Master 2 MIAGE

Mémoire de fin d'études

Interprétabilité et explicabilité des modèles boites noires

Auteur :
Damien **JAIME**

Tuteur :
Emanuel **HYON**

Promotion 2019-2020

Remerciements

LISTE DES ACRONYMES

IA *Intelligence artificielle*

NN *Réseau de neurones (Neuronal Network)*

DNN *Réseau de neurones profond (Deep Neuronal Network)*

TABLE DES MATIÈRES

INTRODUCTION

Nous sommes au quotidien en contact avec une multitude boites noires (c'est-à-dire un système cachant sa logique interne à l'utilisateur). L'utilisateur ne pouvant pas savoir ce qu'il se passe dans le système, de nombreuses questions se posent. Notamment sur la confiance à accorder en ces systèmes, mais aussi d'un point de vue éthique particulièrement sur la protection de la vie privée de l'utilisateur.

Avec l'avènement de l'intelligence artificielle dans l'aide à la prise de décision, la problématique de compréhension de ces systèmes est devenu un des enjeux majeurs de notre génération. En effet, comment peut-on être assuré que notre modèle est réellement fiable ? Notre modèle a-t-il héritée involontairement de biais présent dans ses jeux de données ? Quel degré de compréhension est nécessaire et pour quelle complexité ?
Ces questions sont très importantes dans le domaine médical par exemple où un faux négatif peut être très problématique, car à la différence des faux positifs il n'y a pas de vérification postérieure effectué par l'humain. Par ailleurs, nombreux sont les cas où une intelligence artificielle a hérité de biais involontaire présent dans ses jeux de données d'entraînement engendrant différents types de discriminations. Comme par exemple l'intelligence artificielle de recrutement de l'entreprise Amazon qui rejetait automatiquement les CVs contenant le mot "femme".

Depuis 2016, en Europe, le Règlement Général de l'Union européenne sur la Protection des Données (RGPD) impose un droit à l'explication des décisions algorithmique prise au sujet d'un utilisateur [1]. Or les raisons des décisions de ces algorithmes étant très abstraites même pour la personne l'aillant créée, un problème se pose et les utilisateurs ne peuvent pas réellement utiliser ce droit à l'explication.

Autre les questions d'éthique et de protections des utilisateurs, expliquer

ces boites noire serai d'une très grande aide pour les entreprises afin de créer des produits plus sûr, plus efficacement et les aider à mieux gérer les problèmes de responsabilité.
Pour toutes ces raisons, fournir une explication à un système de décision type boite noire est un sujet de parfaite actualité et suscitant de nombreuses attentes. Le "Gartner hype cycle for Emerging Technologies" de 2019 place ce sujet (Explainable AI) dans le pic des attentes et suppose une maturité dans les cinq à dix ans.

L'objectif de ce mémoire est donc de comprendre comment une intelligence artificielle et plus largement comment un système boite noire est amené à prendre une décision. Pour cela, nous allons analyser les différentes méthodes existantes dans la littérature et les appliquer à plusieurs cas précis de systèmes boites noires. Afin de les comparer, les agréger, les classifier et de mettre en place une méthodologie fournissant une prédiction interprétable et explicable.

Le reste de ce mémoire sera organisé en plusieurs parties. Tout d'abords, nous commencerons avec un état de l'art (Chapitre 1) divisé en deux parties : la première sera axée sur des définitions et présentations d'éléments conceptuel, tandis que la seconds traitera des différentes méthodes et outils existant afin d'expliquer différents types de boites noire. Ensuite, dans le Chapitre 2 nous commencerons par apporter une méthodologie permettant de trouver quelle(s) méthode(s) explicative utiliser en fonction d'un problème donnée (type de boite noire, accès au code source ?, type de données ...). La suite du chapitre 2 consistera donc à appliquer notre méthodologie afin de l'évaluer celons deux critères : sa capacité à rendre un système de décision compréhensible par l'homme ainsi que sa capacité a être utilisée dans une multitude de problématiques.

Chapitre 1

ÉTAT DE L'ART

1.1 Besoin d'explicabilité

Expliquer les modèles de décisions boites noires est un besoin se faisant de plus en plus ressentir et suscitant de nombreux débats et tables rondes dans la communauté scientifique. En effet la non-explicabilité de ces modèles posent des problèmes de fiabilité, d'éthique et de responsabilité, ce floue accentue aussi la méfiance et la peur des personnes envers l'intelligence artificielle.

1.1.1 L'éthique

Tout d'abord, de nombres cas très exposés dans les médias nous montrent que cette incompréhension peut amener à des problèmes du point de vue éthique. La question de l'éthique est souvent ramenée à deux aspects principaux : la discrimination et la protection des données privées des usagers.

La discrimination

Avant de donner des exemples, la question essentielle à se poser est de savoir comment une intelligence artificielle est amenée à prendre des décisions discriminante. Dans un article, Barocas et Selbst distinguent cinq façons par lesquelles une intelligence artificielle pourrait aboutir à une discrimination [7]. Ces cas sont uniquement des cas involontaires et ne prennent pas en compte une discrimination délibérée qui pourrait bien évidemment être possible.

