

EXPLICABILITÉ DES MODÈLES EN ACTUARIAT POUR UN USAGE RESPONSABLE DU MACHINE LEARNING

MOTS CLES : XAI, SHAP, MACHINE LEARNING, ACTUARIAT, DATA SCIENCE, CONFORMITE

Entre exigence déontologique et impératif de gestion des risques, l'explicabilité des modèles constitue un enjeu central pour les actuaires. La démocratisation des techniques de *machine learning*, facilitée notamment par l'essor des chatbots basés sur de grands modèles de langage, devenus d'excellents assistants à la programmation, ainsi que l'évolution des systèmes d'information au sein des organismes d'assurance, ont conduit à l'utilisation de modèles toujours plus complexes. Ceux-ci exploitent à la fois la quantité importante de données internes et les sources externes (open data), dans le but d'affiner les tarifications, d'identifier des classes d'assurés particulièrement à risque ou encore de réaliser des analyses prédictives plus précises.

Cependant, les organismes d'assurance doivent impérativement faire preuve de transparence dans leurs processus décisionnels. Cette transparence est essentielle pour maintenir la confiance des assurés et satisfaire aux exigences des régulateurs. Or, la complexité croissante des modèles statistiques, souvent qualifiés de « boîtes noires », constitue un obstacle majeur à cette transparence. Il devient donc crucial de disposer d'outils permettant de comprendre et d'interpréter la logique sous-jacente à ces modèles.

Cet article, rédigé par le DataLab de GALEA, présente l'utilisation des valeurs de Shapley comme levier pour renforcer l'explicabilité des modèles dans une démarche de transparence, de gestion du risque et de conformité réglementaire.

L'EXPLICABILITÉ : UN ENJEU ACTUARIEL MULTIPLE

Les modèles d'apprentissage statistique (ou *machine learning*, ML) sont devenus des outils incontournables pour l'actuaire. À ce titre, les enjeux associés à leur utilisation croissent de nombreuses exigences propres à la fonction actuarielle.

L'explicabilité des modèles apparaît ainsi comme un élément central : elle constitue à la fois un levier de gestion du risque, en particulier du risque de modèle, une exigence déontologique liée à la maîtrise des outils et méthodes employés, un facteur clé de confiance vis-à-vis des assurés, des équipes dirigeantes et des autorités de supervision, ainsi qu'un impératif de conformité au regard des nouvelles régulations encadrant les systèmes d'intelligence artificielle.

UNE NÉCESSAIRE ADAPTATION DE LA GESTION DU RISQUE DE MODÈLE

Les techniques de ML et, plus largement, les outils d'intelligence artificielle (IA) amplifient des risques déjà existants, dans la mesure où les conséquences négatives potentielles d'un modèle augmentent avec sa complexité. Ces conséquences négatives peuvent découler d'erreurs de conception, de performances médiocres ou d'un usage inapproprié. Le risque de modèle peut se définir comme étant le risque d'obtenir des prévisions ou réponses erronées conduisant soit à de mauvaises décisions (ou

non-optimales), soit à des erreurs d'évaluations financières, comptables ou réglementaires. Le processus de gestion du risque de modèle doit intégrer la spécificité des modèles de ML. Selon l'*International Association of Insurance Supervisors* (IAIS), les pratiques des superviseurs relatives à l'IA tendent à compléter les exigences existantes sur le risque de modèle en mettant davantage l'accent notamment sur la fiabilité, l'explicabilité, la transparence et l'équité.

De manière concrète, les modifications pouvant être apportées au processus de gestion du risque de modèle sont par exemple :

- // La définition claire de l'objectif lors de la phase d'initialisation, du périmètre et des exigences du modèle d'IA, notamment réglementaires.
- // Un panel de tests élargi lors de la phase de conception pour assurer la qualité et la pertinence des données, pour prendre en compte la sensibilité aux hyperparamètres, et pour mesurer la performance par rapport à d'autres modèles.
- // Un respect de l'approche train-validation-test pour la phase de validation du modèle.
- // Une gestion adaptée de l'environnement de travail, des bibliothèques et de leur version, ainsi qu'une documentation dans le respect des conventions du langage de programmation.
- // Enfin, l'exploitation du modèle nécessite un suivi en temps réel afin de détecter s'il les performances du modèle dévient et si

l'alignement obtenu pendant
l'entraînement reste stable.

