Janvier 2022

Tech Sprint sur l'explicabilité des algorithmes d'intelligence artificielle

Rapport de synthèse

Auteur : Laurent Dupont, Pôle Fintech-Innovation, ACPR

ACPR
BANQUE DE FRANCE

Le premier Tech Sprint de l'ACPR s'est déroulé en juin et juillet 2021. Le défi : produire des explications permettant de comprendre le comportement de modèles prédictifs de risque de crédit basés sur de l'intelligence artificielle (IA) et uniquement accessibles en « boîte noire »[1].

Le Pôle Fintech-innovation de l'ACPR a créé, organisé et facilité cet événement, également appelé hackathon réglementaire. Pour ce faire, elle a collaboré avec 4 établissements de crédit volontaires (les « partenaires ») qui ont conçu et entraîné des modèles de Machine Learning (ML) sur un cas d'usage convenu, à savoir la prédiction du risque de défaut de crédits individuels[2].

Les participants du Tech Sprint incluaient des professionnels de fintechs, de banques ou d'autres acteurs financiers, ainsi que des chercheurs et étudiants en *data science* et informatique. Ces participants avaient constitué des équipes jouant le rôle des « analystes ». Leur but premier était d'expliquer le comportement des modèles prédictifs et d'élucider leur nature et leurs caractéristiques.

Ce rapport décrit l'origine du Tech Sprint, l'événement lui-même, puis résume les enseignements majeurs tirés du défi par l'ACPR, et enfin évoque de futurs travaux potentiels sur l'explicabilité de l'IA ou des thèmes adjacents.

Table des matières

[1] Le terme boîte noire désigne un modèle dont le fonctionnement interne est masqué à l'observateur, et pour lequel seules les données d'entrée (dans le cas présent la demande de crédit ou le crédit en cours) et de sortie (par exemple la probabilité de défaut prédite) sont visibles. Par extension, le terme s'applique aussi aux modèles observables mais dont l'architecture est trop complexe pour être totalement assimilable (typiquement un réseau neuronal profond par contraste avec une régression linéaire).

[2] Plus précisément, la prédiction des modèles était soit la probabilité de défaut des crédits en cours, soit un indicateur indirect (en ce cas, la présence d'un individu dans un registre de risque de crédit).

1. Conception du Tech Sprint

1.1. Le rapport de 2020

Le Tech Sprint s'inscrit dans la lignée des travaux précédents du Pôle Fintech-innovation de l'ACPR sur l'IA, et notamment du document de discussion sur la gouvernance de l'IA en finance, publié et soumis à consultation publique en juin 2020. Ce document s'articulait autour de deux axes d'étude : l'évaluation et la gouvernance des algorithmes d'IA.

Notre analyse conduisait ainsi à identifier, dans la conception et l'évaluation des algorithmes et outils d'IA en finance, quatre critères interdépendants propres à limiter les risques liés à l'utilisation de l'IA dans des processus métier : le traitement adéquat des données, la performance de l'algorithme, sa stabilité, et son explicabilité.

Par ailleurs, l'inclusion d'IA influant nécessairement sur la gouvernance des processus associés, nous recommandions de porter l'attention, dès la phase de conception des algorithmes, sur les aspects suivants : leur intégration dans les processus métiers, les interactions entre humain et algorithme, la sécurité et les questions d'externalisation, les processus de validation initiale et continue, enfin leur audit interne ou externe.

La consultation publique, dont une synthèse des résultats a été publiée en décembre 2020, a confirmé la pertinence de ces critères de conception et d'évaluation, ainsi que des principes de gouvernance énoncés.

1.2. Le thème de l'explicabilité

Parmi les principes de conception et d'évaluation de l'IA, l'explicabilité faisait l'objet d'une attention particulière. C'est en effet le principe qui distingue le plus l'IA du reste ; de plus il apparaît comme le principe sous-jacent à un développement maîtrisé de l'IA en finance en rendant possible le contrôle interne et l'audit externe.

Définition

Dans le cadre des travaux de l'ACPR sur l'IA, le concept d'explicabilité recouvre deux questions. D'une part le « comment » c'est à dire la question du fonctionnement de l'algorithme, autrement dit de sa transparence : un enjeu majeur en est l'auditabilité d'une solution algorithmique. D'autre part le « pourquoi », c'est-à-dire la question de l'interprétabilité, avec comme enjeux associés la compréhension du comportement du système par les opérateurs humains qui interagissent avec lui, mais aussi par le client, et l'acceptabilité sociale ou éthique.

La question de l'explicabilité se pose d'ailleurs très concrètement dès la conception de systèmes à base d'IA, pour laquelle un continuum existe entre les deux extrêmes que sont l'IA en boîte noire et l'IA totalement interprétable[3].

[3] Idée promue par exemple par Cynthia Rudin et qui vise à interdire les boîtes noires mais qui s'avère bien souvent un idéal vers lequel tendre plutôt qu'un objectif réaliste. Cf. *"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead"*, Cynthia Rudin (2018).

Niveaux d'explication

Le document de juin 2020 proposait une échelle de 4 niveaux permettant de caractériser une explication algorithmique :

Explication de niveau 1 : observation

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

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

Explication de niveau 2 : justification

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

- soit par la présentation simplifiée d'éléments explicatifs issus de niveaux plus élevés (3 et 4), éventuellement assortis d'explications contrefactuelles;
- soit par la génération par l'algorithme lui-même de justifications obtenues par apprentissage.

Explication de niveau 3 : approximation

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

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

Explication de niveau 4 : réplication

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

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

Audience d'une explication

Le niveau d'explication attendu dépend avant tout de l'audience à qui s'adresse l'explication algorithmique. En effet, les différences de sophistication technique ou métier, mais aussi les motivations intrinsèques à un destinataire particulier du discours explicatif, influent sur la forme de l'explication qu'il est pertinent de proposer. C'est pourquoi un même algorithme pourra être soumis à différents niveaux d'exigence selon que l'on considère un utilisateur final (pour qui une explication devra être directement intelligible) ou un auditeur (qui doit comprendre en détail le fonctionnement technique du système et qui est soumis à des exigences réglementaires fortes).

Notre document sur la gouvernance de l'IA recommandait donc d'adapter l'objectif des explications fournies (et partant, leur forme et leur contenu) à l'audience considérée :

- compréhension du comportement du système par les opérateurs humains qui interagissent avec lui (experts techniques ou expert métier) ;
- compréhension par le client auquel s'appliquent les décisions ou prédictions de l'algorithme ;
- acceptabilité sociale ou éthique de la solution considérée, par exemple afin de prouver l'absence de biais (implicites ou explicites) à caractère discriminatoire dans les décisions prises par l'algorithme ;
- conduite des missions de contrôle interne menées par les établissements assujettis eux-mêmes sur ces systèmes, mais aussi des missions de contrôle menées par l'autorité de supervision (qui peuvent dans chaque cas consister à inspecter des systèmes contenant un ou plusieurs modèles d'IA en « boîte noire »).

1.3. Le cas d'usage du risque de crédit

Le cas d'usage des modèles de risque de crédit est relativement courant parmi les hackathons en ML. Dans le cas du Tech Sprint, le choix a néanmoins convergé sur le risque de crédit à la consommation en raison des nombreux thèmes d'intérêt associés : stabilité financière, consommation des ménages, enjeux commerciaux, et questions socio-économiques telles que l'inclusion financière.

La Commission européenne a entretemps publié (en avril 2021) son projet de réglementation de l'IA. Ce projet proposait une échelle à 4 niveaux de risque et classait les modèles de risque de crédit comme seul cas d'usage à risque élevé du secteur financier, renforçant ainsi encore la pertinence de ce cas d'usage pour l'ACPR.