— **Définition des variables cibles et des étiquettes de classes** : le but d'une intelligence artificielle est de découvrir des corrélations dans des jeux de données. Ainsi, certain choix de variables cibles ou ou d'étiquettes de classes peuvent amener à faire des corrélations discriminatoire. Un exemple trivial serait de considérer un modèle permet-

tant de savoir si un employé fait du bon boulot, l'entreprise choisira de prendre en compte les retards des employés dans son modèle d'évaluation. Les personnes défavorisées habitant souvent loin de leur lieu de travail, ont plus souvent tendance à être en retard à cause des embouteillages et des problèmes de transport. Ces employés habitant loin, ils seront jugés comme faisant du moins bon boulot ce n'est pas forcément le cas. Si on ajoute à cela le fait que les personnes issus de l'immigration sont en moyenne plus pauvre et habitent en moyenne plus loin des centres villes, nous pouvons arriver involontairement à une IA discriminante par le simple fait d'utiliser l'étiquette "retard" dans l'évaluation des performances d'un employé.

— **Les données d'apprentissage** : les données d'apprentissage peuvent contenir des biais qui seront ensuite appris par notre modèle. L'exemple de l'IA de recrutement de l'entreprise Amazon qui rejetait les CVs contenant le mot "femme" évoqué en introduction le montre bien. Cette IA avait appris en analysant tous les profils recrutés par Amzon dans le passé, or les personnes recrutées étaient très majoritairement des hommes. L'IA a donc déduis qu'il était préférable de recruter des hommes.
— **La collecte des données d'apprentissages** : Les lieux dans lesquels sont récoltées les données d'apprentissage sont aussi déterminant. Par exemple si l'on récolte des données concernant la criminalité dans un quartier avec des personnes issus majoritairement de l'immigration, l'IA aura plus tendance à considérer les immigrés comme de potentiels criminels par rapport aux personnes non-imigrées.
— **La sélection des caractéristiques** : Afin que l'IA puisse s'entraîner, il faut lui fournir des données qui sont en réalité une représentation simplifiée de notre monde. Ainsi, son créateur doit faire des choix pour sélectionner les caractéristiques qui constitueront cette représentation. Ce choix peut par concours de circonstance involontairement découler sur de la discrimination il faut donc les choisir avec parcimonie.
— **Données indirect** : Certaines données peuvent inclure des données indirect. Par exemple : *"un jeu de données qui ne contient pas de données explicites sur l'orientation sexuelle peut tout de même la dévoiler. Une étude de 2009 a montré que les liens d'"amis" sur Facebook révèlent l'orientation sexuelle par une méthode de prédiction précise de l'orientation sexuelle des utilisateurs de Facebook fondée sur l'analyse de leurs liens. Le pourcentage d'"amis" s'identifiant comme homosexuels serait fortement corrélé avec l'orientation sexuelle de l'uti-*

lisateur concerné"[1].

Protection des données

L'apprentissage d'une intelligence artificielle requière un très grand nombre de données, parmi lesquelles peuvent se trouver des données personnelles. L'utilisation d'un modèle pourrait donc aussi nécessiter de fournir des données personnelles afin d'effectuer une prédiction. L'utilisation de ces données jugées critiques est problématique, car des restrictions et des règles de protection en découlent. Tel qu'évoqué dans l'introduction, le RGPD par l'addition de plusieurs de ses articles implique un "droit à l'explication". Ce droit a été démontré par Seth Flaxman et Bryce Goodman [1], ce qui a donc pour effet d'accentuer le besoin d'applicabilité des modèles de décision boites noires. En effet, pour le moment ce droit est inaccessible par les usagers et les entreprises éprouvent des problèmes de responsabilités.

1.1.2 Fiabilité et confiance

Deuxièmement, expliquer les décisions prisent par un modèle boite noire permettrai d'augmenter la fiabilité que l'on lui accorde. Le modèle peut très bien fonctionner de manière dans la plupart des cas, mais il est possible qu'il réagisse mal dans un certain cas bien spécifique. Avec l'absence de compréhension du fonctionnement interne de notre modèle, nous pouvons ne pas prévoir ce cas spécifique. Le domaine médical par exemple est en attente d'une plus grande fiabilité de ses modèles en effet, une erreur peut avoir des conséquences graves et mettre en danger des personnes. De plus, un soudage publié par OpinionWay [2] montre que 30% des Français ont peur d'un jugement porté par l'intelligence artificielle dans le domaine financier et 21% dans le domaine médical. L'incompréhension de la logique interne des algorithmes décisionnels tant à accentuer la méfiance envers ces technologies. Fournir une explication permettrai d'augmenter l'acceptation de ces nouvelles technologies.

1.1.3 Performance

Enfin, expliquer la logique interne de la boite noire permettrai d'améliorer le développement de celle-ci, de la rendre plus compétitive et plus performante. Ces explications seront aussi utiles à durant leur développement afin

1. Frederik Zuiderveen Borgesius, Discrimination, intelligence artificielle et décisions algorithmiques
2. https ://www.opinion-way.com/fr/

de comprendre pourquoi notre modèle ne fonctionne pas bien dans certains cas voir même pourquoi il ne converge pas du tout.