LA DÉONTOLOGIE ACTUARIELLE À L'HEURE DE L'IA

Selon le code de déontologie de l'Institut des Actuaires, l'actuaire se doit de « s'assurer de la bonne compréhension des outils et méthodes utilisés pour établir les résultats », résultats qui « engage[nt] sa responsabilité » et dont la « sensibilité [...] aux hypothèses et aux choix de modélisation » doit être appréciée.

Dans ce cadre, l'explicabilité des modèles d'intelligence artificielle devient un enjeu déontologique structurant. En effet, les modèles d'IA, et en particulier ceux basés sur des techniques complexes de ML, sont souvent perçus comme des « boîtes noires » en raison de la difficulté à interpréter leur fonctionnement interne. Si l'actuaire utilise de tels outils sans en comprendre les mécanismes, il contrevient à son devoir fondamental de maîtrise des méthodes employées. En effet, la responsabilité de l'actuaire ne se limite pas à produire des résultats fiables : il doit aussi être capable d'expliquer les fondements de ces résultats, d'en analyser la sensibilité, et de rendre compte de manière transparente.

LA CONFIANCE AVEC TOUTES LES PARTIES PRENANTES

L'explicabilité des modèles d'IA permet d'éviter les deux écueils identifiés par Denis Beau, premier sous-gouverneur de la Banque de France : la défiance systématique ou la confiance aveugle envers la machine. Ce critère de confiance dépend de la position de chaque partie prenante vis-à-vis de la décision prise par le système d'IA. Pour l'assuré, il s'agit de comprendre que la décision assurantielle, qu'elle porte sur la tarification ou l'indemnisation, repose sur des bases solides et justifiables. Pour les actuaires, l'explicabilité facilite la validation des modèles en permettant une meilleure analyse des hypothèses et des résultats. Les instances dirigeantes, quant à elles, ne pourront autoriser la mise en production de systèmes d'IA sans s'assurer d'un niveau suffisant de confiance, notamment en ce qui concerne la stabilité financière et la cohérence avec les valeurs de l'entreprise. Enfin, les organes de supervision évalueront la conformité réglementaire des modèles, en particulier sur les risques de discrimination et le respect des exigences prudentielles. Les outils d'explicabilité sont donc essentiels pour répondre aux attentes de l'ensemble de ces parties-prenantes.

LA CONFORMITÉ DANS UN CADRE RÉGLEMENTAIRE DYNAMIQUE

L'usage de l'intelligence artificielle dans le secteur de l'assurance est avant tout encadré par le droit européen et français en matière de protection des données personnelles. L'article 22 du Règlement général sur la protection des données (RGPD) prévoit le droit pour toute personne de ne pas faire l'objet d'une décision individuelle automatisée produisant des effets juridiques ou significatifs (par exemple : un refus d'assurance ou une augmentation de prime) sans intervention humaine. Dans de tels cas, l'assuré dispose également d'un droit à l'explication portant sur les critères ayant conduit à cette décision.

Au niveau européen, le futur règlement sur l'intelligence artificielle (AI Act), en cours d'adoption, instaure un cadre de régulation fondé sur une classification des systèmes d'IA selon leur niveau de risque. Certains usages seront interdits, tandis que les « systèmes à haut risque » seront soumis à des obligations strictes (exigences de transparence, documentation, contrôle humain, etc.). Seront notamment considérés comme « à haut risque » les systèmes d'IA utilisés pour évaluer les risques ou établir les tarifs en assurance-vie et assurance santé, lorsqu'ils concernent des personnes physiques.

En France, l'Autorité de contrôle prudentiel et de résolution (ACPR) prépare activement la supervision de ces systèmes. Elle prévoit que les assureurs devront mettre en place une gouvernance rigoureuse des algorithmes d'IA, s'inspirant des exigences applicables aux modèles internes dans le cadre de Solvabilité II, tout en tenant compte des spécificités de l'IA. Ses travaux soulignent l'importance de l'évaluation approfondie des modèles, de la qualité des données, de la détection de biais discriminatoires, ainsi que de l'exigence d'explicabilité, indispensable à la maîtrise et à la supervision des décisions algorithmiques.

QUELLE EXPLICABILITÉ ?

