux de rejet. A l'inverse, en utilisant des données administratives détaillées sur les prêts hypothécaires américains, Fuster *et al.* (2022) montrent que l'utilisation d'algorithmes de ML augmente la disparité des taux d'intérêt entre les emprunteurs blancs/asiatiques et les emprunteurs noirs/hispaniques.

42. Du fait de cette justification, le projet de règlement ne concerne pas à ce stade les modèles d'évaluation pour les crédits aux personnes morales, i.e., entreprises ou organisations, mais rien ne garantit que cette distinction soit maintenue à l'avenir.

Ces conclusions divergentes mettent en exergue la difficulté d'évaluer l'équité d'un modèle de risque de crédit, et notamment (1) d'identifier les variables qui doivent être considérées comme des attributs protégés et pour lesquelles les banques doivent démontrer l'absence de discrimination, (2) de définir statistiquement le concept d'équité et de mesurer les éventuelles discriminations, et (3) de comprendre l'origine d'une violation de l'équité. Concernant le premier point, certaines législations listent précisément les sources de discrimination qui doivent être évaluées. Par exemple, aux Etats-Unis, la loi sur le crédit à la consommation (*Equal Credit Opportunity Act* ou ECOA) interdit toute discrimination opérée sur la base de l'origine raciale, du sexe, de la couleur, de l'âge, de la nationalité, de l'état civil ou de la perception de revenus provenant d'un programme d'aide publique. A l'inverse en Europe, l'AI act ne propose qu'une liste d'exemples de sources possibles de discrimination (e.g., origine ethnique, handicap, âge et orientation sexuelle). Par ailleurs, la plupart de ces sources ne peuvent pas être testées en pratique puisque, pour pouvoir tester si un modèle génère un biais de discrimination par rapport à une caractéristique, il faut détenir des données sur cette variable. C'est pourquoi, en pratique seules les discriminations liées au genre, à l'âge ou à la localisation géographique, peuvent être évaluées par les banques dans leur processus de validation de modèle. La deuxième difficulté porte sur la définition statistique qu'il convient de donner à la notion d'équité. Il existe des dizaines de définitions statistiques de l'équité (Verma et Rubin, 2018), certaines étant incompatibles entre elles (Barocas *et al.*, 2022). [43] De plus, pour une définition donnée, se pose la question de savoir à partir de quel seuil de déviation, on doit considérer un modèle comme étant non équitable et générateur de discrimination. Dans ce contexte, Hurlin *et al.* (2022) proposent une procédure générale d'inférence, permettant de tester l'hypothèse d'équité ou de non-équité au sein de chaque classe de risque. Enfin, la dernière question porte sur l'origine d'une éventuelle discrimination dans un modèle. Dans le cas d'une discrimination indirecte, i.e., non liée à l'introduction directe de la variable protégée dans le modèle, il convient d'identifier les variables légitimes qui permettent de trianguliser l'attribut protégé et sont responsables de la discrimination. Cette identification est d'autant plus compliquée que le modèle de ML est opaque et fortement flexible. Une solution consiste à mobiliser des outils dérivés des méthodes de ML interprétable, comme les *Fairness Partial Dependence Plots* (FPDP) récemment proposés par Hurlin *et al.* (2022).

43. La définition la plus couramment utilisée est celle de la parité statistique, qui implique l'égalité des probabilités d'être classé comme bon type dans les groupes d'individus présentant l'attribut protégé et ceux qui ne le présentent pas.

Il est important de noter qu'un biais de discrimination peut apparaître de façon tout à fait involontaire, du fait d'une discrimination indirecte (par proxy). Pour autant, le modélisateur et la banque ne pourront arguer du caractère involontaire de la discrimination ou de la non-interprétabilité du modèle, pour se dédouaner de leurs responsabilités juridiques. Aux États-Unis, ce principe a été récemment réaffirmé par le *Consumer Financial Protection Bureau* dans une décision de mai 2022. Cette décision confirme que la loi fédérale ECOA exige que les banques puissent expliquer aux demandeurs les raisons spécifiques pour lesquelles elles refusent une demande de crédit, même si le créancier s'appuie sur des modèles de crédit utilisant des algorithmes complexes et non interprétables. Toutefois, il convient de bien dissocier les questions d'interprétabilité et d'équité (EBA, 2020) : un modèle peut être interprétable, ou rendu interprétable ex-post par une technique de ML interprétable, et pour autant générer des biais de discrimination. A l'inverse, un modèle boîte noire n'est pas nécessairement à l'origine de ce type de biais de discrimination, qui peuvent préexister dans les données.

