Interprétabilité du modèle d'apprentissage automatique avec AWS

AWS Conseils prescriptifs

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Conseils prescriptifs: Interprétabilité du modèle d'apprentissage automatique avec AWS

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Les marques et la présentation commerciale d’Amazon ne peuvent être utilisées en relation avec un produit ou un service qui n’est pas d’Amazon, d’une manière susceptible de créer une confusion parmi les clients, ou d’une manière qui dénigre ou discrédite Amazon. Toutes les autres marques commerciales qui ne sont pas la propriété d’Amazon appartiennent à leurs propriétaires respectifs, qui peuvent ou non être affiliés ou connectés à Amazon, ou sponsorisés par Amazon.

Table of Contents

Introduction 1
Résultats commerciaux ciblés 1
Présentation de 3
Bonnes pratiques 6
Interprétabilité locale 7
Pour les modèles d'ensembles d'arbres, utilisez tree SHAP 7
Pour les réseaux neuronaux et les modèles différentiables, utilisez des gradients et une conductance intégrés 7
Dans tous les autres cas, utilisez Kernel SHAP 7
Visualisations recommandées 8
Interprétabilité globale 10
Interprétabilité sur AWS 12
FAQ 13
Il existe de nombreuses méthodes pour déterminer l'importance des fonctionnalités qui ne sont pas abordées ici. Pourquoi ne sont-ils pas mentionnés ? 13
Quels sont les points faibles des méthodes préconisées ? 13
Ressources 14
Historique du document 16
..... xvii

Interprétabilité du modèle d'apprentissage automatique avec AWS

Adewale Akinfaderin, Matthew Chasse, Michele Donini et Benjamin Fenker, Amazon Web Services ()AWS

Février 2022 (historique du document)

Il est plus facile pour les utilisateurs finaux d'utiliser des algorithmes d'apprentissage automatique de manière responsable lorsqu'ils peuvent comprendre pourquoi un modèle fait une prédiction spécifique. Pour les développeurs de modèles, une meilleure compréhension de la manière dont un modèle fait des prédictions peut faciliter l'ingénierie et la sélection des fonctionnalités. Il n'existe pas de définition standard de ce que signifie expliquer un modèle, sauf qu'une explication doit être une condition préalable à des normes telles que la confiance, la robustesse, la causalité, l'informativité, la transférabilité du modèle et une prise de décision juste et éthique. Il existe des méthodes courantes pour générer des interprétations, mais elles présentent des faiblesses et des forces différentes. Cela n'est pas surprenant : en général, l'heuristique ou l'ensemble d'hypothèses simplificatrices que vous utilisez pour interpréter un modèle complexe peut simultanément être une source d'imprécision pour l'interprétation.

Ce guide fournit des conseils généraux sur les méthodes d'interprétabilité des modèles pour les praticiens de l'apprentissage automatique. Par souci de concision, le guide omet de nombreux détails et spécificités de mise en œuvre, et fournit des références pour vous aider à étudier des cas d'utilisation spécifiques de manière plus approfondie.

Résultats commerciaux ciblés

Dans certains cas, les réglementations telles que celles des secteurs de la santé et de la finance exigent l'interprétabilité du modèle en tant que résultat commercial souhaité. Les interprétations des modèles fournissent également des informations supplémentaires que les développeurs de modèles et les utilisateurs peuvent utiliser. Les autres résultats commerciaux ciblés liés à l'utilisation de l'interprétabilité des modèles sont les suivants :

- Justifiez les décisions importantes (par exemple, dans les domaines de la santé et de la finance) qui ont une incidence sur le bien-être des clients lorsque l'équité est essentielle.
- Contrôlez les inexactitudes et les distorsions des modèles lors de la prise de décisions commerciales.