L'explicabilité désigne la capacité à rendre compréhensible le fonctionnement et les décisions d'un algorithme. Elle vise à fournir des éléments techniques permettant d'analyser, d'évaluer et de justifier les résultats produits par un modèle. Elle est intimement liée à l'interprétabilité, qui renvoie davantage à la compréhension par un public non expert, et s'appuie sur des notions connexes telles que la transparence (accès aux mécanismes internes) et l'auditabilité (capacité à évaluer empiriquement et analytiquement le système).

OBJECTIFS

Selon l'ACPRⁱ, rendre un modèle explicable a pour objectif de pouvoir répondre aux questions suivantes :

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

Autrement dit, la question de l'explicabilité des algorithmes concerne à la fois les raisons de leurs décisions (« pourquoi ») et les mécanismes qui les sous-tendent (« comment »).

MESURER L'EXPLICABILITÉ

Pour répondre à ces questions l'explication doit idéalement être : précise, complète, compréhensible, succincte, actionnable, robuste, et réutilisable. Les acteurs devront arbitrer entre performance et explicabilité, deux exigences qui peuvent s'avérer difficiles à concilier en pratique. Ce compromis est d'autant plus complexe que la définition de métriques appropriées pour évaluer chacune de qualités citées ci-dessus reste un sujet ouvert, malgré l'existence de premières propositions.

Une échelle à 4 paliers a été proposée par l'ACPR pour classifier les modèles selon leur niveau d'explicabilité :

- // Le premier niveau correspond au niveau d'observation : il est possible de répondre techniquement à la question « Que fait l'algorithme ? »
- // Le deuxième niveau est celui de la justification : ce niveau est atteint quand une réponse peut être apportée à la question : « Pourquoi l'algorithme donne-t-il tel résultat (en général ou dans une situation précise) ? »
- // Le troisième niveau est le niveau de l'approximation : il est possible de fournir une réponse, à la question : « Comment fonctionne l'algorithme ? ».
- // Le quatrième et dernier niveau est celui de la réplication : à ce stade, l'algorithme est complètement maîtrisé, il est possible de prouver que l'algorithme fonctionne correctement.

VISION GLOBALE ET LOCALE DE L'EXPLICABILITÉ

Deux approches complémentaires peuvent être mobilisées pour répondre au second

niveau d'explication. L'explicabilité globale vise à fournir une vue d'ensemble du fonctionnement du modèle : elle décrit les principes de l'algorithme, son mode d'utilisation, ainsi que ses limites, liées à son architecture, à sa phase d'entraînement ou à la nature des données utilisées. L'explicabilité locale, quant à elle, s'attache à expliquer une décision particulière prise par le modèle, par exemple : « Pourquoi l'individu i se voit attribuer le tarif y_i ? ». Ces deux approches sont indispensables autant que complémentaires et doivent donc être envisagées conjointement.

Des techniques et outils issus du monde académique ont été développés pour répondre aux différents niveaux d'explicabilité, à l'échelle globale et à l'échelle locale. En effet, l'explicabilité des algorithmes d'intelligence artificielle, désignée sous le terme d'XAI (*eXplainable AI*), constitue depuis quelques années un domaine de recherche riche et en plein essor, dans lequel les acteurs peuvent puiser pour faire face aux nombreux enjeux auxquels ils sont confrontés.

SHAP : UNE SOLUTION ROBUSTE ET EFFICACE

Parmi les méthodes XAI les plus utilisées figure en bonne place la méthode SHAP (*SHapley Additive exPlanation*) introduite par Lundberg et Lee en 2017ⁱⁱⁱ. Cette méthode repose sur les valeurs de Shapley outils décrits en 1953 par Shapley dans le contexte de la théorie des jeux coopératifs. L'idée générale est de pouvoir mesurer le mérite d'un joueur donné dans un ensemble de joueurs. La méthode mesure donc la contribution d'un joueur au résultat final en tenant compte de ses interactions avec les autres joueurs. La méthode SHAP est :

- // Agnostique : elle peut être utilisée quel que soit le modèle d'apprentissage ;
- // Post-hoc : elle est calibrée sur des modèles déjà entraînés ;
- // Locale : elle est calculée au niveau individuel de chacune des prédictions.