1.2 Interprétabilité et explicabilité

Pour commencer, il est important de définir et de saisir la différence entre les thermes interprétabilité et explicabilité. Nous allons pour cela nous baser sur la définition donnée par Tim Miller concernant l'interprétabilité [2] : "*the degree to which an observer can understand the cause of a decision*". L'interprétabilité consiste donc à fournir une explication à une prise de décision qui soit compréhensive par l'homme. Il sera possible de déterminer explicitement les caractéristiques les plus importants pour la prise de décision. L'explicabilité d'un modèle est le fait de "*rendre compte explicitement à partir de données et caractéristiques connues de la situation*". Ce qui consiste donc à mettre en relation les valeurs prises par certaines caractéristiques et leurs conséquences sur la prédiction. [3]
Créer un modèle interprétable implique de prendre en compte plusieurs facteurs :

Interprétabilité globale et local L'interprétabilité d'un modèle est dite *globale*, lorsque l'on comprend la logique de la totalité du modèle et que nous sommes en mesure d'expliquer toutes les solutions possibles de notre modèle. À contrario, elle sera dite *local*, lorsqu'il est possible d'expliquer seulement une ou plusieurs solutions spécifiques. Un problème complexe pourra ainsi être découpé en un sous problème plus simple afin de pouvoir expliquer une partie des solutions comme le montre la figure 1.1 ci-dessous.

Limatation temporel Le temps que l'on peut allouer pour fournir une explication à notre modèle est aussi à prendre en compte. En effets, fournir une explication peut prendre du temps et cette explication peut être longue à appréhender pour un être humain. Dans certains contextes où la prise de décision devra être effectuée rapidement, il sera préférable d'avoir une explication simple, compréhensible et fournis rapidement par la machine. Dans d'autres cas, nous pourrons prendre le temps d'aborder une explication plus complexe et détaillée.

La cible de l'explication La nature de l'expertise de l'utilisateur est aussi un facteur à prendre en compte dans le choix de l'explication que nous voulons lui apporter. En effets, un expert dans le domaine

3. https ://perso.math.univ-toulouse.fr/mllaw/home/statisticien/explicabilite-des-decisions-algorithmiques/

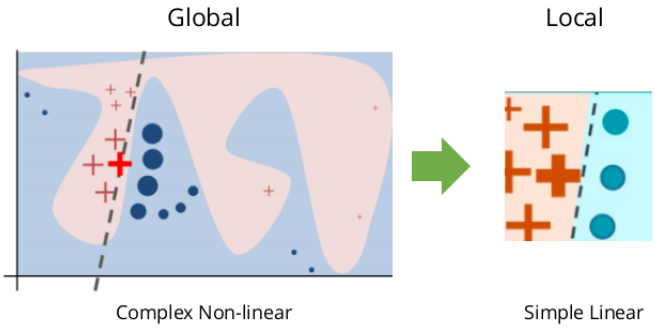


Figure 1.1 – Un problème global complexe pouvant être expliqué à échelle local. *Source : www.kdnuggets.com*

aura tendance à préférer une explication exhaustive et précises qu'une explication simple et opaque et inversement pour une personne moins à l'aise.

1.2.1 Pré-requis d'un modèle interprétable

En plus de prendre en compte les facteurs précédemment évoqués, un modèle interprétable doit être capable de satisfaire une liste de choses souhaitées. L'article "A survey of methods for explaining black box models"[6] met en avant, après une analyse de différents états de l'art traitant de ce sujet, les desiderata d'un modèle interpretable :

Interpretabilité Dans quelle mesure le modèle ou la prédiction sont compréhensifs par l'homme. Ce sujet est encore en débat afin de savoir comment mesurer cette interprétabilité.

Précision Dans quelle mesure le modèle interprétable prédit avec précision les différentes instances. La précision d'un modèle peut être faite avec le *score de précision* (accuracy score), il s'agit simplement d'un rapport entre les observations correctement prédites et les observations totales. La précision est l'indicateur de base afin de calculer la précision d'un modèle, ce score fonctionne mieux si les faux positifs et les faux négatifs ont un coût similaire. Si le coût des faux positifs et des faux négatifs est très différent, il vaut mieux regarder le *F1-score*. Le F1-score est la moyenne pondérée de la précision et du rappel.

$$F1Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Où la *précision* est le rapport des observations positives correctement prédites au total des observations positives prévues. Et le *recall* (rappel) est le rapport des observations positives correctement prédites à toutes les observations dans la classe réelle.

Éthique Si notre modèle traite des données personnelles, il devra garantir en plus une protection contre toutes formes de discrimination ainsi qu'une protection de la vie privée des personnes concernées.

Ces différents aspects jouent un rôle important quant à la confiance qu'un utilisateur va apporter à notre modèle interprétable. De plus d'autres notions viennent s'ajouter, notamment pour les modèles d'exploration de données et d'apprentissage automatique. Il est important de respecter des critères tels que la *robustesse*, la *causalité*, l'*évolutivité* et la *généralité*. Cela signifie qu'un modèle doit garantir un certain niveau de performance indépendamment des données d'entrée (robustesse), et que les changements d'entrée dû à une perturbation affectent le comportement du modèle (causalité). Enfin, étant donné que nous pouvons utiliser le même modèle avec une multitude données d'entrée et dans différents cas d'applications, il est préférable d'avoir des modèles portables possédant un minimum de restriction (évolutivité, généralité).

1.3 Approches utilisées

Face à ce besoin d'explicabilité grandissant, les entreprises peuvent préférer se tourner vers des modèles moins performant mais plus explicable afin de contourner la problématique de boite noire. Ainsi, nombreuse sont les entreprises qui se détournent de l'apprentissage profond au profil de l'apprentissage par arbre de décision ou les modèles linéaires par exemple qui fournissent un résultat plus compréhensibles et interprétable.