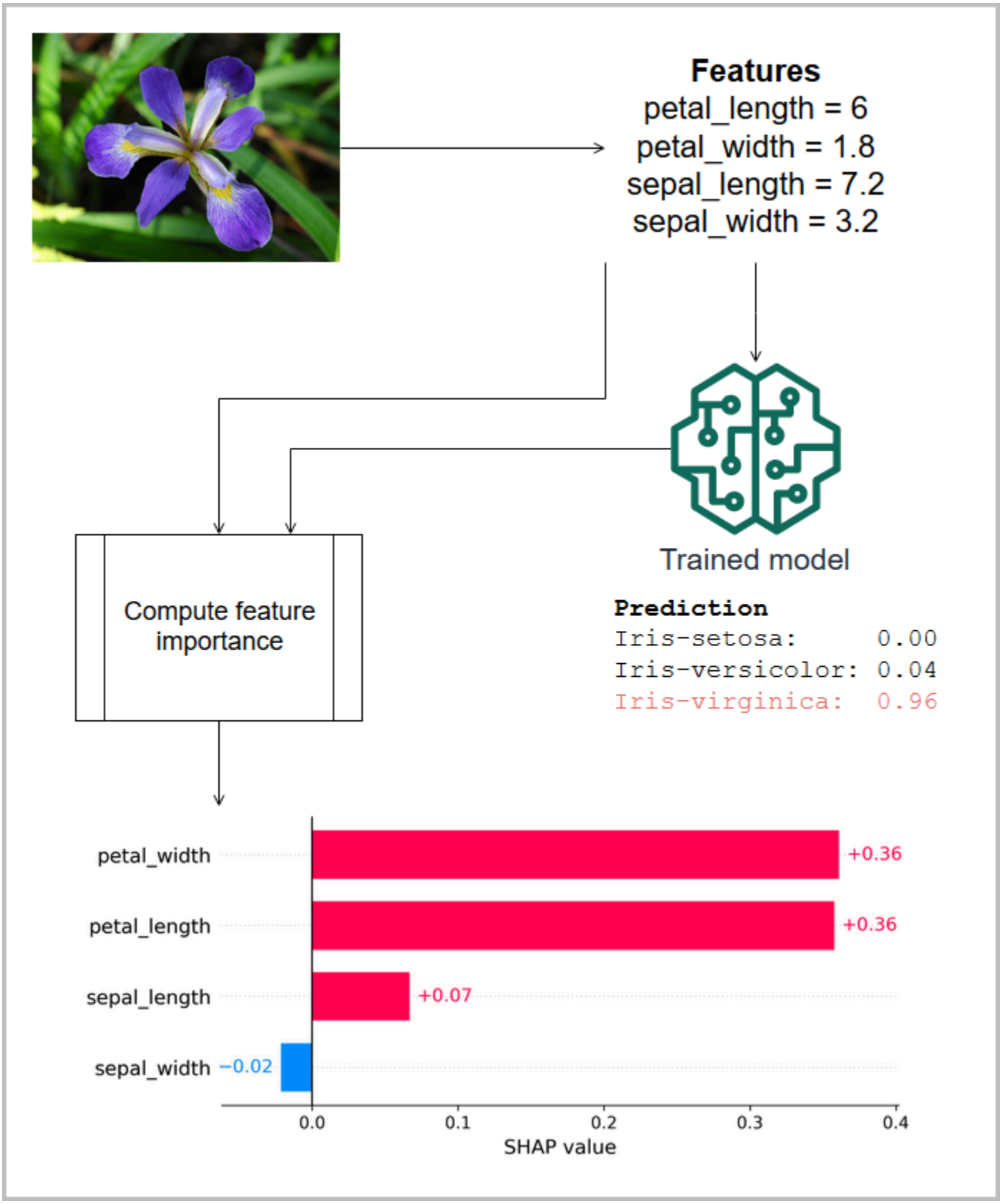
- Améliorez et accélérez le développement de modèles et l'ingénierie des fonctionnalités lorsque les interprétations des modèles sont utilisées par les data scientists.
- Découvrez les raisons des comportements généraux des modèles et fournissez de nouvelles informations sur les données et le modèle.

Ces résultats commerciaux sont directement liés aux quatre raisons d'explicabilité identifiées dans [1].

Présentation de

Il n'existe pas de définition universellement acceptée de ce qu'est un modèle interprétable ou de l'information adéquate en tant qu'interprétation d'un modèle. Ce guide se concentre sur la notion couramment utilisée d'importance des caractéristiques, selon laquelle un score d'importance pour chaque caractéristique en entrée est utilisé pour interpréter la façon dont elle affecte les sorties du modèle. Cette méthode fournit des informations mais nécessite également de la prudence. Les scores d'importance des fonctionnalités peuvent être trompeurs et doivent être analysés avec soin, y compris une validation par des experts en la matière si possible. Plus précisément, nous vous conseillons de ne pas vous fier aux scores d'importance des fonctionnalités sans vérification, car des interprétations erronées peuvent mener à de mauvaises décisions commerciales.