LE FONCTIONNEMENT DE SHAP

Concrètement, SHAP calcule pour chaque variable l'impact moyen (marginal) sur la prédiction en comparant les sorties du modèle avec et sans cette variable dans toutes les combinaisons possibles de variables. Pour un cas avec 4 variables X_1, X_2, X_3, X_4 , l'ensemble des combinaisons S possible pour étudier l'impact de X_1 est :

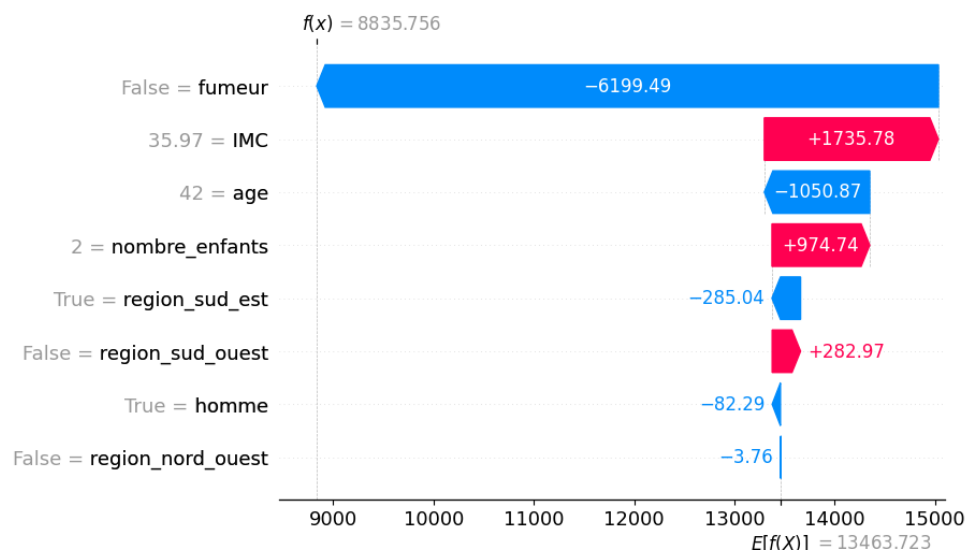


Figure 1 Waterfall plot : Explication locale de la prédiction de l'algorithme pour un assuré donné

$\{\{X_2\}, \{X_3\}, \{X_4\}, \{X_2, X_3\}, \{X_2, X_4\}, \{X_3, X_4\}, \{X_2, X_3, X_4\}\}$.

Pour chacune des combinaisons $s \in S$, la différence est calculée entre la prédiction $f_{s \cup X_1}$ basée sur les variables de s et de X_1 et la prédiction f_s basée seulement sur les variables de s . Cette différence correspond à une contribution marginale

En moyennant ces différences sur S avec une pondération relative au cardinal de chaque s , on obtient la valeur de Shapley pour X_1 , c'est-à-dire sa contribution à la prédiction relativement aux autres variables.

Finalement, la somme des valeurs de Shapley de toutes les variables égalise la prédiction du modèle relative à une valeur de référence (la sortie moyenne du modèle). Cela garantit une attribution additive et cohérente (axiome d'efficacité) et une répartition « juste » des contributions entre variables.

La méthode SHAP repose sur les principes mathématiques rigoureux des valeurs de Shapley, ce qui lui confère une base théorique solide. Les propriétés associées assurent par exemple qu'une variable n'ayant aucun impact sur la prédiction se verra attribuer une contribution nulle, ou que deux variables identiques recevront le même poids. Ces garanties confèrent de la « fiabilité » aux explications. De plus, la somme des SHAP values correspond toujours exactement à la prédiction du modèle, assurant une cohérence globale.

Et bien que la complexité computationnelle du calcul naïf soit une fonction exponentielle du nombre de variables, il existe des solutions efficaces qui garantissent l'exactitude des explications pour les modèles d'arbres (*RandomForest*, *XGBoost*, *LightGBM*, etc.).

SHAP EN PRATIQUE

Concernant les modèles d'IA en assurance, dans un cadre d'apprentissage supervisé, les variables explicatives sont assimilées aux joueurs et le résultat du jeu correspond à la prédiction de l'algorithme. La méthode SHAP calcule donc un score pour chaque variable explicative qui permet de comprendre comment chaque variable x_i pour l'individu i a participé à la décision \hat{y}_i , estimation par le modèle prédictif de y_i .

L'application proposée dans le présent article porte sur le jeu de données « *insurance* » du livre « *Machine Learning with R* » de Brett Lantz dont les données sont en accès libres sur github^{iv}. Le jeu de données contient 1338 lignes, chacune correspondant à un assuré sinistré sur un contrat frais de santé aux Etats-Unis, et de 7 colonnes : l'âge, le sexe, l'indice de masse corporel (IMC), le nombre d'enfant(s), l'information fumeur/non-fumeur, la région, et le coût annuel pour l'assureur.