1.4. Enjeux

Le but principal promu par l'organisation du Tech Sprint à l'été 2021 était d'éclairer les défis réglementaires liés à l'IA et au ML, depuis la gestion des risques associés jusqu'à la protection des consommateurs en passant par la gouvernance des processus métier impactés.

Au plan technique, le Tech Sprint visait à explorer quelles méthodes explicatives s'appliquent à un cas d'usage concret, et quels types d'explications sont le plus adaptées à une variété de parties prenantes (auditeurs, experts métier, experts techniques, clientèle). Il a également contribué à la politique de l'ACPR consistant à promouvoir le partage de connaissances et à initier des travaux collaboratifs entre acteurs du secteur financier.

Parallèlement, le Pôle Fintech-innovation avait entamé une réflexion concernant l'audit, sur place et sur pièces, de systèmes à base d'IA – un domaine de recherche embryonnaire pour lequel un scénario d'application pratique semblait nécessaire.

2. L'événement

2.1. Objectifs

Le diagramme suivant, extrait du Guide de l'Analyste, résumait à l'intention des participants les objectifs du Tech Sprint :

- l'objectif principal d'explicabilité ;
- un objectif secondaire ou « bonus » de détection de biais ;
- pour éviter les malentendus, le guide de l'analyse précisait également que, à la différence de beaucoup de *hackathons* classiques, la mesure de performance et la mesure de bien-fondé (parfois appelée justification algorithmique) ne faisaient pas partie des objectifs (les « non-objectifs »).

2.2. Déroulé

La phase préparatoire du Tech Sprint a duré environ 6 mois, durant lesquels le Pôle Fintech-innovation et les banques partenaires du Tech Sprint ont collaboré à la conception et à l'apprentissage des modèles prédictifs. Ces modèles ont été sélectionnés pour leurs structures diverses et différents degrés de complexité, tout en demeurant représentatifs des modèles actuellement ou bientôt déployés en production par les établissements bancaires.

En fin de phase préparatoire, l'ACPR a déployé ces modèles en mode « boîte noire » pour le Tech Sprint : l'ACPR a installé les modèles, une fois entraînés par chaque banque partenaire sur un jeu de

ses données réelles, sur l'infrastructure de Cloud interne de la Banque de France. Les participants n'auraient ainsi lors du défi aucun accès direct ni aux modèles ni aux données d'entraînement, mais uniquement à une API[4] qui leur permettait d'interroger chaque modèle à partir d'un ensemble de caractéristiques d'un ou plusieurs clients, leur renvoyant la ou les probabilités de défaut estimées par le modèle. On notera aussi que chaque banque partenaire avait fourni un jeu de données de test (anonymisées mais représentatives) afin que les participants puissent entrer de plain-pied dans le défi d'explicabilité sans nécessairement passer par l'étape chronophage de génération de données synthétiques.

L'événement Tech Sprint proprement dit comportait deux sessions, afin de permettre à la fois aux équipes de professionnels et à celles constituées d'étudiants de participer au défi d'explicabilité. Chaque session s'étendait sur deux jours.

Lors du premier jour, au format typique d'un hackathon, les analystes assemblés en équipes œuvraient à expliquer les boîtes noires, assistés par l'équipe du Pôle Fintech-innovation et par les ressources qui leur avaient été fournies[5]. La première journée comportait aussi des activités en lien avec le thème telles qu'un exposé sur les facteurs sociocognitifs en jeu dans les explications d'IA.

La seconde journée permettait à chaque équipe d'analystes de préparer puis présenter ses travaux face à un jury composé de cadres dirigeants de la Banque de France et de l'ACPR. Les présentations étaient limitées à 5 minutes mais leur format libre permettait à chaque équipe d'exprimer sa créativité : slides, démo interactive, ou dans le cas d'une équipe une performance théâtralisée des explications produites par leur méthode.

Dans chaque session, les 3 équipes ayant montré les travaux les plus convaincants étaient récompensées par le jury sur la base des critères suivants :

- leurs accomplissements techniques ;
- leur innovation scientifique et méthodologique ;
- la qualité de leur restitution ;
- leur contribution aux enjeux métier et réglementaires (par exemple assister les experts métier dans leur interprétation des décisions algorithmiques, ou faciliter la maintenance de l'algorithme en décelant un comportement attendu).

2.3. Défi proposé

Un chemin semé d'embûches

La difficulté de la compétition amicale proposée par l'ACPR ne saurait être surestimée. En effet le défi avait été conçu sous une triple focale technique (un challenge générique d'explicabilité), fonctionnelle (le cas d'usage concret des modèles de risque de crédit à la consommation) et réglementaire (la situation d'audit en boîte noire).

Le Tech Sprint mêlait donc des obstacles de diverses natures, notamment les suivants :

- Le temps de latence lié à l'hébergement et à la construction (*round-trip time* + temps d'exécution des requêtes) peut impacter la méthode explicative utilisée en limitant le nombre maximal de prédictions par unité de temps.

[4] Pour *Application Programmable Interface*, permettant une interrogation programmatique d'un service logiciel.
[5] À commencer par le Guide de l'Analyste, envoyé à toutes les équipes sélectionnées et qui contenait de nombreuses ressources et liens utiles.

- Des données de test imparfaites – tant en qualité qu'en complétude – (précisément car les jeux de test mis à disposition par les banques partenaires étaient représentatifs des données terrain réelles) posent la question de l'opportunité de la génération de données synthétiques.
- Les outils d'explicabilité disponibles présentent des limitations. Ainsi certaines solutions courantes telles que DiCE et SHAP sont très coûteuses en temps de calcul, avec un phénomène d'aggravation dans le cas d'un modèle complexe qui est encore moins interprétable que des modèles simples tout en présentant un temps d'inférence potentiellement plus élevé.
- Un certain niveau d'expertise métier est nécessaire à la conceptualisation des explications, c'est à dire à la production d'explications utilisant des concepts de plus haut niveau sémantique[6] que les simples variables prédictives.
- Challenge sans doute le plus ardu, la situation d'explication en boîte noire impliquait une ignorance totale des algorithmes d'entraînement des modèles étudiés.

Nécessaire pluridisciplinarité

Le Tech Sprint a démontré les bénéfices tangibles d'une approche interdisciplinaire à l'explication algorithmique. En effet, les équipes combinant expertise en *data science* et *data engineering* avec des capacités de visualisation de données ont fourni les explications les plus exhaustives et précises, ainsi que les rendus visuels les plus convaincants de ces explications. Une connaissance des concepts des sciences sociocognitives peut s'avérer également précieuse pour adapter les explications à différents types d'audience.

Une participante résume les objectifs du Tech Sprint ACPR sur l'explicabilité de l'intelligence artificielle.