Apprentissage par arbre de décision

L'apprentissage par arbre de décision exploite un arbre qui a pour noeuds des conditions, les arrêtes correspondent à la valeur d'une variable d'entrée et les feuilles correspondent aux différents labels possible. La figure 1.2 montre un exemple trivial d'arbre de décision. Ainsi, la décision sera simplement expliquée en exprimant les chemins de l'arbre empruntés. Il pourra par exemple être exprimé sous la forme d'un if-then :

if condition1, condition2, condition3 then outcome

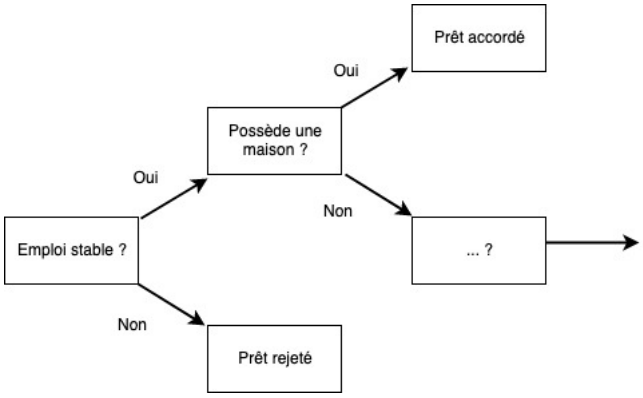


Figure 1.2 – Exemple d'un arbre de décision

Mais, la recherche dans ce domaine n'a pas évoluée depuis un certain temps et l'éventail des algorithmes disponibles est assez limité.

Sur-couche explicative

Le but étant d'essayer de fournir à l'utilisateur des éléments approximatifs permettant de comprendre son modèle boite noire, comme par exemple identifier les variables d'entrée les plus importantes dans la prise de décision du modèle. Comme le montre la figure 1.3, la sur-couche explicative venant se greffer après la prédiction du modèle. Différents outils proposent une approche permettant de livrer des éléments de compréhension approximatif pour un modèle boite noire simple. Nous allons présenter les deux outils les plus populaires, LIME et SHAP.

1.3.1 LIME

LIME signifie Local Interpretable Model-Agnostic Explanations. Comme son nom l'indique, l'objectif est de fournir une interprétation local à un modèle de classification ou de régression (model agnostic). Le but étant de créer un modèle de substitution formé pour fournir une prédiction approximative de notre boite noire cible. L'idée est de sonder la boite noire autant de fois que nécessaire en faisant varier les données d'entrées afin de comprendre pourquoi le modèle fournis une telle prédiction. Par exemple dans la figure 1.4 tirée de l'article original de la présentation de LIME[4] : L'image originelle (a) est donnée à notre boite noire qui prédit "Guitare électrique", "Guitare acoustique" et "Labrador". Nous constatons que le modèle s'est trompé sur la reconnaissance de la guitare électrique et nous utilisons LIME afin de comprendre ce qu'il s'est passé. LIME va prendre notre image (a) et la dériver de

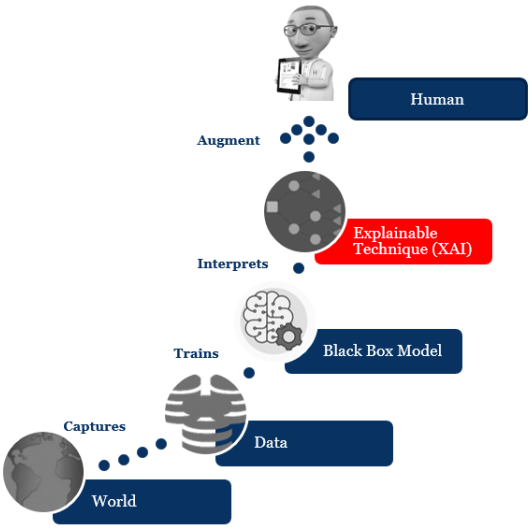
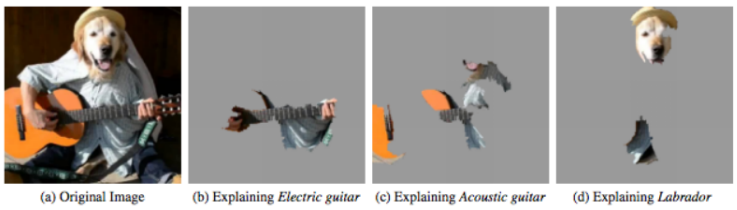


Figure 1.3 – Position de la sur-couche explicative dans le processus de prédiction. *Source : www.kdnuggets.com*

plusieurs façons en cachant certaines parties et envoyer ces nouvelles images à notre modèle. Le but étant de trouver les parties qui une fois cachées font que le modèle ne prédit plus la même chose. Puis une fois les trois classes trouvées, LIME nous envoie une explication où l'on voit les parties de l'image aillant joué un rôle dans la prédiction de chaque labels de notre modèle.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Figure 1.4 – Source : LIME Paper

1.3.2 SHAP

SHAP signifie SHapley Additive exPlanations, c'est une méthode permettant de fournir une interprétation local à une prédiction. Cette méthode est

basée sur la valeur de Shapley issus de la théorie des jeux, il est donc nécessaire dans un premier temps d'expliquer brièvement cette valeur de Shapley.