L'objectif est de prédire ce coût en fonction des caractéristiques de l'assuré. Après un traitement rapide des données, un modèle de *Gradient Boosting* est entraîné. Le modèle SHAP est ensuite appliqué à ce modèle.

La Figure 1 présente les résultats issus de la méthode SHAP, obtenus à l'aide de la bibliothèque Python du même nom. Le profil analysé correspond à un assuré de 42 ans, résidant dans le sud-est, père de deux enfants, non-fumeur, avec un IMC de 35,97. La prédiction moyenne du modèle est $E[f(X)] = 13463$. Le graphique permet de visualiser la décomposition de la prédiction individuelle $f(x_i)$ en contributions additives de chaque

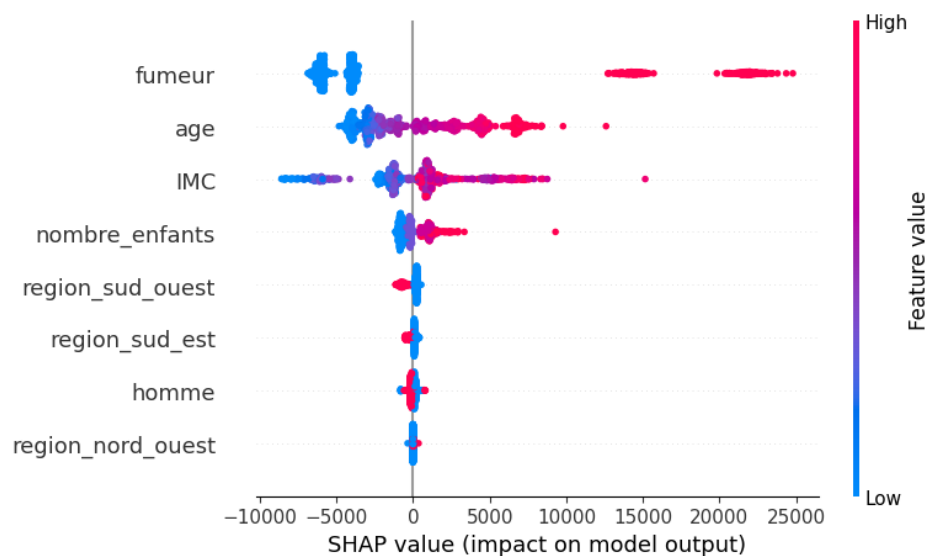


Figure 2 Summary plot : Explication globale du modèle par la représentation de l'ensemble des valeurs de Shapley pour chacune des variables

variable explicative, représentées par des flèches horizontales colorées : en rouge lorsqu'une variable augmente la prédiction, en bleu lorsqu'elle la diminue. Le graphique se lit de bas en haut. Le point de départ est la prédiction moyenne égale à 13463. Ne pas habiter dans le Nord-Ouest (*region_nord_ouest = False*) contribue à réduire la prédiction de 4. Le fait d'être un homme réduit de 82. Ne pas habiter dans le Sud-Ouest augmente la prédiction de 975. Et ainsi de suite pour arriver à une prédiction individuelle de 8836. Parmi les caractéristiques, le statut de non-fumeur ressort comme le facteur ayant l'impact le plus important, ce que traduit la longueur de sa flèche. Cette variable contribue de manière significative à une réduction de la prédiction : le modèle apprend ici qu'être non-fumeur est associé à des coûts de santé moindres.

Une analyse globale peut être réalisée en agrégeant les explications locales. La Figure 2 illustre cette démarche en regroupant l'ensemble des valeurs de Shapley calculées dans un graphique synthétique, où chaque point représente un assuré. Les variables explicatives sont classées de haut en bas selon leur importance décroissante dans les prédictions du modèle. L'effet de la variable « fumeur/non-fumeur » est très explicite, en effet la séparation est nette entre les points rouges représentant les assurés fumeurs et les points bleus représentant les non-fumeurs. Être fumeur entraîne systématiquement une augmentation de la prédiction, traduisant un surcoût de santé capté par le modèle. L'effet de l'âge est également bien identifiable : plus les points sont rouges (indiquant des âges plus élevés), plus leur contribution à la hausse de la prédiction est importante. La variable

« sexe » a un effet relativement limité, en effet les points sont centrés autour de zéro. Les hommes sont en rouge et les femmes en bleu.