Dévoilement des boîtes noires

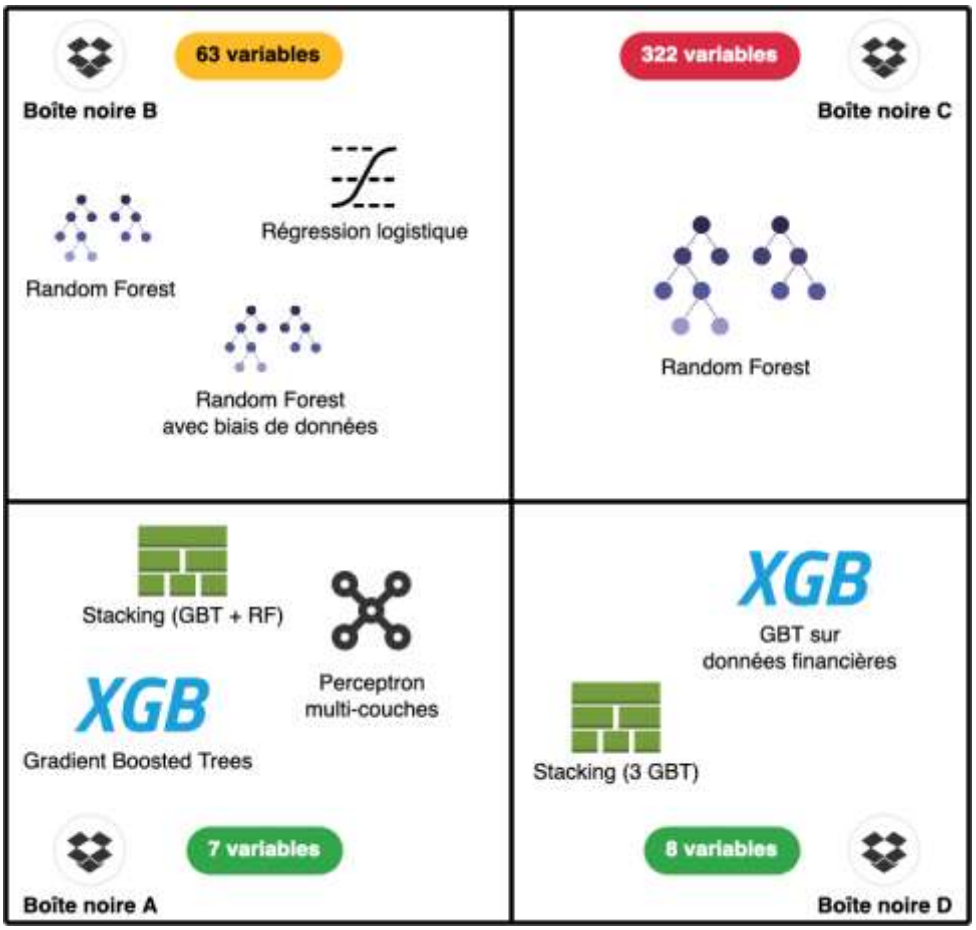
Le contenu de chaque boîte noire a été dévoilé lors de la journée de restitution, suite à la présentation des travaux des participants afin de garder le mystère intact.

[6] C'est-à-dire des concepts faisant sens pour les experts du domaine métier (ici le crédit à la consommation en général, et l'environnement d'un établissement bancaire en particulier) et non pas pour les experts techniques tels que data scientists et ingénieurs logiciels (qui manipulent des variables prédictives de plus bas niveau sémantique).

Comme déjà indiqué, les modèles proposés représentaient une variété de classes d'algorithmes (du linéaire au réseau neuronal, avec aussi de l'assemblage de modèles), d'API (prenant un seul ou une liste de clients en entrée), de schémas de données, de mode d'hébergement (cloud privé Banque de France ou sur une infrastructure externe). Néanmoins les modèles étaient réalistes et représentatifs de ce qui est soit en production, soit en expérimentation dans les banques partenaires.

Le contenu de chaque boîte noire peut être résumé ainsi et dans le diagramme qui suit :

- La boîte noire A contenait 3 modèles avec 7 variables prédictives : un modèle XGBoost, un modèle MLP (perceptron multi-couche), un modèle agrégeant un XGBoost et une forêt aléatoire par une régression logistique.
- La boîte noire B contenait 3 modèles avec 63 variables prédictives : un modèle réglementaire basé sur un assemblage de régressions logistiques, et une paire de forêts aléatoire qu'il était suggéré d'analyser ensemble afin de déterminer ce qui les distinguait (en l'occurrence un biais introduit dans les données d'apprentissage).
- La boîte noire C contenait un modèle avec 322 variables prédictives : une forêt aléatoire.
- La boîte noire D contenait 2 modèles avec 8 variables prédictives : un modèle de type GBT (*Gradient Boosted Tree*) simple et un modèle agrégé de 3 GBT. Une différence métier était présente dans cette boîte noire, la variable cible étant la probabilité d'appartenance à un fichier de risque de crédit (FICP ou FCC) maintenu par la Banque de France.

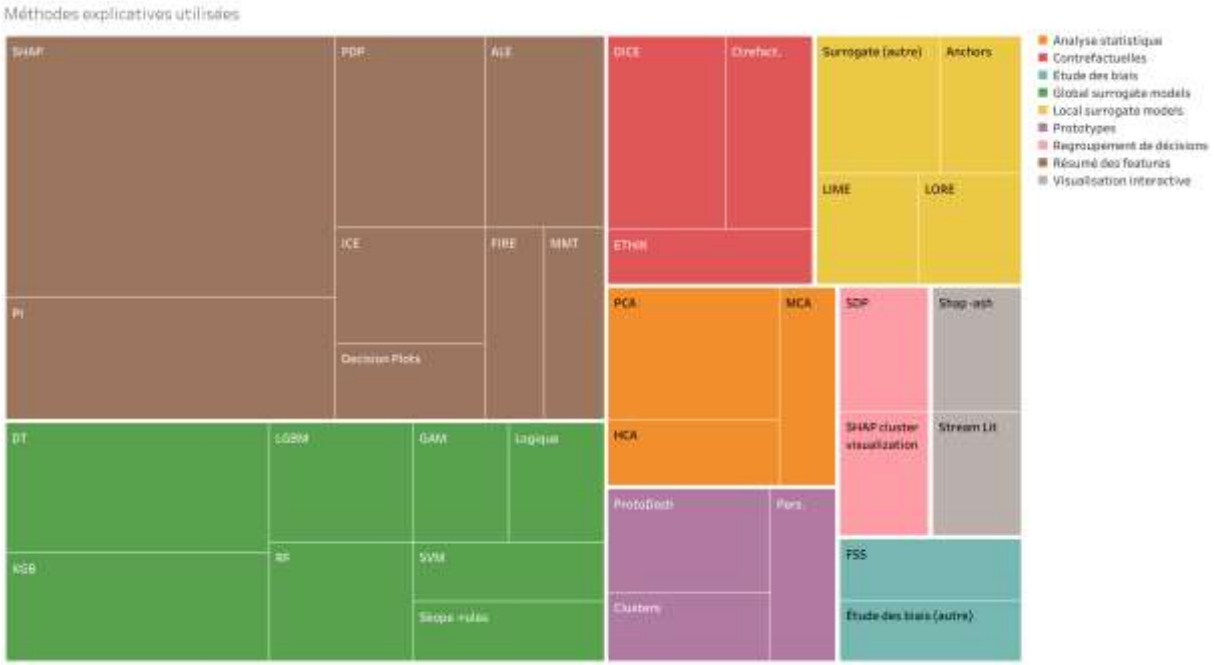


Par ailleurs, les modèles différaient par le traitement des valeurs manquantes en entrée : une imputation était opérée sur toutes les variables le cas échéant, mais selon un mécanisme plus ou moins rudimentaire ; la liste des variables prédictives utilisées *de facto* était fournie pour certains modèles mais pas pour tous, auquel cas un défi supplémentaire consistait à les inférer.

3. Méthodologie explicative

3.1. Méthodes utilisées pour expliquer un modèle

Le diagramme suivant montre les méthodes explicatives mises en œuvre lors du Tech Sprint.



1. Méthodes explicatives utilisées lors du Tech Sprint

On y constate que les équipes participantes ont déployé un arsenal méthodologique très large. Néanmoins quelques méthodes comme SHAP se dégagent, dont certaines limitations d'ordre technique (temps de calcul) ont été décrites précédemment mais dont l'intérêt méthodologique est aussi relativisé par certains experts[7].