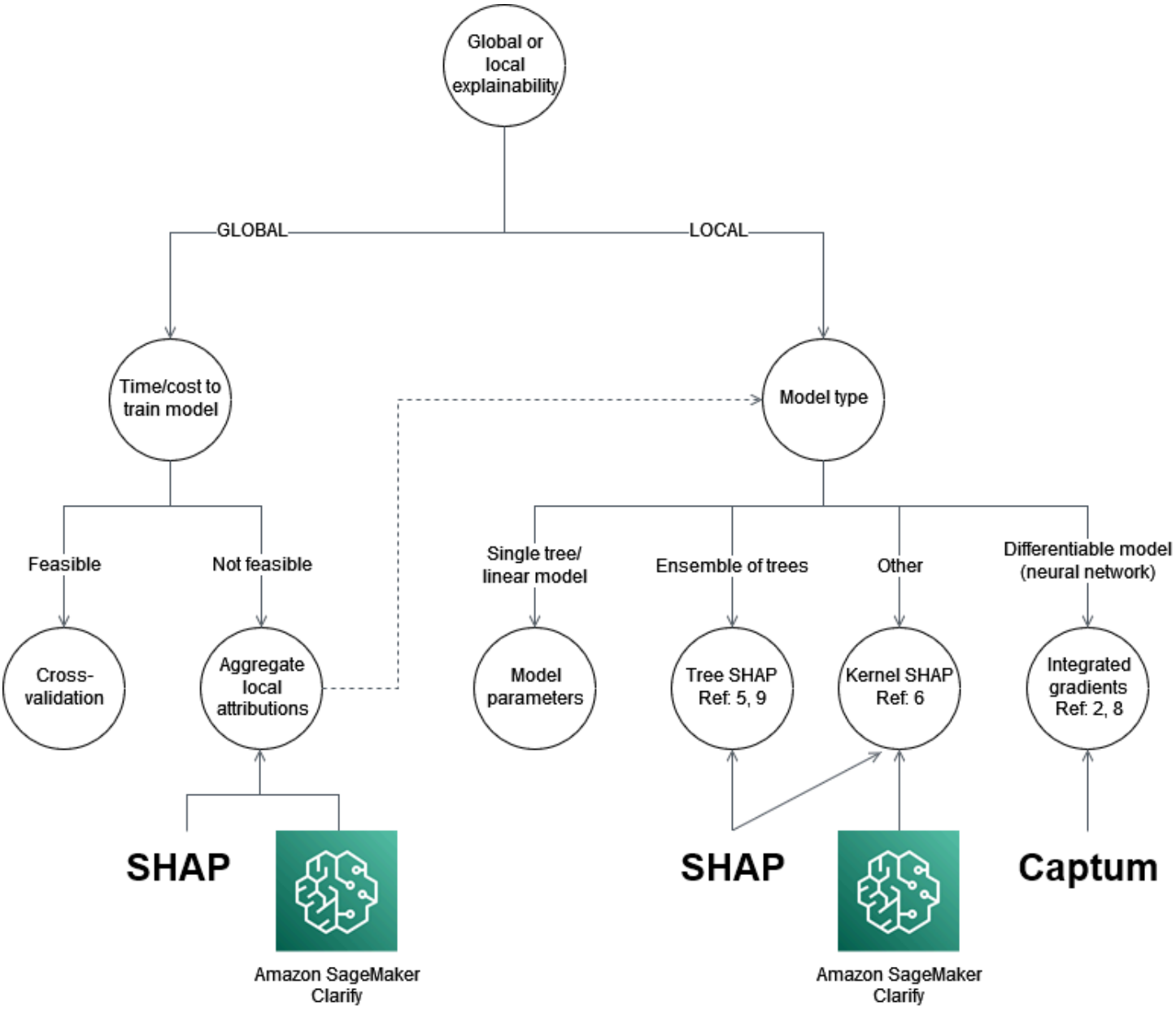
Dans l'illustration suivante, les caractéristiques mesurées d'un iris sont transmises à un modèle qui prédit l'espèce de la plante, et les importances des caractéristiques associées (attributions SHAP) pour cette prédiction sont affichées. Dans ce cas, la longueur des pétales, la largeur des pétales et la longueur des sépales contribuent toutes positivement à la classification d'Iris virginica, mais la largeur des sépales a une contribution négative. (Ces informations sont basées sur le jeu de données sur l'iris de [4].)


Features
petal_length = 6
petal_width = 1.8
sepal_length = 7.2
sepal_width = 3.2
Trained model
Compute feature importance
Prediction
Iris-setosa: 0.00
Iris-versicolor: 0.04
Iris-virginica: 0.96
petal_width
petal_length
sepal_length
sepal_width
+0.36
+0.36
+0.07
−0.02
0.0
0.1
0.2
0.3
0.4
SHAP value

Les scores d'importance des caractéristiques peuvent être globaux, indiquant que le score est valide pour le modèle pour toutes les entrées, ou locaux, indiquant que le score s'applique à une seule sortie du modèle. Les scores d'importance des caractéristiques locales sont souvent mis à l'échelle et additionnés pour produire la valeur de sortie du modèle, et sont donc appelés attributions. Les modèles simples sont considérés comme plus interprétables, car les effets des caractéristiques d'entrée sur les résultats du modèle sont plus faciles à comprendre. Par exemple, dans un modèle de régression linéaire, les magnitudes des coefficients fournissent un score global d'importance des caractéristiques, et pour une prédiction donnée, l'attribution d'une caractéristique locale est le produit de son coefficient et de la valeur de la caractéristique. En l'absence d'un score d'importance des entités locales direct pour une prédiction, vous pouvez calculer un score d'importance à partir d'un ensemble d'entités en entrée de référence afin de comprendre la contribution d'une entité par rapport à la ligne de base.

Bonnes pratiques

L'interprétabilité des modèles est un domaine d'étude en pleine expansion, et les meilleures pratiques doivent être comprises dans ce contexte. Cette section fournit les recommandations actuelles les plus susceptibles d'aboutir à des résultats positifs dans un court laps de temps. Vous pouvez les utiliser avec d'autres méthodes pour interroger complètement les modèles lorsque le temps le permet. L'organigramme suivant présente AWS des recommandations pour trouver la meilleure méthode d'interprétabilité. Les sections qui suivent traitent de ces méthodes plus en détail.



Interprétabilité locale

Les méthodes les plus populaires pour l'interprétabilité locale de modèles complexes sont basées soit sur les explications additives de Shapley (SHAP) [8], soit sur des dégradés intégrés [11]. Chaque méthode comporte un certain nombre de variantes spécifiques à un type de modèle.