La valeur de Shapley introduit par Shapley en 1953, permet de repartir les gains équitablement dans un jeu coopératif. Le but étant que tous les joueurs coopérant ensemble reçoivent un certain gain en fonction de leur contribution.

SHAP reprends cette idée afin d'expliquer toutes sortes de modèle de machine learning, le but étant d'associer une valeur égale à sa contribution dans la prédiction pour chaque caractéristique d'entrée. Prenons par exemple un modèle permettant de prédire le prix d'un logement. Une multitude de caractéristiques sont données en entrés à notre modèle afin d'estimer le prix du logement, le but de SHAP est donc de définir pour chacune de ces caractéristiques leur impact monétaire sur le prix final du logement. Il interrogera donc la boite noire avec une multitude de cas différents afin d'isoler le prix de chaque caractéristique. Prenons le cas de la figure 1.5 où l'on veut expliquer la prévision de 310 000 euros pour un logement de 50m2, au premier étage, près d'un parc et où les animaux sont interdits. SHAP va donc recréer la même l'instance et la donner en entrée à la boite noire en autorisant les animaux afin d'en déduire le coût (dans notre exemple 10 000 euros). Cette opération sera effectuée sur toutes les caractéristiques du problème afin de trouver la contribution de chacune. Cela peut sembler trivial dans cet exemple mais il peut exister des dépendances entre les caractéristiques qui modifient leurs coût en fonction de la présence ou non d'une ou plusieurs autre(s) caractéristique(s).

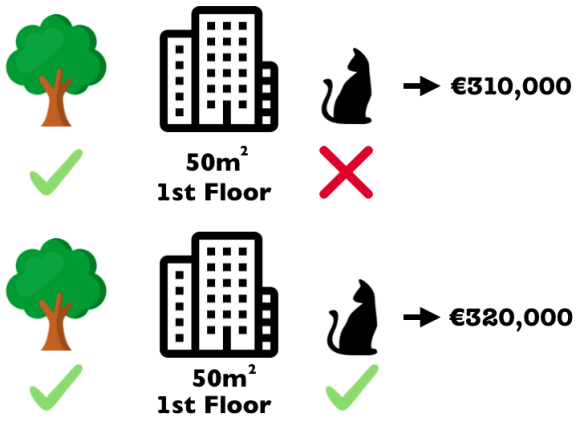


Figure 1.5 – Exemple de l'explication du coût d'un logement avec SHAP. *Source : Christoph Molnar, Interpretable Machine Learning*

1.3.3 Limites de ces implémentations

Il existe aujourd'hui de nombreuses implémentations de ces méthodes qui sont beaucoup utilisées, mais elles essuient aussi quelques critiques et ne conviennent pas dans tous les cas de figures. Premièrement, ces algorithmes représentent un coût non négligeable. Ces méthodes basées sur des simulations et interrogeant notre modèle plusieurs fois pour une seule prédiction, peuvent poser des problèmes de performances lorsque l'on manipule un très grand nombre de données. Aussi la pertinence des explications fournies sont sujettes à débat, en effet ce sont des approximations et rien ne nous garantis que ces explications correspondent réellement au fonctionnement de notre modèle. Ces explications pourraient même donner l'effet inverse à celui escompté, en effet on peut arriver dans des cas où l'on fait confiance à notre modèle grâce à ces explications alors qu'elles ne sont pas fondées et ainsi prendre une décision critique appuyée sur une explication erroné.

Il est donc nécessaire d'utiliser ces algorithmes avec parcimonie et en aillant conscience qu'ils ne fournissent que des approximations de notre problème. Mais ce genre d'approches sont très bénéfiques pour la recherche, nous permettent d'apporter de nouvelles solutions et de mettre en lumière de nouvelles problématiques. Il est donc maintenant nécessaire de définir les différents types de problème à expliquer ainsi que les différents explicateurs possible afin de bien comprendre quelles différents types d'algorithmes explicatifs existantes.

1.4 Ouvrir les boites noires

L'article "A survey of methods for explaining black box models"[6] passe en revue cinquante-quatre méthodes aillant pour but d'ouvrir différentes boites noires, en nous fournissant une classification exposée dans un tableau disponible en annexe 1 et 2 page ? ? ?. Nous aurons l'occasion de reparler de ce tableau plus tard. Cette revue permet de distinguer différents types de problèmes, d'explicateurs, de boites noires, de données d'entrée ainsi que différentes restrictions sur le modèle (accès au code, aux données...). Nous allons commencer par expliquer ces différentes caractéristiques.

1.4.1 Types de problèmes

Explication du modèle

Ce problème consiste à fournir un modèle interprétable et transparent capable d'imiter le comportement d'une boite noire et de nous fournir un prédicat compréhensible par l'homme.
Étant donnée un prédicateur de boite noire b et un ensemble de données D, le problème d'explication du modèle consiste à trouver une fonction f telle que f(b,D)=c où c est un prédicateur compréhensible capable d'imiter le comportement de b et dérivable afin d'obtenir une explication.

Explication du résultats

Ce problème consiste à fournir un résultat interprétable, c'est-à-dire que le modèle devra fournir le résultat avec une explication sur les raisons qui l'ont poussé à donner cette prédiction. Il n'est pas nécessaire d'expliquer la logique interne du système mais seulement le processus de décision pour une instance donnée (interprétation local).