La catégorie la plus fréquemment utilisée – incluant SHAP – est celle de *résumé* des *features*, ou calcul d'importance des variables : ces méthodes permettent de classer les variables des plus au moins prédictives, ou de visualiser l'effet d'une voire de deux variables sur la valeur estimée (en l'occurrence la PD).

De nombreuses équipes ont employé des modèles globaux de substitution (*global surrogate models*) : il s'agit de modèles entraînés à reproduire, avec le plus de précision possible, les prédictions du modèle étudié en boîte noire – tout en étant fondamentalement plus interprétables. Le recours si fréquent aux modèles globaux de substitution est surprenant car leur apprentissage nécessite une grande disponibilité technique des APIs (par conséquent il peut s'avérer très chronophage), pour un gain potentiellement limité voire nul si le modèle en boîte noire sous-jacent se révèle être relativement simple (par exemple un arbre de décision). Ce choix est néanmoins probablement motivé par

[7] Comme le montre la citation suivante : *"Mathematical problems arise when Shapley values are used for feature importance and that the solutions to mitigate these necessarily induce further complexity, such as the need for causal reasoning […] Shapley values are not a natural solution to the human-centric goals of explainability."* Elizabeth Kumar (2020)

l'impossibilité, dans le cadre du scénario proposé d'inspection en boîte noire, d'inférer la classe du modèle étudié et donc d'adopter des méthodes plus ciblées.

Des méthodes d'approximation locales, ou modèles locaux de substitution (*local surrogate models*), ont été utilisées. Ces méthodes incluent la technique LIME très couramment employée en explicabilité, toutefois elles présentent l'inconvénient d'être potentiellement moins fidèles pour des modèles non linéaires ou présentant des discontinuités.

On peut aussi noter l'utilisation par plusieurs équipes de méthodes basées sur des prototypes, représentant de façon synthétique un sous-domaine de l'ensemble du domaine possible des données d'entrée. Cette utilisation est, dans le cas d'usage considéré, en accord avec le principe de segmentation des consommateurs de crédit. On note en revanche que les méthodes dites à base de « critiques », c'est-à-dire faisant ressortir les zones de points peu représentées dans les données d'apprentissage, ont été négligées ; elles auraient toutefois pu révéler des anomalies pertinentes dans des zones de faible densité du domaine de données d'entrée.

Quelques méthodes purement statistiques méritent d'être soulignées, telles des méthodes d'analyse des composantes principales mais aussi une mesure d'importance des variables obtenue par permutation de façon assez originale. Cela montre la pertinence d'appliquer des méthodes purement statistiques existantes – y compris les plus communes – soit telles quelles soit en y apportant des variations.

Des méthodes de regroupement de décisions ont été mises en œuvre, telles que SDP *(Same Decision Probability)* ou le clustering SHAP, dont l'utilité sera décrite dans la section « Principe d'intelligibilité ».

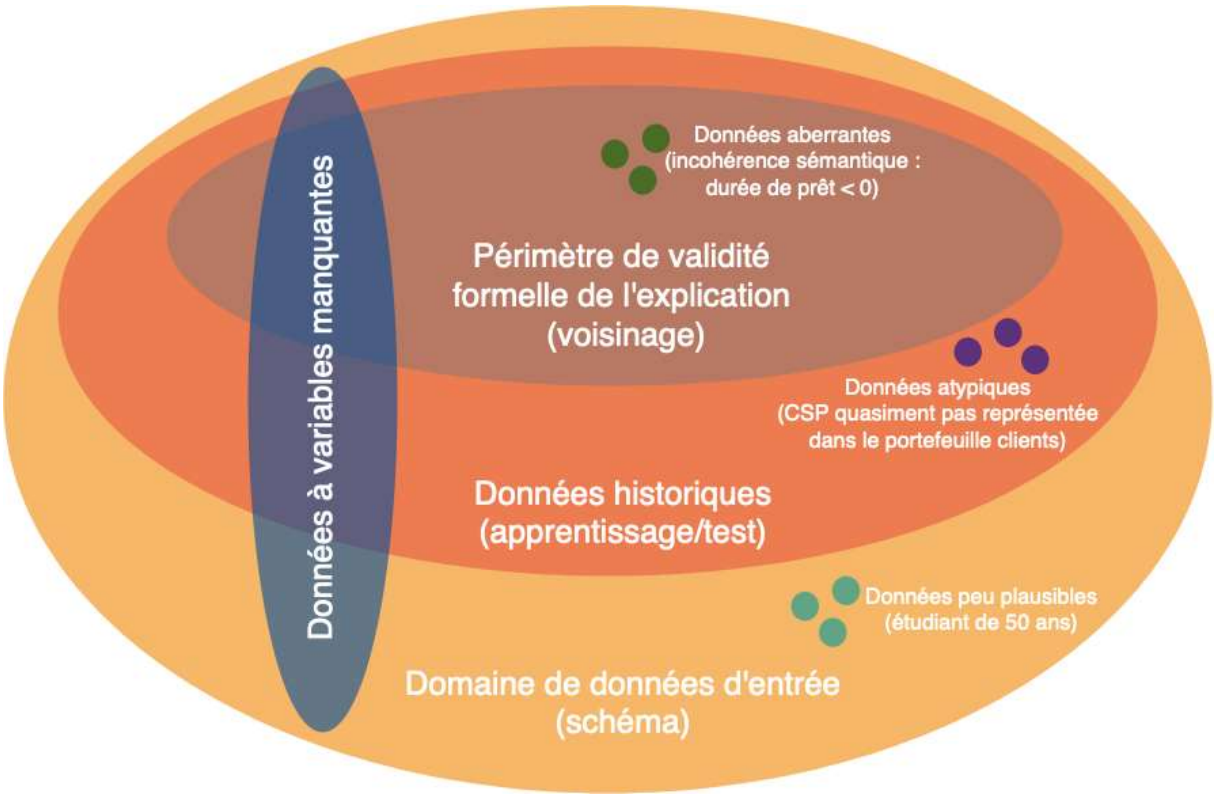
Enfin, dans la perspective de la restitution des explications à leurs destinataires, l'accent a été mis par les participants davantage sur la production d'explications aussi intelligibles que possible, et moins sur la visualisation – éventuellement interactive – des explications produites. Ce dernier point sera évoqué dans la section « Principe d'interactivité ».

3.2. Expliquer le modèle sur « toutes » les données

Le pouvoir de généralisation des explications produites est un objectif évident de toute méthode explicative d'algorithmes. La généralisation vise à expliquer le modèle sur autant de données d'entrée que possible, idéalement sur « toutes » les données. La question demeure toutefois du niveau de généralisation le plus approprié :

a) Une explication devrait-elle demeurer valide uniquement au sein de la distribution des données d'entraînement, ou également dans les régions du domaine d'entrée jamais observées ?
b) Convient-il d'expliquer le comportement du modèle sur des données aberrantes sur le plan de la cohérence ou de la sémantique ?
c) Faut-il expliquer son comportement sur des données peu plausibles ?
d) Faut-il l'expliquer sur des points de données ayant des attributs manquants ?

Le schéma suivant résume les différents périmètres possibles pour le pouvoir de généralisation d'une explication algorithmique :



2. Périmètres possibles pour le pouvoir de généralisation d'une explication

Cette question a été couverte par les travaux de certains analystes, qui ont soulevé des points subtils tels que l'applicabilité d'un méthode explicative idéale à des points atypiques (aux sens b et c ci-dessus). Par exemple, les données peu plausibles du point de vue métier (telles qu'un étudiant de 50 ans) peuvent induire des effets significatifs sur le comportement du modèle, mais dont l'intérêt comme objet d'explication ne fait pas consensus.