7 Conclusion

Plusieurs conclusions et recommandations émergent de notre analyse. Tout d'abord, le ML présente d'indéniables avantages dans le contexte IRB. Utilisées en amont ou en aval de la phase de différenciation des risques, ces méthodes permettent d'importants gains de productivité et d'efficacité, comme par exemple dans la préparation des données, la transformation des variables, ou dans la construction de modèles challengers. Utilisés dans la phase centrale de différenciation des risques, certains modèles de ML, notamment les modèles d'ensemble homogènes, permettent d'obtenir des estimations des paramètres de risque (PD ou LGD) plus précises que celles fournies par les approches paramétriques usuelles. Toutefois, ces gains ont tendance à plafonner avec la complexité des modèles. Ainsi, dans la littérature académique, il devient de plus en plus difficile d'améliorer significativement les performances des modèles de scoring bancaire en développant de nouvelles méthodes de classification. Enfin, l'utilisation de certains modèles de ML pour la différenciation des risques, offre des perspectives de réductions significatives des fonds propres réglementaires, comprises entre 2% et 20% suivant les études, tout en respectant les critères de validation imposés par le régulateur. A ce stade, aucune méthode de ML ne semble se dégager en la matière, le choix de l'algorithme de ML influençant à la fois la probabilité d'obtenir l'approbation réglementaire

et le niveau des exigences en fonds propres.

Mais l'adoption du ML soulève également plusieurs défis. Le premier est celui de l'acceptabilité de ces méthodes, à la fois par le régulateur et par les banques. Lorsque les méthodes de ML sont utilisées pour la validation des modèles internes ou en amont de la différenciation des risques, cette acceptabilité semble, au final, poser peu de problèmes. Il n'en va pas de même concernant la différenciation des risques. Comment la banque pourrait-elle justifier l'utilisation d'une forêt aléatoire ou d'un modèle XGBoost, en lieu et place du modèle standard de la régression logistique ? Si ce changement génère une trop forte diminution des exigences en fonds propres réglementaires, il pourrait être sujet à une suspicion d'arbitrage réglementaire de la part du superviseur ou des équipes de validation interne. Ce risque est d'autant plus grand, qu'à méthode de ML donnée, le choix des hyperparamètres peut affecter de façon assez peu transparente la complexité du modèle et ses performances prédictives. Si à l'inverse, ce changement génère une faible diminution (voire une augmentation) du capital réglementaire, la banque n'aura aucun intérêt à adopter le ML et à refondre ses modèles internes. Par conséquent, seuls des critères objectifs de mesure statistique des gains en termes de précision dans les estimations des paramètres de risque peuvent résoudre ce dilemme. Mais là encore, la littérature académique montre clairement que les modèles de ML qui permettent de mieux estimer les paramètres de risque au sens des critères statistiques usuels (AUC, sensibilité, précision, etc.), ne sont pas toujours nécessairement ceux qui permettent de réduire les RWA ou de passer les tests de validation permettant d'obtenir l'approbation du superviseur.

Le deuxième défi lié à l'utilisation du ML dans le contexte IRB est celui de l'interprétabilité et de l'explicabilité des modèles de ML. Il est évident qu'il n'est pas possible de confier la détermination des fonds propres réglementaires d'une banque à un modèle de type boîte noire. Dans ce contexte, la démarche usuelle consiste à utiliser des modèles de ML non interprétables couplés à des techniques permettant de rendre leurs prédictions et leur fonctionnement général interprétables ex-post. Le problème est qu'il n'existe pas *une* explication universelle du fonctionnement d'un modèle complexe. Au contraire, il existe une pléthore de méthodes d'interprétabilité fournissant tout autant d'explications sur le fonctionnement global d'un modèle ou sur les décisions prises par ce modèle au niveau d'un emprunteur ou d'un contrat. La question se pose alors de savoir quel type d'interprétations et d'explications attend le superviseur concernant le fonctionnement des modèles internes. C'est sans

aucun doute cette incertitude sur la procédure de validation réglementaire des modèles de ML qui explique qu'aujourd'hui ces outils sont encore très peu utilisés par les banques dans les contextes IRB ou IFRS9.

Nous montrons qu'une façon de sortir de cette situation consiste à promouvoir l'utilisation de modèles de ML nativement interprétables dans le contexte IRB. Des travaux théoriques récents et des études empiriques dans le domaine du scoring de crédit remettent en cause l'idée d'un nécessaire arbitrage entre performance et interprétabilité. Utilisés dans le contexte de la différenciation des risques, des modèles de ML interprétables peuvent tout à fait fournir des estimations des paramètres de risque aussi précises que les modèles de type boîte noire, habituellement considérés comme plus performants. Même si ces nouvelles approches permettent de lever les freins au développement du ML liés à la question de l'interprétabilité, n'en demeurent pas moins les défis de la gouvernance, de la formation des personnels, des risques opérationnels et de l'évaluation de l'équité algorithmique, notamment en lien avec l'AI act.