Pour les modèles d'ensembles d'arbres, utilisez tree SHAP

Dans le cas des modèles basés sur des arbres, la programmation dynamique permet un calcul rapide et exact des valeurs de Shapley pour chaque caractéristique. Il s'agit de l'approche recommandée pour les interprétations locales dans les modèles d'ensembles d'arbres. (Voir 7], la mise en œuvre est à https://github.com/slundberg/shap.)

Pour les réseaux neuronaux et les modèles différentiables, utilisez des gradients et une conductance intégrés

Les gradients intégrés fournissent un moyen simple de calculer les attributions de caractéristiques dans les réseaux neuronaux. La conductance s'appuie sur des gradients intégrés pour vous aider à interpréter les attributions provenant de parties de réseaux neuronaux telles que les couches et les neurones individuels. (Voir [3,11], la mise en œuvre est terminée https://captum.ai/.) Vous ne pouvez pas utiliser ces méthodes sur des modèles sans utiliser de dégradé ; dans ce cas, vous pouvez utiliser Kernel SHAP (décrit dans la section suivante) à la place. Lorsque le dégradé est disponible, les attributions de gradient intégrées peuvent être calculées plus rapidement que les attributions de Kernel SHAP. L'un des défis liés à l'utilisation de dégradés intégrés consiste à choisir le meilleur point de base pour obtenir une interprétation. Par exemple, si le point de base d'un modèle d'image est l'image dont l'intensité est nulle dans tous les pixels, les zones importantes d'une image qui sont plus sombres peuvent ne pas avoir d'attributions correspondant à l'intuition humaine. Une approche pour résoudre ce problème consiste à utiliser plusieurs attributions de points de base et à les additionner. Cela fait partie de l'approche adoptée dans la méthode d'attribution de caractéristiques XRAI pour les images [5], où les attributions de gradient intégrées qui utilisent une image de référence noire et une image de référence blanche sont additionnées pour produire des attributions plus cohérentes.

Dans tous les autres cas, utilisez Kernel SHAP

Vous pouvez utiliser Kernel SHAP pour calculer les attributions de fonctionnalités pour n'importe quel modèle, mais il s'agit d'une approximation du calcul des valeurs complètes de Shapley et cela reste coûteux en termes de calcul (voir [8]). Les ressources informatiques requises pour Kernel

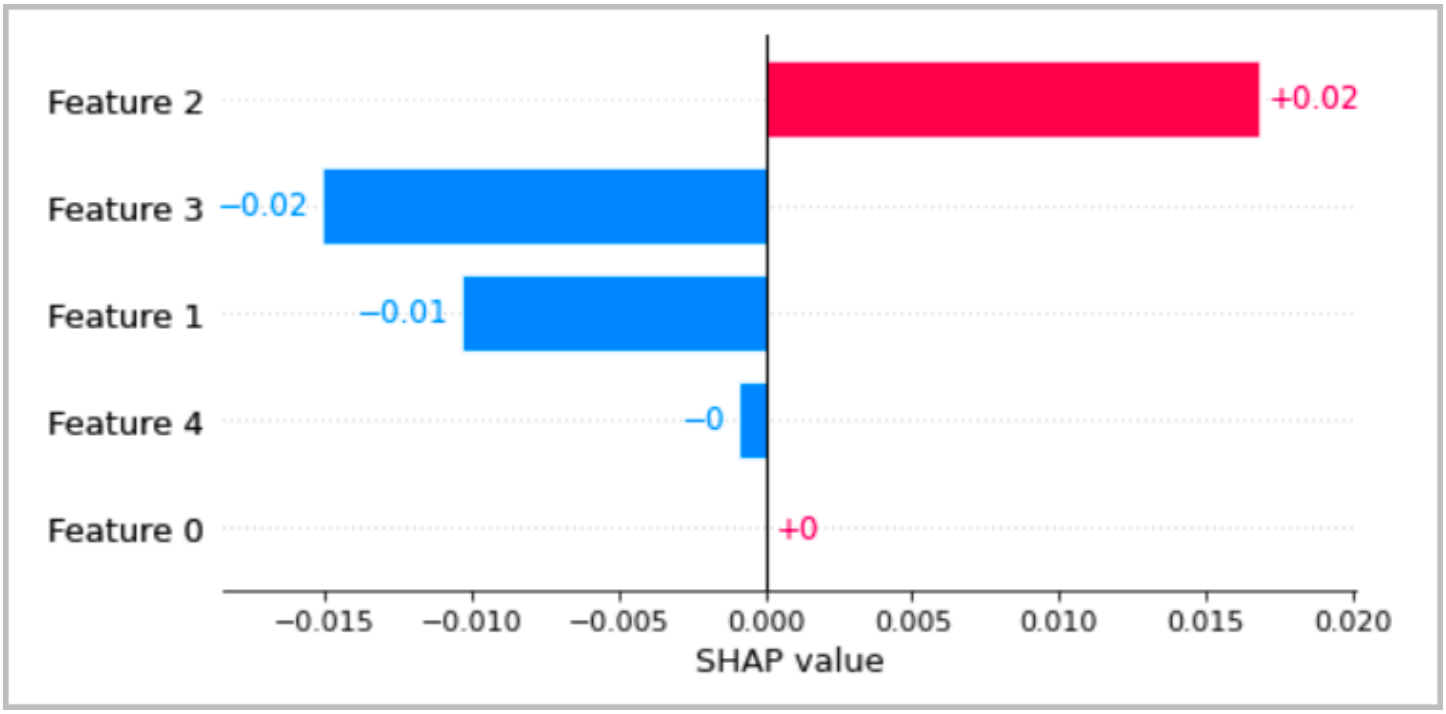
SHAP augmentent rapidement avec le nombre de fonctionnalités. Cela nécessite des méthodes d'approximation qui peuvent réduire la fidélité, la répétabilité et la robustesse des explications. Amazon SageMaker AI Clarify propose des méthodes pratiques qui déploient des conteneurs prédéfinis pour calculer les valeurs Kernal SHAP dans des instances distinctes. (Pour un exemple, consultez le GitHub référentiel Fairness and Explainability with SageMaker AI Clarify.)