Un problème lié au pouvoir de généralisation, également souligné par les analystes, est le manque de réalisme des données utilisées par certaines méthodes, telles que les diagrammes PDP (*Partial Dependence Plots*): en effet, de par leur utilisation de distributions marginales, les PDP masquent la distribution des données réelles (ce qui peut conduire à sur-interpréter l'effet du modèle dans des zones de faible densité), en outre ils présupposent l'absence de corrélation entre la variable prédictive représentée et les autres variables.

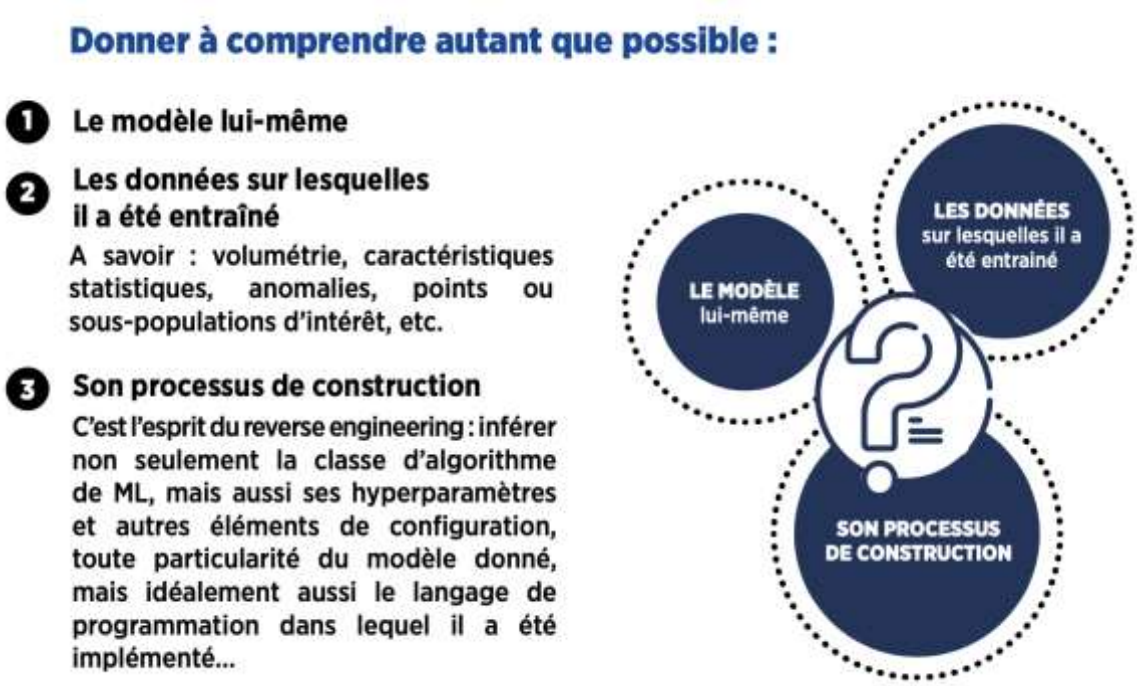
3.3. Expliquer au-delà du modèle

Les participants se sont donc essentiellement attachés à expliquer le comportement des modèles eux-mêmes, répondant à des questions telles que :

- Pourquoi telle prédiction individuelle ?
- Comment telle prédiction individuelle peut-elle être changée ?
- Comment le modèle se comporte-t-il dans certaines zones (anormales, avec peu de données, etc.) ?
- Comment des perturbations (changements mineurs) dans les données d'entrée affectent-elles la prédiction en sortie ?

Ils ont donc, aiguillés en cela par le Guide de l'Analyste qui leur avait été distribué, éludé la question de la justification (« Où le modèle a-t-il raison / tort ? ») et ont rarement abordé la question bonus de l'équité algorithmique (« Le modèle est-il sans biais ou équitable (*fair*) ? »).

Mais outre l'explicabilité du modèle lui-même, le Guide de l'Analyste indiquait l'intérêt d'expliquer les données d'apprentissage ainsi que le processus de construction :



3. Énoncé des objectifs d'explication "avancés"

Le reste de cette section décrit les travaux réalisés en ce sens et les difficultés rencontrées.

Expliquer les données d'apprentissage

La question est ici de savoir ce qui peut être inféré sur les données utilisées pour entraîner puis interroger les modèles (domaine de valeurs de chaque variable prédictive, leur distribution statistique, etc.), y compris pour les variables n'intervenant pas ou peu dans les prédictions.

À cet égard, les analystes se sont souvent bornés à calculer l'importance des variables dans les modèles, sans inférer de connaissance sur les données utilisées en apprentissage.

Au mieux ils ont glané des indices, en supposant (sans le prouver) que des comportements étonnants – par exemple un plateau de risque au-delà de 10 ans de durée d'emprunt – étaient liés à une trop faible densité des données.

En outre, afin de constituer un challenge d'explication des données d'apprentissage, une paire de modèles avait été proposée dans le Tech Sprint, les deux modèles ne différant que par les jeux d'entraînement : clients en défaut sans incident bancaire surpondérés dans le modèle B_3 par rapport au modèle B_2. Cette différence pouvait être déduite de la moindre importance des variables prédictives d'incident dans le modèle B_3, même si cette dernière observation aurait pu être attribuable à une

différence structurelle entre les deux modèles[8]. Ce challenge s'est avéré trop ardu dans le temps imparti pour le Tech Sprint.

Expliquer le processus d'apprentissage

La question est alors de savoir ce que l'on peut inférer de l'algorithme d'apprentissage : sa composition générale (est-ce une régression linéaire, un réseau de neurones, etc.), plus fine (nombre de paramètres et hyperparamètres), ou très détaillée (valeurs des paramètres et hyperparamètres).

Des méthodes très diverses, certaines inventives, ont été mises en œuvre par les participants dans ce but : visualisations de type PDP et ICE (*Independent Conditional Expectations*) pour détecter les arbres de décision, « hack » de la documentation des boîtes noires pour en déduire le contenu, etc.

En conclusion, l'inférence de l'algorithme d'apprentissage est extrêmement difficile pour un modèle en boîte noire. On peut supposer qu'une solution fiable et efficace (qui n'existe pas encore à notre connaissance) pour réaliser cette inférence combinera probablement des heuristiques complexes afin d'opérer sur les types d'algorithmes les plus courants.

3.4. Enseignements méthodologiques

Cette section énumère les principaux enseignements méthodologiques du premier Tech Sprint organisé par l'ACPR.

Ces leçons peuvent être résumées ainsi : une mission d'audit portant sur un algorithme d'IA nécessite d'assembler une équipe dotée d'une variété de méthodes, d'adopter une approche agile et de la maintenir sur toute la durée de la mission, en particulier en présence de modèles de ML accessibles uniquement en boîte noire.

« Acheter ou construire » une solution d'explicabilité

La question du « *build vs. buy* » (construire une solution ou l'acheter sur étagère) est bien connue de l'industrie logicielle. Le Tech Sprint a montré qu'elle se posait aussi en explicabilité de l'IA.

Ainsi plusieurs équipes, représentant des fournisseurs de technologie de petite ou moyenne taille, ont abordé le défi au moyen de leur propre plateforme de *data science*. Une telle plateforme permet généralement d'exécuter une analyse d'un modèle en mode plus ou moins « presse-bouton ». Les bénéfices associés sont typiquement :

- la robustesse d'une solution industrielle (le processus de production d'explications est testé et peut en cas d'erreur être aisément relancé) ;
- le traitement parallélisable des modèles (car ils sont traités de façon identique) ;
- des fonctionnalités prêtes à l'emploi telles que la génération automatique de rapports sur les modèles.