Références

ACPR : Artificial intelligence : Challenges for the financial sector. Discussion papers publication, December, 2018.

ACPR : Governance of artificial intelligence in finance. Discussion papers publication, November, 2020.

H. J. Allen : "Just Say No" to machine learning in IRB models. *Oxford Business Law Blog*, 2021. URL https://blogs.law.ox.ac.uk/business-law-blog/blog/2021/12/just-say-no-machine-learning-irb-models.

A. Alonso et J. Carbó : Machine learning in credit risk : Measuring the dilemma between prediction and supervisory cost. *Working Paper, Banco de Espana*, 2020.

E. Altman, G. Marco et F. Varetto : Corporate distress diagnosis : Comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking and Finance*, 18:505–529, 1994.

D. W. Apley et J. Zhu : Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 82(4):1059–1086, 06 2020.

B. Baesens, T. V. Gestel, S. Viaene, M. Stepanova, J. Suykens et J. Vanthienen : Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

S. Barocas, M. Hardt et A. Narayanan : *Fairness and Machine Learning*. 2022. http://www.fairmlbook.org.

R. Bartlett, A. Morse, R. Stanton et N. Wallace : Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*, 143:30–56, 2022.

T. Bellotti et J. Crook : Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1):171–182, 2012. ISSN 0169-2070.

T. Berg, V. Burg, A. Gombović et M. Puri : On the rise of fintechs : Credit scoring using digital footprints. *Review of Financial Studies*, 33(7):2845–2897, 2020.

J. Bergstra et Y. Bengio : Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

P. Bracke, A. Datta, C. Jung et S. Sen : Machine learning explainability in finance : An application to default risk analysis. Bank of England, Staff Working Paper No. 816, 2019.

L. Breiman : Bagging predictors. *Machine Learning*, 26:123–140, 1996.

L. Breiman : Random forest. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. Friedman, C. Stone et R. Olshen : *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

C. Carter et J. Catlett : Assessing credit card applications using machine learning. *IEEE Expert*, 2:71–79, 1987.

A. Charpentier, E. Flachaire et A. Ly : Econometrics and machine learning. *Economics and Statistics*, 505-506:147–169, 2018.

C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang et T. Wang : An interpretable model with globally consistent explanations for credit risk. *arXiv*, 2018. URL https://doi.org/10.48550/arXiv.1811.12615.

T. Chen et C. Guestrin : XGBoost : A scalable tree boosting system. *In Krishnapuram et al. (eds)*, p. 785–794. 2016.

Y. Chen, R. Calabrese et B. Martin-Barragan : Effects of imbalanced datasets on interpretable machine learning. *Working Paper, University of Edimburgh*, 2022.

L. Clerc, A. Moraglia et S. Peyron : Les néobanques vont-elles bouleverser leur secteur d'activité ? *Revue d'économie financière*, 135(3):165–180, 2020.

X. Dastile, T. Celik et M. Potsane : Statistical and machine learning models in credit scoring : A systematic literature survey. *Applied Soft Computing*, 91:106263, 2020.

E. Dumitrescu, S. Hué, C. Hurlin et S. Tokpavi : Machine learning for credit scoring : Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3):1178–1192, 2022.

EBA : Report on Big Data and advanced analytics. European Banking Authority, January, 2020.

EBA : Discussion paper on machine learning for IRB models. European Banking Authority, November, 2021.

EC : White paper on artificial intelligence : A European approach to excellence and trust. European Commission, February, 2020.

EC : Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence, COM/2021/206 final. European Commission, April, 2021.

E. Flachaire, G. Hacheme, S. Hué et S. Laurent : GAM(L)A : An econometric model for interpretable machine learning. *arXiv*, 2022. URL https://doi.org/10.48550/arXiv.2203.11691.

H. Fraisse et M. Laporte : Return on investment on artificial intelligence : The case of bank capital requirement. *Journal of Banking & Finance*, 138:106401, 2022.

Y. Freund et R. Schapire : A decision-theoric generalization of on-line learning and application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J. Friedman : Greedy function approximation : A gradient boosting machine. *Annals of statistics*, p. 1189–1232, 2001.

A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai et A. Walther : Predictably unequal ? The effects of machine learning on credit markets. *Journal of Finance*, 77:5–47, 2022.

N. Grennepois, A. Alvirescu et M. Bombail : Using random forest for credit risk models. *Deloitte Risk Advisory*, 2018.

B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir et W. Lemahieu : Deep learning for credit scoring : Do or don't ? *European Journal of Operational Research*, 295(1):292–305, 2021.

D. Guégan et B. Hassani : Regulatory learning : How to supervise machine learning models ? an application to credit scoring. *The Journal of Finance and Data Science*, 4 (3):157–171, 2018.

C. Hurlin, J. Leymarie et A. Patin : Loss functions for loss given default model comparison. *European Journal of Operational Research*, 268(1):348–360, 2018.