Pour les modèles à arbre unique, les variables fractionnées et les valeurs foliaires fournissent un modèle immédiatement explicable, et les méthodes décrites précédemment ne fournissent pas d'informations supplémentaires. De même, pour les modèles linéaires, les coefficients fournissent une explication claire du comportement du modèle. (Le SHAP et les méthodes de gradient intégré renvoient toutes deux des contributions déterminées par les coefficients.)

Le SHAP et les méthodes intégrées basées sur les dégradés présentent tous deux des faiblesses. SHAP exige que les attributions soient dérivées d'une moyenne pondérée de toutes les combinaisons de fonctionnalités. Les attributions obtenues de cette manière peuvent être trompeuses lors de l'estimation de l'importance des caractéristiques s'il existe une forte interaction entre les caractéristiques. Les méthodes basées sur des gradients intégrés peuvent être difficiles à interpréter en raison du grand nombre de dimensions présentes dans les grands réseaux neuronaux, et ces méthodes sont sensibles au choix d'un point de base. Plus généralement, les modèles peuvent utiliser les fonctionnalités de manière inattendue pour atteindre un certain niveau de performance, lequel peut varier en fonction du modèle. L'importance des fonctionnalités dépend toujours du modèle.

Visualisations recommandées

Le tableau suivant présente plusieurs méthodes recommandées pour visualiser les interprétations locales abordées dans les sections précédentes. Pour les données tabulaires, nous recommandons un graphique à barres simple qui montre les attributions, afin qu'elles puissent être facilement comparées et utilisées pour déduire la manière dont le modèle fait des prédictions.



Pour les données texte, l'intégration de jetons entraîne un grand nombre d'entrées scalaires. Les méthodes recommandées dans les sections précédentes produisent une attribution pour chaque dimension de l'intégration et pour chaque sortie. Afin de synthétiser ces informations dans une visualisation, les attributions d'un jeton donné peuvent être additionnées. L'exemple suivant montre la somme des attributions du modèle de réponse aux BERT-based questions qui a été entraîné sur le jeu de données SQUAD. Dans ce cas, l'étiquette prévue et vraie est le symbole du mot « france ».

Legend: Negative Neutral Positive

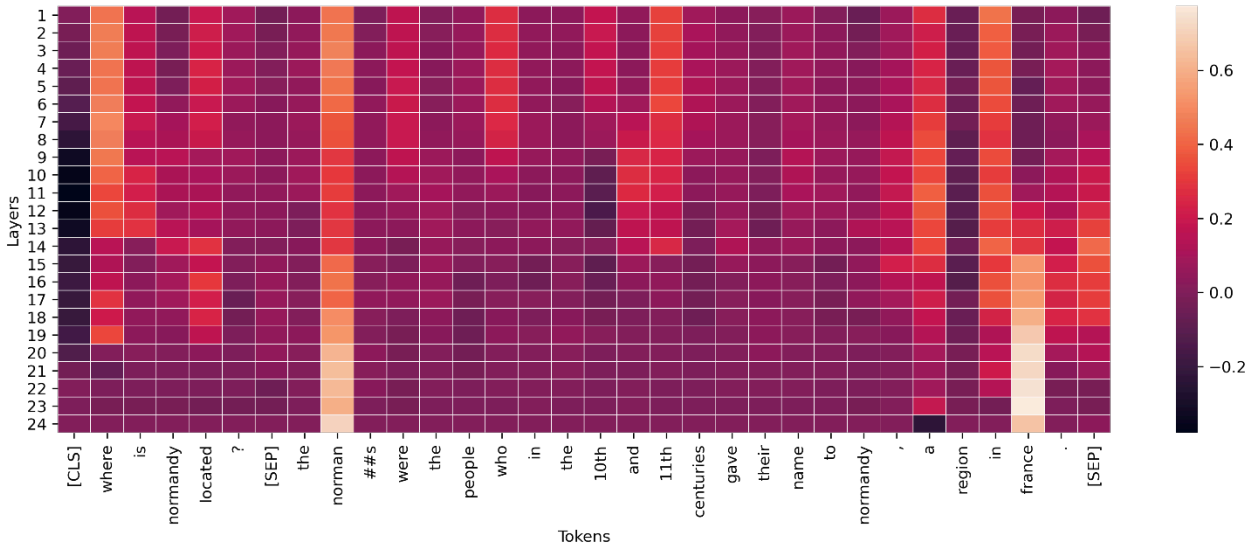
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
29	29 (1.00)	29	4.39	[CLS] where is normandy located ? [SEP] the norman ##s were the people who in the 10th and 11th centuries gave their name to normandy , a region in france . [SEP]

Sinon, la norme vectorielle des attributions de jetons peut être attribuée sous forme de valeur d'attribution totale, comme indiqué dans l'exemple suivant.