En revanche, ces outils ne permettent généralement pas de traiter différemment des modèles comportant moins de 10 variables prédictives et ceux en comportant plusieurs centaines, ne fournissent que rarement une couche d'abstraction sémantique des modèles étudiés, et disposent de capacités de visualisation limitées ou assez rigides.

[8] Cette hypothèse alternative pouvait à son tour être infirmée par une analyse comparée des mesures d'importance de variables entre les deux modèles.

D'autres équipes ont au contraire développé leur solution explicative *ab initio*, faisant souvent appel à des composants logiciels *open source*. Ces équipes étaient ainsi davantage en mesure d'ajuster les explications produites :

- au cas d'usage spécifique, en intégrant par exemple de la connaissance métier dans les explications produites sous forme de concepts de haut niveau ou d'une abstraction métier au-delà des variables brutes ;
- à la situation d'inspection en boîte noire, en réalisant par exemple des développements différents selon que le modèle permettait grâce à un temps de réponse très faible de construire un modèle de substitution, ou qu'une analyse multivariée des variables prédictives était plus appropriée.

Les inconvénients des approches ad-hoc étaient bien sûr le temps de conception et de développement, qui s'accommodait mal du format court du Tech Sprint, la relative fragilité de l'outil résultant et l'impossibilité de traiter l'ensemble des 9 modèles proposés avec un soin égal.

L'état de l'art de la R&D

Le Tech Sprint fut aussi l'occasion d'illustrer l'expertise et le savoir-faire français en IA explicable, et plus généralement en *data science*.

Outre les méthodes les plus courantes, les participants ont en effet mis à contribution l'état de l'art en explicabilité de l'IA, incluant leurs propres inventions – confirmant ainsi l'intensité et la qualité de la recherche française en la matière. Parmi ces techniques :

- La méthodes des « Active Coalitions of Variables » consiste à ne calculer les valeurs de Shapley que pour les variables les plus influentes, en garantissant la robustesse des explications résultantes, fournissant ainsi des explications plus faciles à interpréter et visualiser (au prix toutefois d'une contrainte d'additivité des explications, non adaptable à tous les cas d'usage).
- La bibliothèque Shapash, créée par le Groupe MAIF, restitue des éléments explicatifs tels que l'importance des variables prédictives sous une forme visuelle et accessible à tous.
- La méthode Skope-rules, dont les concepteurs incluent le Groupe BPCE, vise à apprendre des règles de décision simples et logiques permettant d'optimiser la détection des zones d'entrée associées à une prédiction donnée en sortie.

Il est à noter que l'innovation de ces travaux concerne autant la construction des explications que la qualité de leur restitution (notamment via un critère de concision détaillé dans la section « Principe d'intelligibilité »).

Simplicité n'est pas synonyme de transparence

Le Tech Sprint a illustré que, lorsqu'ils sont exposés en boîte noire, des modèles à base de forêt aléatoire étaient parfois aussi difficilement explicables que des modèles plus complexes (modèles agrégés ou GBM).

Plutôt que la classe d'algorithme utilisée, la complexité des données d'entrée – notamment le nombre de variables prédictives – était le facteur-clef de la difficulté présentée par une boîte noire. Cette observation s'applique à la fois :

- À la génération d'explications. En effet, il est beaucoup plus difficile de sélectionner les variables pertinentes lorsque les facteurs prédictifs potentiels sont en nombre élevé. En particulier, l'effet d'une variable donnée varie alors souvent de façon non monotone entre deux explications locales – ou entre une explication locale et une explication globale –, conduisant à des explications généralement moins robustes.

- À la réception des explications. En effet, une prédiction impliquant une multitude de facteurs est plus difficile à motiver, surtout lorsque ces facteurs sont corrélés aux plans statistique et sémantique. L'établissement de liens de causalité est implicitement attendu par le destinataire de l'explication, ce qui est infaisable dans les cas de modèles de ML pur n'embarquant aucune contrainte de causalité.

4. Restitution des explications

Le Tech Sprint nécessitait donc de mettre en œuvre une ou plusieurs méthodes explicatives qui soient adaptées, tant au plan scientifique que de la faisabilité technique, au défi proposé d'audit en boîte noire. Mais la qualité de la restitution de ces travaux était un critère d'évaluation aussi important que la méthodologie, l'objectif final étant la bonne compréhension de l'explication par son ou ses destinataires supposés.

Cette section tente de résumer les moyens employés par les participants afin de restituer de façon adéquate les explications produites au cours de leurs travaux. Deux principes-clefs en ressortent :

- un souci de rendre les explications aussi intelligibles que possible en limitant la charge cognitive associée à leur assimilation ;
- la conception d'interfaces permettant d'interagir avec le destinataire des explications.

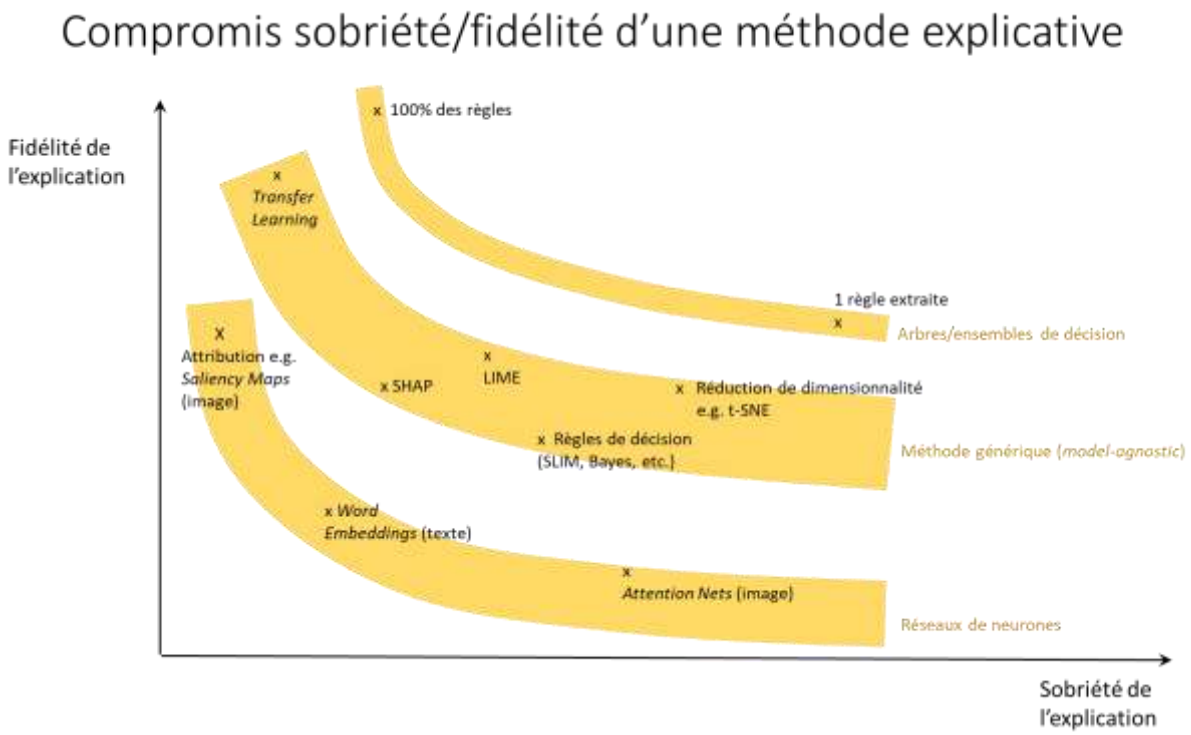
4.1. Principe d'intelligibilité

Le document de gouvernance de l'IA publié par l'ACPR précisait qu'une explication algorithmique doit s'avérer intelligible pour l'audience concernée, adaptée au cas d'usage, et proportionnée au risque porté par le processus. Cette section examine le premier critère, celui d'intelligibilité d'une explication par son destinataire.