Inspection de la boîte noire

Ce problème consiste à fournir une représentation visuelle ou textuelle afin de comprendre le fonctionnement interne de notre boite noire.
Étant donnée un prédicateur de boite noire b et un ensemble de données D, le problème d'explication du modèle consiste à trouver une fonction f telle que f(b,D)=V où V est une représentation du fonctionnement de la boite noire.

Conception transparente

Ce problème consiste à fournir un modèle transparent qui soit directement interprétable globalement ou localement.

1.4.2 Types d'explicateurs

- **Arbre de décision (Decision Tree)** : Expliqué dans la section 1.3 l'arbre de décision exploite un arbre qui a pour noeuds des conditions, les arrêts correspondent à la valeur d'une variable d'entrée et les feuilles correspondent aux différents labels possible.
- **Règles de décision (Decion Rules)** : Utilisé pour expliquer le modèle, le résultat ainsi que pour la conception transparente. Il est aussi

possible de transformer un arbre en un ensemble de règles. Les règles de décisions sont simplement des conditions IF-THEN.

— **Importance des fonctionnalités (Features Importance)** : Solution souvent utilisée, elle consiste à trouver les entrées de la boite noire pour lesquelles les poids sont les plus importants. Par exemple, pour une classification d'image, de trouver les pixels les plus importants dans l'entrée pour en arriver à une prédiction. Comme le montre la figure 1.6 tirée de l'article *"Explainable Artificial Intelligence : Understanding, Visualizing and Interpreting Deep Learning Models"* [8].

Figure 1.6 – Importance des fonctionnalités : explication de la prédiction "coq"

— **Masque saillant (Salient Mask)** Généralement utilisé pour expliquer localement les réseaux de neurones profond (DNN), le salient mask permet de mettre en évidence visuellement les parties déterminantes de l'entrée analysée. L'article *"Real Time Image Saliency for Black Box Classifiers"*[5] décrit son fonctionnement et la figure 1.7 montrant un exemple de salient mask est tirée de cet article.

— **Analyse de sensibilité (Sensitivity Analysis)** : Généralement utilisée pour l'inspection de boite noire. L'analyse de sensibilité consiste à évaluer l'incertitude statistique du résultat d'une boîte noire avec les différentes sources d'incertitude dans ses entrés. En d'autres thermes, cela consiste à modifier des variables d'entrées afin de voir si cela a un impact sur le résultat en sorti et donc de savoir si elles affectent notre prédiction.

— **Diagramme de dépendance partielle (Partial Dependence Plot)** : Ces graphiques permettent de comprendre l'effet d'une ou deux variables d'entrée sur la sortie du modèle. Le but étant de montrer si

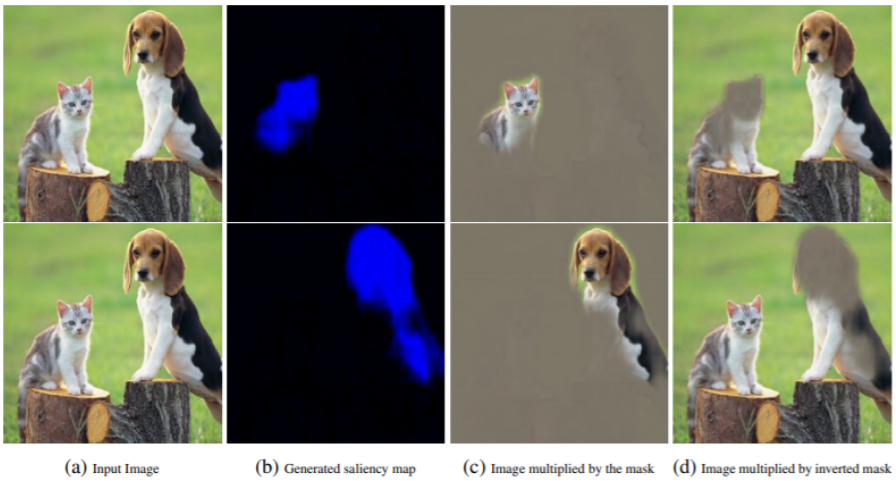


Figure 1.7 – Masque saillant : explication de la prédiction chien et chat

la relation entre une caractéristique d'entrée et la sortie est linéaire, monotone ou plus complexe. Par exemple, appliqué à un modèle de régression linéaire, les tracés de dépendance partielle montrent toujours une relation linéaire. Nous somme limité par une ou deux variables à la fois, car une variable donne une représentation en 2 dimensions du problème et deux variables fournissent donc une représentation 3 en dimensions. Par exemple, la figure 1.8 tirée du livre [3] montre trois diagrammes de dépendances différents sur trois valeurs d'entrées (la température, l'humidité et la vitesse du vent) et leurs impactes linéaire sur la prédiction du nombre de vélos.

— **Sélection de prototype (Prototype Selection)** : Cet explicateur consiste à retourner, avec le résultat, un exemple très similaire à l'enregistrement classifié, afin de préciser avec quel critère la prédiction a été renvoyée.

— **Activation des neurones (Neurons Activation)** : L'analyse des réseaux de neurones permet aussi de comprendre son comportement. Cela consiste à analyser les neurones activés pour chaque entrés passées en argument à notre modèle.