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
29	29 (1.00)	29	4.71	[CLS] where is normandy located ? [SEP] the norman ##s were the people who in the 10th and 11th centuries gave their name to normandy , a region in france . [SEP]

Pour les couches intermédiaires des modèles d'apprentissage profond, des agrégations similaires peuvent être appliquées aux conductances à des fins de visualisation, comme indiqué dans l'exemple suivant. Cette norme vectorielle de la conductance des jetons pour les couches de transformateurs montre l'activation éventuelle de la prédiction des jetons finaux (« france »).

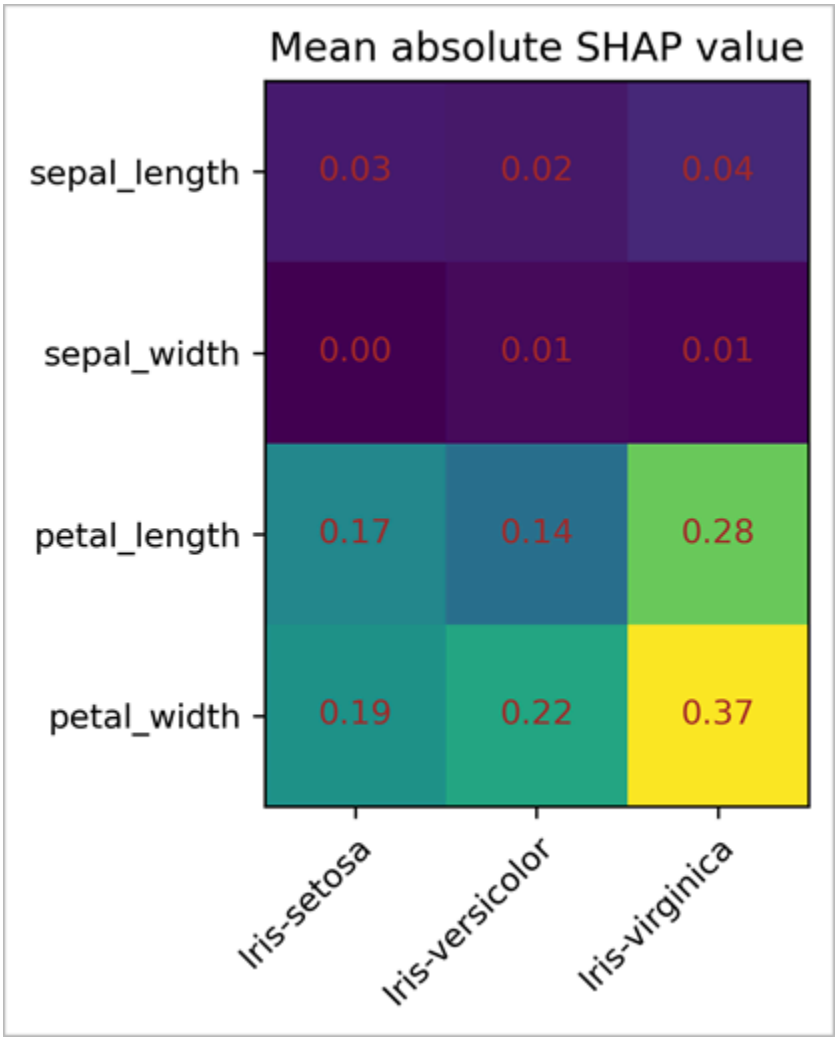


Les vecteurs d'activation conceptuelle fournissent une méthode pour étudier plus en détail les réseaux de neurones profonds [6]. Cette méthode extrait les entités d'une couche d'un réseau déjà entraîné et entraîne un classificateur linéaire sur ces entités afin de tirer des conclusions sur les informations contenues dans la couche. Par exemple, vous souhaiterez peut-être déterminer quelle couche d'un modèle de BERT-based langage contient le plus d'informations sur les parties du discours. Dans ce cas, vous pouvez entraîner un modèle linéaire de partie du discours sur chaque sortie de couche et estimer grossièrement que le classificateur le plus performant est associé à la couche contenant le plus d'informations partielles. Bien que nous ne recommandions pas cette méthode comme méthode principale d'interprétation des réseaux neuronaux, elle peut être une option pour une étude plus détaillée et une aide à la conception de l'architecture du réseau.