Arbitrage sobriété / fidélité

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

Le diagramme suivant, extrait du document de 2020, schématise l'arbitrage sobriété/fidélité d'une explication selon le type d'algorithme de ML considéré et le type d'explication choisi. Y sont figurés quelques « couloirs » d'arbitrage illustrant que pour un même type d'algorithme, certaines méthodes explicatives vont dévier de la courbe de tendance générale.



4. Arbitrage entre sobriété et fidélité d'une méthode explicative

Degré de concision

La littérature des sciences cognitives concernant la production et la réception d'explications algorithmiques met au premier plan la prise en compte des limites cognitives du destinataire humain de ces explications. Parmi les principes couramment acceptés permettant cette prise en compte, figurent la limitation du nombre de « bloc cognitifs »[9] inclus dans l'explication et leur séquençage temporel, la visualisation des explications à la demande et leur passage en arrière-plan après un certain laps de temps.

La concision des éléments explicatifs présentés au destinataire à un instant donné est donc reconnue comme un élément-clef d'une bonne explication.

Or la concision d'une explication est, comme expliqué dans la section précédente, en tension avec l'objectif de fidélité au modèle ; d'autre part elle est notoirement difficile à mesurer car reposant sur la notion ambiguë de *chunk* cognitif. Un participant du Tech Sprint l'a exprimé ainsi : « La métrique de concision est à adapter à chaque cas d'usage, en fonction des parties prenantes. »

Les équipes participant au Tech Sprint se sont efforcées d'adapter la concision des explications au cas d'usage qui leur était présenté. La méthode généralement utilisée consistait à grouper les variables soit par catégorie sémantique soit par comportement :

[9] Le terme consacré par la littérature est « *cognitive chunk* »: un nombre de blocs cognitifs entre 3 et 5, en tout cas inférieur à 7, est préconisé (voir l'article fondateur de G. A. Miller : *"The magical number seven, plus or minus two: Some limits on our capacity for processing information"*). La définition d'un bloc cognitif dépend du contexte précis, mais correspond généralement à un *cluster* conceptuel, c'est-à-dire un groupe de concepts similaires entre eux et sémantiquement éloignés des autres concepts évoqués par l'explication.

- La catégorisation sémantique consiste par exemple à présenter ensemble les variables financières (revenus du foyer, mensualités de crédit, etc.), les variables sociodémographiques (situation maritale, situation professionnelle, etc.) et les variables de tenue de compte (date et fréquence des incidents de remboursement, etc.). Sa limitation est d'agréger des variables aux comportements potentiellement différents, ainsi qu'un découpage parfois arbitraire (dans l'exemple précédent, où classer la situation de logement ?)
- Le groupement de variables par comportement est donc souvent préférable, conduisant à regrouper des variables variant similairement à la hausse ou à la baisse – et qui s'avèrent d'ailleurs souvent sémantiquement liées (par exemple dans le cas d'usage du Tech Sprint, les variables concernant le ou les crédits immobiliers détenus par le client se trouvent fréquemment corrélées aux variables analogues concernant les crédits à la consommation).

Une autre méthode, singulièrement innovante, a tenté de concilier concision et robustesse : cette approche permet de mesurer la stabilité d'un groupe de variable, et donc d'inférer le petit groupe de variables « stable » au sens de maintenir la décision à un certain niveau de probabilité donné[10]. Le relâchement de contraintes indiqué par le seuil probabiliste conduit à diminuer le nombre de variables présentées dans une explication donnée en ne sacrifiant que celles à effet négligeable ou n'affectant qu'une petite minorité d'individus.

Combler le fossé entre explications locales et globales

Certains participants ont tenté de combler le fossé existant entre explications locales et explications globales, et de pallier ainsi leurs limitations respectives. En effet, les explications locales sont fidèles au modèle sous-jacent qu'elles tentent d'approximer ou dont elles révèlent les variables les plus importantes, mais toujours dans le voisinage d'un point d'entrée : une explication ne permet ainsi d'expliquer qu'une seule prédiction à la fois. Quant aux explications globales, elles tentent d'approximer le comportement général du modèle, de nouveau soit en construisant une approximation soit en indiquant les variables prépondérantes, ce qui conduit pour tous les modèles à l'exception des plus simples à une représentation peu fidèle de la réalité.

Trois approches distinctes ont permis à certaines équipes de définir des cadres explicatifs allant au-delà de cette dichotomie entre local et global : les *personae*, le *clustering* simple, et le *clustering* de valeurs de Shapley.

Les *personae*

Cette approche consiste à construire de toutes pièces un archétype, ou avatar représentatif d'un segment de population. Il ne s'agit pas forcément d'un individu réel (et unique), et ses caractéristiques sont généralement construites en combinant des critères statistiques (pour obtenir des *personae* aussi largement représentatives que possible) et des critères métier (pour ne pas regrouper des profils de clients n'ayant rien à voir entre eux).

L'intérêt de cette méthode est de proposer un petit nombre de profils de consommateurs réalistes, et d'expliquer le comportement du modèle sur chaque profil. La représentation mentale des destinataires de l'explication s'en trouve d'autant facilitée car on imagine le traitement réservé à un individu spécifique. En revanche, le rattachement d'un individu quelconque à une *persona* peut être approximatif, car ces *personae* sont définies en petit nombre : le comportement du modèle sur des individus éloignés de la *persona* considérée sera en fait différent de ce qu'indique l'analyse, d'où un

[10] C'est la méthode des *Active Coalitions of Variables*, utilisant une technique dite SDP (*Same Decision Probability*).

manque relatif de fidélité de cette méthode (idem pour l'explication concernant des points de données situés dans des zones de faible densité et donc généralement non pris en compte par la liste de *personae* disponibles).

Le *clustering* simple

Dans cette approche (la plus classique des trois), on regroupe les individus en fonction des similarités de leurs caractéristiques intrinsèques, par exemple les individus âgés de 30 à 45 ans et propriétaires de leur logement. La définition des classes (ou "clusters") en question est une propriété émergente des jeux de données (d'apprentissage et de test) visant à résumer les zones de forte densité de ces données, et non des critères définis à dire d'expert comme dans le cas des *personae*.

L'avantage de cette méthode par rapport à la construction de *personae* est que les groupes résultants reflètent la distribution statistique des données d'entrée : un groupe ou cluster ne contiendra a priori pas de points de données trop éloignés les uns des autres ; de plus on peut généralement associer un individu (réel et non archétypique comme pour les *personae*) représentatif dans chaque cluster. Inversement, les clusters peuvent être plus nombreux que les *personae* définies à dire d'expert, et les points de données représentatifs sont rarement les profils les plus pertinents du point de vue métier.

Clustering de valeurs de Shapley

Dans cette approche, on regroupe les individus dont les variables les plus sensibles sont similaires. Cela n'implique pas la similitude de leurs caractéristiques intrinsèques : par exemple les individus âgés de moins de 25 ans ou de plus de 60 ans, n'ayant aucun crédit en cours ou en ayant au moins 3, peuvent se retrouver groupés par cette méthode en raison de l'aspect déterminant des variables "âge" et "nombre de crédits en cours". Un outil dans cette catégorie, développé par une équipe du Tech Sprint, produit des explications dites « explications régionales ».