C. Hurlin et C. Pérignon : Machine learning et nouvelles sources de données pour le scoring de crédit. *Revue d'économie financière*, 135(3):21–50, 2019.

C. Hurlin, C. Pérignon et S. Saurin : The fairness of credit scoring models. *arXiv*, 2022. URL https://doi.org/10.48550/arXiv.2205.10200.

S. Hué, C. Hurlin, C. Pérignon et S. Saurin : Explainable performance. *arXiv*, 2022. URL https://doi.org/10.48550/arXiv.2212.05866.

IIF : Machine learning in credit risk. Institute of International Finance, 2nd Edition Summary Report, August, 2019.

IIF : Survey report on machine learning uses in credit risk and AML applications. Institute of International Finance and EY, Public Summary, December, 2022.

J. Jagtiani et C. Lemieux : The roles of alternative data and machine learning in fintech lending : Evidence from the LendingClub consumer platform. *Financial Management*, 48 (4):1009–1029, 2019.

S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu et H. Lakkaraju : The disagreement problem in explainable machine learning : A practitioner's perspective. *arXiv*, 2022. URL https://doi.org/10.48550/arXiv.2202.01602.

S. Lessmann, B. Baesens, H.-V. Seow et L. C. Thomas : Benchmarking state-of-the-art classification algorithms for credit scoring : An update of research. *European Journal of Operational Research*, 247(1):124 – 136, 2015.

G. Loterman, I. Brown, D. Martens, C. Mues et B. Baesens : Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28 (1):161–170, 2012.

S. M. Lundberg et S.-I. Lee : A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4768–4777, 2017.

P. Makowski : Credit scoring branches out. *The Credit World*, 75:30–37, 1985.

A. Markov, Z. Seleznyova et V. Lapshin : Credit scoring methods : Latest trends and points to consider. *The Journal of Finance and Data Science*, 8:180–201, 2022.

C. Molnar : *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

C. Molnar, G. Casalicchio et B. Bischl : Interpretable machine learning – a brief history, state-of-the-art and challenges. *arXiv*, 2022. URL https://doi.org/10.48550/arXiv.2010.09337.

S. Mullainathan et J. Spiess : Machine learning : An applied econometric approach. *Journal of Economic Perspectives*, 31:87–106, 2017.

M. Oskarsdottir, C. Bravo, C. Sarraute, J. Vanthienen et B. Baesens : The value of big data for credit scoring : Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74:26–39, 2019.

P. Ould : Improving IRB and RWA calculations with machine learning. *FICO Decision Blog*, February, 2022. URL https://www.fico.com/blogs/improving-irb-and-rwa-calculations-machine-learning.

L. E. Papke et J. M. Wooldridge : Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6):619–632, 1996.

M. Ribeiro, S. Singh et C. Guestrin : "Why should I trust you ?" : Explaining the predictions of any classifier. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, p. 97–101, juin 2016.

J.-C. Rochet et J. Tirole : Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029, 06 2003.

C. Rudin : Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

L. Semenova, C. Rudin et R. Parr : On the existence of simpler machine learning models. *In 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, p.

1827–1858, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522.

L. S. Shapley : A value for n-person games. *In Contributions to the Theory of Games*, p. 307–317. 1953.

V. Srinivasan et Y. Kim : Credit granting : A comparative analysis of classification procedures. *Journal of Finance*, 42:665–683, 1987.

A. Sudjianto, J. Qiu, M. Li et J. Chen : Linear iterative feature embedding : An ensemble framework for interpretable model. *arXiv*, 2021. URL https://doi.org/10.48550/arXiv.2103.09983.

A. Sudjianto et A. Zhang : Designing inherently interpretable machine learning models. *arXiv*, 2021. URL https://doi.org/10.48550/arXiv.2111.01743.

K. Tam et M. Kiang : Managerial applications of neural networks : The case of bank failure predictions. *Management Science*, 38:926–947, 1992.

L. Thomas : A survey of credit and behavioural scoring : Forecasting financial risk of lending to customers. *International Journal of Forecasting*, 16:149–172, 2000.

R. Tibshirani : Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246.

E. N. Tong, C. Mues, I. Brown et L. C. Thomas : Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research*, 252 (3):910–920, 2016. ISSN 0377-2217.

V. Vapnik : The nature of statistical learning theory. Springer Science & Business Media, 2013.

T. Verdonck, B. Baesens, M. Óskarsdóttir et S. vanden Broucke : Special issue on feature engineering editorial. *Machine Learning*, 2021.

S. Verma et J. Rubin : Fairness Definitions Explained. *In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, p. 1–7, 2018.

X. Yao, J. Crook et G. Andreeva : Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, 263(2):679–689, 2017. ISSN 0377-2217.

H. Zou : The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.