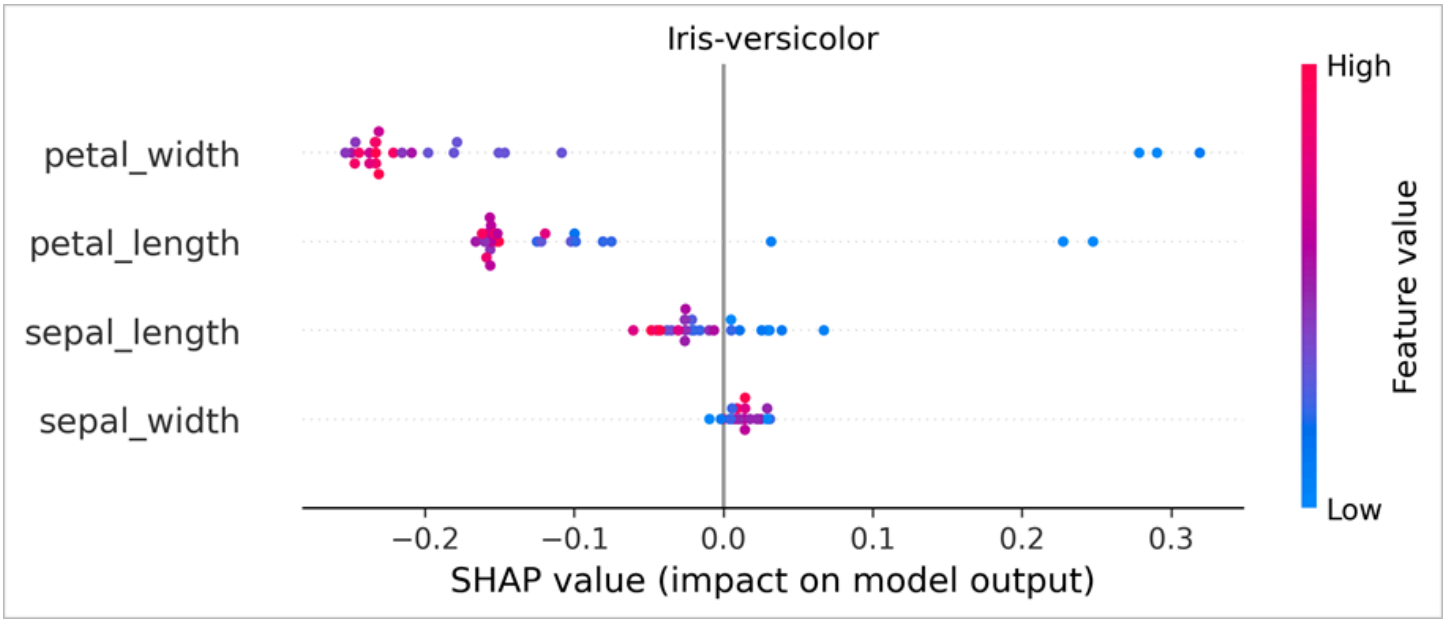
Interprétabilité globale

Comprendre comment les fonctionnalités contribuent à la sortie globale d'un modèle fournit des informations générales utiles pour la sélection des fonctionnalités et le développement du modèle. Pour mesurer l'effet de l'ajout d'une nouvelle fonctionnalité, vous exécutez généralement une validation croisée avec et sans la fonctionnalité. Cependant, il est souvent impossible d'exécuter une validation croisée pour toutes les combinaisons de fonctionnalités et tous les types de modèles considérés en raison des coûts de calcul. D'autres méthodes permettant de déterminer l'importance des caractéristiques sont donc utiles pour prendre des décisions rapides. Notre recommandation pour déterminer les attributions globales des fonctionnalités est d'agréger les scores d'attribution des fonctionnalités locales recommandés dans la section précédente pour toutes les données. Nous recommandons également de calculer la modification du score de validation croisée lorsqu'une fonctionnalité est supprimée, si les contraintes de temps et de calcul le permettent. L'exemple suivant

illustre l'agrégation des scores d'attribution locaux. Il fait la moyenne des magnitudes des valeurs SHAP pour le modèle de classification de l'iris (à partir de l'aperçu) et les trace sous forme de carte thermique. Vous pouvez constater que les mesures des sépales ne jouent pas un rôle important dans le modèle de détermination de la classe d'iris.



Pour une sortie de modèle spécifiée, la collection de valeurs SHAP entre les instances d'évaluation peut être visualisée dans un diagramme en essaim d'abeilles, comme illustré dans le schéma suivant (pour un sous-ensemble de données du jeu de données iris [4]). Vous pouvez voir ici que l'`petal_width`attribut a le plus grand effet sur le résultat du modèle pour la classe`Iris-versicolor`, et qu'une `petal_width` valeur élevée contribue négativement à la prédiction de la classe. Lorsque plusieurs points de données ont une valeur d'attribution de caractéristique identique ou très similaire, les points sont empilés pour indiquer la prévalence la plus élevée à cet emplacement.



Interprétabilité sur AWS

Vous pouvez utiliser des instances Jupyter gérées par Amazon SageMaker AI pour installer facilement des modules Python via Conda et. `pip` Pour plus d'informations sur les packages Python pour SHAP et les méthodes intégrées basées sur les dégradés, consultez la section Ressources. Pour les petites tâches et les tests locaux sur une instance SageMaker AI Jupyter, l'utilisation des méthodes de ces packages Python peut être suffisante. Si vous utilisez un modèle géré par l' SageMaker SageMaker IA, AI Clarify propose des méthodes pratiques pour lancer Kernel SHAP sur une instance dédiée et décharger les calculs pendant qu'un développeur de modèles continue de travailler sur son instance Jupyter. Pour plus d'informations, consultez la section Créer des lignes de base d'attributs d'entités et des rapports d'explicabilité dans la documentation de l' SageMaker IA.