Cette méthode se distingue des autres en ce que les individus sont regroupés en fonction de leur traitement par le modèle, et non en fonction de caractéristiques intrinsèques qui peuvent n'avoir aucun rôle significatif dans le modèle prédictif. À l'inverse, la représentation mentale et la description des clusters est moins intuitive et plus complexe que pour les autres méthodes, car ils sont définis en termes d'importance plus ou moins grande de chaque variable, et non des valeurs de ces variables.

4.2. Principe d'interactivité

Si la section précédente « Principe d'intelligibilité » partait de la nécessaire prise en compte des limites cognitives du destinataire humain d'explications algorithmiques, un autre principe établi par les sciences cognitives consiste à reproduire ou accompagner le processus humain de formulation d'une explication.

Il s'agit alors d'adapter l'explication aux objectifs du destinataire, de sélectionner les facteurs prédictifs selon des critères proches des critères utilisés par les humains (notamment l'anomalie et l'intentionnalité), d'ajuster en mode itératif les éléments présentés en fonction de la réaction de l'utilisateur, et de privilégier les variables « actionnables ». Certains de ces points ont particulièrement été étudiés par les participants du Tech Sprint et font l'objet de cette section.

Caractère « actionnable »

Les travaux des participants ont également éclairé la question du caractère actionnable à prévoir pour les explications algorithmiques. Les variables actionnables sont ici les facteurs prédictifs qui peuvent être modifiés par un changement dans le comportement du consommateur de services financiers : par

exemple, le montant du prêt demandé peut être actionnable dans certains cas, mais pas pour un consommateur ayant un projet précis ; le revenu du foyer n'est pour sa part nullement actionnable.

Certaines équipes ont estimé important d'inclure des variables actionnables tant dans les explications destinées aux consommateurs eux-mêmes que dans celles utilisées par les conseillers clientèle. À l'inverse, d'autres équipes ont mis en évidence des situations nécessitant l'inclusion de variables non actionnables : si par exemple une demande de prêt a été rejetée au moins partiellement en raison de l'âge du demandeur, cela doit être communiqué de façon transparente étant donné l'interdiction de prendre des décisions d'octroi de crédit basées systématiquement et uniquement sur l'âge.

Contextualisation

Enfin, les travaux du Tech Sprint ont souligné qu'une explication n'est pas une information autoporteuse qui pourrait être interprétée sans contexte additionnel. Par exemple, les visualisations devraient être accompagnées de connaissance métier. Cela peut inclure une description de la sémantique de chaque variable prédictive – ce qu'elle signifie dans la réalité et pas juste dans le modèle. L'information concernant le modèle lui-même peut aussi s'avérer pertinente dans certains cas, par exemple pour expliciter la différence entre un taux d'intérêt moyen et le taux d'intérêt accordé au consommateur.

« Dialogue » explicatif

Les participants ont exploré par différents moyens le processus fondamentalement interactif de réception d'une explication.

Certains ont proposé des fonctions de navigation à l'intérieur d'une explication : visualisation des variables importantes « à la demande » (en cliquant sur une variable, une visualisation détaillée de sa distribution et de ses effets est affichée), zoom sur une sous-population ou même sur un individu pour voir les explications locales de sa prédiction, filtrage interactif du nombre et du type de variables (par exemple pour retenir ou non les variables non-actionnables), etc.

Une équipe a prévu dans son mécanisme explicatif une *feedback loop* (boucle de rétroaction) afin d'améliorer les explications en continu au cours de leur utilisation, notamment en mettant à jour les concepts intégrés dans les règles de décision présentées au destinataire.

5. Sujets de travaux futurs

L'ACPR a choisi d'axer son premier Tech Sprint autour de l'un des principes-clefs de la gouvernance de l'IA préalablement identifiés, à savoir l'explicabilité des modèles prédictifs.

Parmi les potentiels futurs travaux de l'ACPR liés à l'explicabilité de l'IA, on peut citer les suivants :

- L'ACPR pourrait poursuivre ses travaux sur l'explicabilité de l'IA en se focalisant sur d'autres principes de conception et de gouvernance identifiés comme essentiels mais dont l'étude est encore largement limitée au plan théorique, tels que l'équité algorithmique.
- Le Pôle Fintech-innovation étudie par ailleurs les interactions entre opérateurs humains et algorithmes d'IA, tant sous un angle académique (revue de littérature sur les facteurs humains en jeu dans la réception d'une explication algorithmique) que sous un angle pratique (étude expérimentale sur les *robo-advisors* en assurance-vie).

Par ailleurs, le Pôle Fintech-innovation étudie également l'audit d'algorithmes à base d'IA en général. En particulier, la question se pose du processus d'évaluation adéquat : sur la base de bonnes pratiques ou d'exigences réglementaires de haut niveau, l'analyse d'un cas d'usage concret permettrait de déterminer la méthodologie et l'outillage les plus appropriés à cette évaluation. Un point crucial concerne les interactions entre humain et algorithme : comme l'ont souligné le document de gouvernance de l'ACPR mais aussi le projet de réglementation sur l'IA de la Commission européenne, l'humain doit rester au premier plan de l'opération d'un système à base d'IA, d'où la nécessité d'inclure ses actions et son comportement dans le protocole d'évaluation.

On notera enfin que, dans la suite du Tech Sprint, l'une des équipes lauréates travaille à étendre ses travaux réalisés lors de l'événement en construisant une application de démonstration des méthodes utilisées, opérant sur des jeux de données utilisés lors du Tech Sprint et au-delà. Cette prolongation des effets du Tech Sprint de l'ACPR sur les travaux d'acteurs de la place est un signe positif, qui confirme l'intérêt de tels exercices pour promouvoir les échanges et la collaboration entre acteurs du secteur, sous l'égide de l'ACPR.

Annexe : exemples d'explications produites

Cette section présente quelques exemples d'explications réalisées par les participants du Tech Sprint. Pour chaque profil de destinataire concerné, un exemple est donné dans le cas d'une approche par modèle de substitution, et un exemple dans le cas d'une méthode post-hoc (de type résumé de *features*, éventuellement combinée à une autre technique), ces deux catégories constituant la majorité des méthodes mises en œuvre par les participants.

Pour le data scientist

Objectifs : comprendre la nature du modèle, ses performances, le traitement des valeurs manquantes.

Explications par modèle de substitution : arbre relationnel complet.

Explications par méthode post-hoc : SDP local avec probabilité des estimations + valeurs de Shapley avec incertitudes.

Pour l'expert métier

Objectifs : comprendre ce qui influence la décision, aider le client à éviter de faire défaut.

Explications par modèle de substitution : les règles les plus sûres.

Explications par méthode post-hoc : règles métier générales (*Skope rules*) + valeurs de Shapley locales.

Pour l'auditeur

Objectifs : catégoriser les clients, comprendre l'ensemble du processus et l'objectif du modèle, ses limites, comprendre si le modèle fait ce pour quoi il est conçu.

Explications par modèle de substitution : difficile à définir, encore plus à entraîner (surtout en audit externe).

Explications par méthode post-hoc : explications globales + focus sur les régions de « stress ».

Pour le consommateur

Objectifs : comprendre pourquoi la décision a été prise, comment son comportement influe sur l'algorithme, etc.

Explications par modèle de substitution : explication en langage naturel sur la base des règles les plus sûres. Par exemple : « L'avis est favorable car vous souhaitez racheter ou reprendre un prêt immobilier résidentiel et aucune écriture n'a été écartée sur votre CAV pour les 3 derniers mois. »

Explications par méthode post-hoc : exemples contrefactuels + importance locale des variables.