L'explicateur varies en fonction de notre type de problème, de notre type de boite noire, de notre type de données d'entrée et de nos différentes restrictions sur le modèle (accès au code, aux données...). Nous allons dans le chapitre suivant revenir sur le tableau évoqué précédemment [6] afin de définir quel explicateur utiliser et dans quels cas.

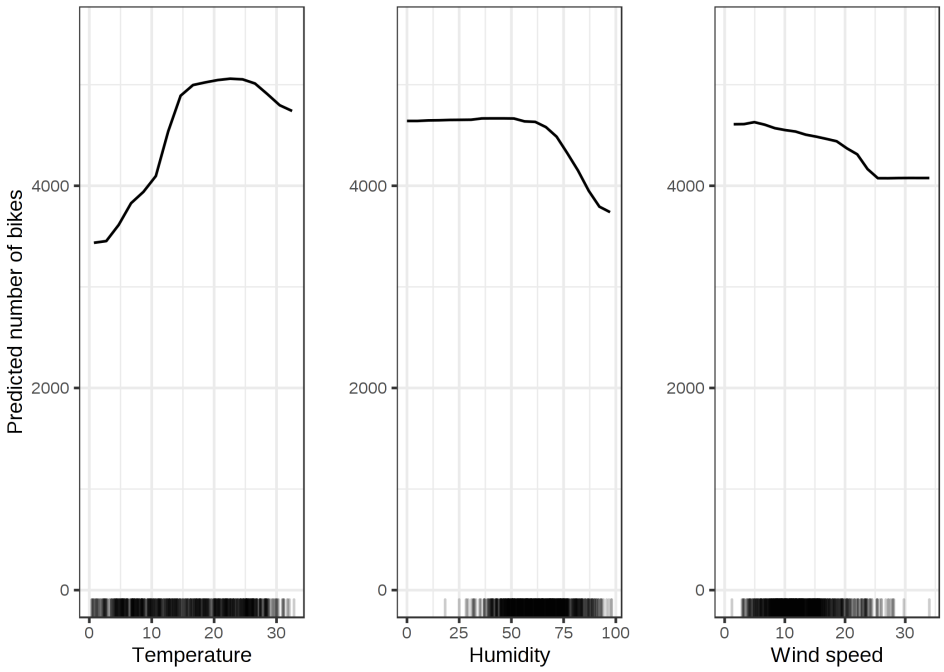


Figure 1.8 – Exemple de diagramme de dépendance partielle.