FAQ

Il existe de nombreuses méthodes pour déterminer l'importance des fonctionnalités qui ne sont pas abordées ici. Pourquoi ne sont-ils pas mentionnés ?

Ce guide met l'accent sur ce que nous pensons être les méthodes les plus efficaces et les plus directes pour l'interprétabilité des modèles. D'autres méthodes présentent des avantages en termes de rapidité et de facilité de calcul, et peuvent être appropriées selon le modèle. Les directives contenues dans cet article sont prescriptives et non proscriptives.

Quels sont les points faibles des méthodes préconisées ?

SHAP nécessite des attributions dérivées d'une moyenne pondérée de toutes les combinaisons de fonctionnalités. Les attributions obtenues de cette manière peuvent être trompeuses lorsqu'il s'agit d'estimer l'importance des caractéristiques lorsqu'il existe de fortes interactions entre les caractéristiques. Les méthodes basées sur des gradients intégrés peuvent être difficiles à interpréter en raison du grand nombre de dimensions présentes dans les grands réseaux neuronaux. Les modèles peuvent utiliser les fonctionnalités de manière inattendue pour atteindre un certain niveau de performance, qui peut varier en fonction du modèle. L'importance des fonctionnalités dépend donc toujours du modèle.

Ressources

Références

1. Adadi, Amina et Mohammed Berrada. 2018. « Un coup d'œil à l'intérieur de la boîte noire : une enquête sur l'intelligence artificielle explicable (XAI). » Accès IEEE 6 : 52138 — 52160.
2. Ancône, Marco, Enea Ceolini, Cengiz Oztireli et Markus Gross. 2018. « Vers une meilleure compréhension des méthodes d'attribution basées sur le gradient pour les réseaux de neurones profonds. » Actes de la Conférence internationale sur les représentations de l'apprentissage (ICLR). Archives : 1711.06104.
3. Dhamdhere, Kedar, Mukund Sundararajan et Qiqi Yan. 2018. « Quelle est l'importance d'un neurone ? » Actes de la trente-sixième Conférence internationale sur le Machine Learning (ICML). Archives : 1805.12233.
4. Dua, Dheeru et Casey Graff. 2019. Référentiel de machine learning de l'Université de Californie à Irvine (UCI) [http://archive.ics.uci.edu/ml]. Irvine, Californie : Université de Californie, École d'information et d'informatique.
5. Kapishnikov, Andrei, Tolga Bolukbasi, Fernanda Viegas et Michael Terry. 2019. « XRAI : de meilleures attributions selon les régions. » Actes de la conférence internationale IEEE/CVF sur la vision par ordinateur (ICCV) : 4948—4957. Archives : 1906.02825.
6. Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas et Rory Sayres. 2018. « L'interprétabilité au-delà de l'attribution de fonctionnalités : tests quantitatifs avec des vecteurs d'activation de concepts (TCAV). » Archives : 1711.11279.
7. Lundberg, Scott M., Gabriel G. Erion et Su-In Lee. 2019. « Attribution de fonctionnalités individualisée cohérente pour les ensembles d'arbres. » Archives : 1802.03888.
8. Lundberg, Scott M. et Su-In Lee. 2017. « Une approche unifiée pour interpréter les prédictions des modèles ». Progrès des systèmes de traitement de l'information neuronale (NIPS) 30. Archives : 1705.07874.
9. Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev et Percy Liang. 2016. « SQu AD : plus de 100 000 questions pour la compréhension automatique du texte. » Archives : 1606.05250.
10. Ribeiro, Marco T., Sameer Singh et Carlos Guestrin. 2016. « Pourquoi devrais-je te faire confiance ? » : Expliquer les prédictions de n'importe quel classificateur. » KDD '16 : Actes de la 22e conférence internationale ACM SIGKDD sur la découverte des connaissances et l'exploration de données : 1135-1144. Archives : 1602.04938.

11. Sundararajan, Mukund, Ankur Taly et Qiqi Yan. 2017. « Attribution axiomatique pour les réseaux profonds ». Actes de la 34e Conférence internationale sur le Machine Learning 70 : 3319-3328. Archives : 1703.01365.

Packages logiciels externes

- FORME : https://github.com/slundberg/ forme
- Captum : https://captum.ai/

Lectures supplémentaires

- Amazon SageMaker AI clarifie l'explicabilité du modèle (documentation sur l'SageMaker IA)
- Référentiel Amazon SageMaker AI Clarify (GitHub)
- Molnar, Christoph. Apprentissage automatique interprétable. Un guide pour rendre les modèles de boîtes noires explicables, 2019.

Historique du document

Le tableau suivant décrit les modifications importantes apportées à ce guide. Pour être averti des mises à jour à venir, abonnez-vous à un fil RSS.

Modification	Description	Date
Publication initiale	—	28 février 2022

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.