La bibliothèque SHAP permet d'approfondir l'analyse en explorant les effets d'interaction entre les variables. La Figure 3 illustre, par exemple, que l'impact d'un IMC élevé (points rouges) sur la prédiction est d'autant plus marqué chez les assurés fumeurs (fumeur = 1).

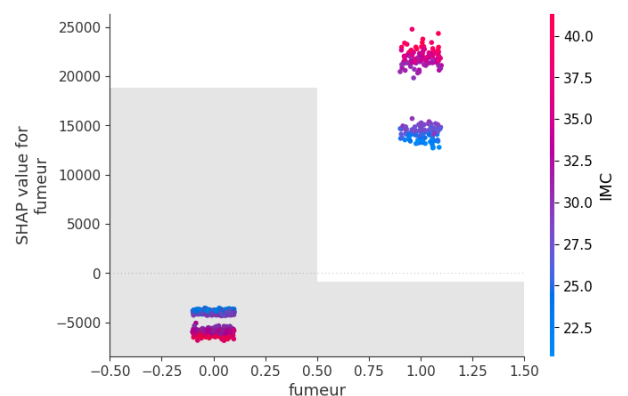


Figure 3 Scatter plot : Illustration de l'effet croisé de deux variables

LES ATOUTS DE SHAP

SHAP se révèle donc être un outil :

- // Facilement interprétable : la méthode produit des scores explicites (positifs ou négatifs) pour chaque variable, ce qui rend les explications intuitives, facilement visualisables.
- // Cohérent : elle permet d'interpréter les influences globales (importance des variables, effets principaux) et locales

- (effets individuels) dans un cadre unique.
- // Mathématiquement robuste : avec des garanties mathématiques.

Cette méthodologie peut être utilisée dans un cadre assurantiel tout en respectant les exigences RGPD d'accéder à des « informations pertinentes sur la logique sous-jacente »^v des traitements automatisés concernant les assurés, en traduisant par exemple chaque valeur SHAP en phrase simple : « votre prime est augmentée de X€ à cause de l'option A et réduite de Y€ grâce au profil B ». De même, face à l'audit interne et aux instances de supervision, les actuaires disposent avec les résultats SHAP d'un outil permettant de justifier la conformité du modèle aux règles métiers et d'attester de l'absence de discrimination.

Dans le cadre de la conformité à l'IA Act, SHAP sera un outil-clé pour répondre aux obligations réglementaires. En effet, SHAP permet de documenter la traçabilité de la prise de décision en documentant comment chaque variable entre en compte, et de produire des rapports compatibles avec les exigences réglementaires.

CONCLUSION

La conformité s'étend désormais aux modèles eux-mêmes, plaçant l'actuaire au cœur des enjeux de transparence et d'interprétabilité. Il lui revient d'être le garant de la clarté des modèles, dans un contexte où la responsabilité des dirigeants peut être engagée en cas de non-conformité. En rendant chaque prédiction explicable, les actuaires renforcent la confiance dans leurs travaux et répondent aux exigences déontologiques, techniques et réglementaires.

L'outil SHAP constitue, à cet égard, un levier précieux pour l'audit et la gestion des risques : il offre une lecture claire du comportement du modèle, enrichit la compréhension du risque modélisé et facilite l'identification d'éventuelles anomalies ainsi que l'alignement avec les attentes métiers.

Les consultantes et consultants du DataLab de GALEA sont à votre disposition pour répondre à vos questions sur cette note et vous aider à mettre en place des systèmes d'IA de confiance performants dans le respect du cadre réglementaire en vigueur. Nous vous accompagnons également dans le cadre de formation sur l'IA et la data science en assurance.



<https://www.galea-associes.eu/>

Pour aller plus loin :

Explicabilité en Intelligence Artificielle ; vers une IA Responsable :
<https://hal.science/hal-03936135>

SHAP for Actuaries: Explain any Model :
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389797

ⁱ <https://www.iais.org/uploads/2023/12/Regulation-and-supervision-of-AI-ML-a-thematic-review.pdf>

ⁱⁱ https://acpr.banque-france.fr/system/files/2025-02/20200612_gouvernance_evaluation_ia.pdf

iii

<https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>

iv <https://github.com/stedy/Machine-Learning-with-R-datasets>

v https://www.edpb.europa.eu/sme-data-protection-guide/respect-individuals-rights_fr