Chapitre 2

APPLICATION

2.1 Annexes

Annexe .1 : Revue des méthodes explicatives

Name	Ref.	Authors	Year	Problem	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
Trepan	[20]	Craven et al.	1996	Model Expl.	DT	NN	TAB	✓				✓
-	[50]	Krishnan et al.	1999	Model Expl.	DT	NN	TAB	✓		✓		✓
DecText	[9]	Boz	2002	Model Expl.	DT	NN	TAB	✓	✓			✓
GPDT	[39]	Johansson et al.	2009	Model Expl.	DT	NN	TAB	✓	✓	✓		✓
Tree Metrics	[14]	Chipman et al.	1998	Model Expl.	DT	TE	TAB					✓
CCM	[23]	Domingos et al.	1998	Model Expl.	DT	TE	TAB	✓	✓			✓
-	[29]	Gibbons et al.	2013	Model Expl.	DT	TE	TAB	✓	✓			
STA	[114]	Zhou et al.	2016	Model Expl.	DT	TE	TAB		✓			
CDT	[87]	Schetinin et al.	2007	Model Expl.	DT	TE	TAB			✓		
-	[32]	Hara et al.	2016	Model Expl.	DT	TE	TAB		✓	✓		✓
TSP	[94]	Tan et al.	2016	Model Expl.	DT	TE	TAB					✓
Conj Rules	[19]	Craven et al.	1994	Model Expl.	DR	NN	TAB		✓			
G-REX	[37]	Johansson et al.	2003	Model Expl.	DR	NN	TAB	✓	✓	✓		
REFNE	[115]	Zhou et al.	2003	Model Expl.	DR	NN	TAB	✓	✓	✓		✓
RxREN	[6]	Augasta et al.	2012	Model Expl.	DR	NN	TAB		✓	✓		✓
SVM+P	[70]	Nunez et al.	2002	Model Expl.	DR	SVM	TAB			✓		✓
-	[28]	Fung et al.	2005	Model Expl.	DR	SVM	TAB			✓		✓
inTrees	[22]	Deng	2014	Model Expl.	DR	TE	TAB			✓		✓
-	[61]	Lou et al.	2013	Model Expl.	FI	AGN	TAB	✓		✓	✓	✓
GoldenEye	[33]	Henelius et al.	2014	Model Expl.	FI	AGN	TAB	✓	✓	✓	✓	✓
PALM	[51]	Krishnan et al.	2017	Model Expl.	DT	AGN	ANY	✓		✓		✓
-	[97]	Tolomei et al.	2017	Model Expl.	FI	TE	TAB			✓		
-	[108]	Xu et al.	2015	Outcome Expl.	SM	DNN	IMG			✓	✓	✓
-	[25]	Fong et al.	2017	Outcome Expl.	SM	DNN	IMG			✓		✓
CAM	[113]	Zhou et al.	2016	Outcome Expl.	SM	DNN	IMG			✓	✓	✓
Grad-CAM	[89]	Selvaraju et al.	2016	Outcome Expl.	SM	DNN	IMG			✓	✓	✓
-	[56]	Lei et al.	2016	Outcome Expl.	SM	DNN	TXT			✓		✓
LIME	[83]	Ribeiro et al.	2016	Outcome Expl.	FI	AGN	ANY	✓	✓	✓	✓	✓
MES	[98]	Turner et al.	2016	Outcome Expl.	DR	AGN	ANY	✓		✓		✓
NID	[71]	Olden et al.	2002	Inspection	SA	NN	TAB			✓		
GDP	[7]	Baehrens	2010	Inspection	SA	AGN	TAB	✓		✓		✓
IG	[92]	Sundararajan	2017	Inspection	SA	DNN	ANY			✓		✓
VEC	[16]	Cortez et al.	2011	Inspection	SA	AGN	TAB	✓		✓		✓
VIN	[35]	Hooker	2004	Inspection	PDP	AGN	TAB	✓		✓		✓
ICE	[30]	Goldstein et al.	2015	Inspection	PDP	AGN	TAB	✓		✓	✓	✓
Prospector	[48]	Krause et al.	2016	Inspection	PDP	AGN	TAB	✓		✓		✓
Auditing	[2]	Adler et al.	2016	Inspection	PDP	AGN	TAB	✓		✓	✓	✓
OPIA	[1]	Adebayo et al.	2016	Inspection	PDP	AGN	TAB	✓		✓		
-	[110]	Yosinski et al.	2015	Inspection	NA	DNN	IMG			✓		✓
TreeView	[96]	Thiagarajan et al.	2016	Inspection	DT	DNN	TAB			✓		✓
IP	[90]	Shwartz et al.	2017	Inspection	NA	DNN	TAB			✓		
-	[81]	Radford	2017	Inspection	NA	DNN	TXT			✓		
CPAR	[109]	Yin et al.	2003	Transp. Design	DR	-	TAB					✓
FRL	[102]	Wang et al.	2015	Transp. Design	DR	-	TAB			✓	✓	✓
BRL	[57]	Letham et al.	2015	Transp. Design	DR	-	TAB			✓		
TLBR	[91]	Su et al.	2015	Transp. Design	DR	-	TAB			✓		✓
IDS	[53]	Lakkaraju et al.	2016	Transp. Design	DR	-	TAB			✓		
Rule Set	[104]	Wang et al.	2016	Transp. Design	DR	-	TAB			✓	✓	✓
1Rule	[64]	Malioutov et al.	2017	Transp. Design	DR	-	TAB			✓		✓
PS	[8]	Bien et al.	2011	Transp. Design	PS	-	ANY			✓		✓
BCM	[44]	Kim et al.	2014	Transp. Design	PS	-	ANY			✓		✓
-	[63]	Mahendran et al.	2015	Transp. Design	PS	-	IMG			✓	✓	✓
-	[47]	Kononenko et al.	2010	Transp. Design	FI	-	TAB			✓		✓
OT-SpAMs	[103]	Wang et al.	2015	Transp. Design	DT	-	TAB			✓	✓	✓

Figure 1 – Tableau résumant l'ensemble des méthodes expliquant les boites noires présent dans la littérature. Description en annexe 2. Tiré de [6]

Annexe .2 : Description des méthodes explicatives

Feature	Description
Problem	Model Explanation, Outcome Explanation, Black Box Inspection, Transparent Design
Explanator	DT - Decision Tree, DR - Decision Rules, FI - Features Importance, SM - Saliency Masks, SA - Sensitivity Analysis, PDP - Partial Dependence Plot, NA - Neurons Activation, PS - Prototype Selection
Black Box	NN - Neural Network, TE - Tree Ensemble, SVM - Support Vector Machines, DNN - Deep Neural Network, AGN - AGNostic black box
Data Type	TAB - TABular, IMG - IMaGe, TXT - TeXT, ANY - ANY type of data
General	Indicates if an explanatory approach can be generalized for every black box, i.e., it does not consider peculiarities of the black box to produce the explanation
Random	Indicates if any kind of random perturbation or permutation of the original dataset is required for the explanation
Examples	Indicates if example of explanations are shown in the paper
Code	Indicates if the source code is available
Dataset	Indicates if the datasets used in the experiments are available

Figure 2 – Description du tableau présent en annexe 1. Tiré de [6]

BIBLIOGRAPHIE

[1] Seth Flaxman Bryce Goodman. *European Union regulations on algorithmic decision-making and a "right to explanation"*. 2016.

[2] Tim Miller. *Explanation in Artificial Intelligence : Insights from the Social Sciences*. 2018.

[3] Christoph Molnar. *Interpretable Machine Learning*. 2019.

[4] C.Guestrin MT.Ribeiro S.Singh. *"Why Should I Trust You ?" Explaining the Predictions of Any Classifier*. 2016.

[5] Yarin Gal Piotr Dabkowski. *Real Time Image Saliency for Black Box Classifiers*. 2017.

[6] S.Ruggieri R.Guidotti A.Monreale. *A Survey Of Methods For Explaining Black Box Models*. 2018.

[7] Andrew D. Selbst Solon Barocas. *Big Data's Disparate Impact*. 2016.

[8] KR.Müller W.Samek T.Wiegand. *Explainable Artificial Intelligence : Understanding, Visualizing and Interpreting Deep Learning Models*. 2017.

TABLE DES FIGURES

Université Paris-Nanterre
200 Avenue de la République
92000 Nanterre