tech SPRINT

30.06 › 01.07

ACPR | Banque de France

DÉFI EXPLICABILITÉ

GUIDE DE L'ANALYSTE

Attendus du défi & documents utiles

tech SPRINT

30.06 › 01.07

ACPR | Banque de France

DÉFI EXPLICABILITÉ

Ensemble, contribuons à une IA de confiance.

PRÉAMBULE

NOTE AUX ANALYSTES

Ce guide contient autant d'informations que possible afin de vous préparer au Tech Sprint. Les Animateurs de l'ACPR seront également disponibles le jour du défi pour vous guider et vous prodiguer conseils et aide en cas de blocage.

Son objectif :

- Mettre toutes les équipes d'Analystes sur un pied d'égalité
- Permettre aux moins avertis en matière de Machine Learning et de modèles de risques de crédit, de se former rapidement aux méthodologies d'explication du ML
- Proposer l'ensemble des librairies utiles pour le Défi (à télécharger en amont sur vos machines).
- Rappeler les attendus du Défi d'explicabilité
- Eclairer les Analystes non spécialistes du domaine sur les attendus généraux d'une explication algorithmique (qu'est-ce qu'une bonne explication, en fonction de l'audience notamment)
- Exposer les critères généraux sur lesquels seront évalués vos travaux.

SOMMAIRE

Attendus du Défi d'explicabilité

Liens et références

DÉFI EXPLICABILITÉ

PRÉREQUIS DU DÉFI

NÉCESSAIRE "MUST HAVE"
Interrogation des API

UTILE "NICE-TO-HAVE"
Génération de données synthétiques

SUPERFLUS "NEED-NOT-HAVE"
Conception de modèles de risque de crédit
Réglementation des modèles de risque

1

EXPLIQUER LES MODÈLES
de risque de crédit basés sur du Machine Learning.

ÉCLAIRER LES ENJEUX RÉGLEMENTAIRES
à la lumière de ce qui est faisable techniquement en explicabilité du ML

DÉFI EXPLICABILITÉ

PRÉREQUIS NÉCESSAIRE
"MUST-HAVE"

Interrogation d'API

Etre en mesure d'envoyer les requêtes aux APIs donnant accès aux modèles en boîte noire.

RISQUE

Passer beaucoup de temps sur l'écriture du code d'envoi des requêtes (et de récupération et parsing des réponses).
Ceci n'est pas l'objet du Défi.

Certes, souvent quelques lignes de code suffisent, mais si on n'a jamais pratiqué l'appel à des APIs de type REST on ne sera pas forcément à l'aise, on hésitera sur le langage à choisir (ligne de commande curl, script Python, etc.).

NOTRE RECOMMANDATION

Arriver le jour J en sachant :

Construire et manipuler du JSON (sachant qu'il s'agit d'objets simples, avec peu ou pas de *nested objects*)[1]

Envoyer une requête http unique à un *endpoint* POST en mode synchrone[2]

Envoyer, en mode batch, une série de requêtes http à un *endpoint* POST en mode synchrone[3]

Récupérer, stocker puis analyser les réponses aux requêtes[4].

POUR VOUS AIDER

[1] **Tutoriel vidéo**
Manipulation de JSON en Python.

[2] **Tutoriel**
Commande curl

[3] **Convertisseur de curl**
vers les langages les plus courants.

[4] Compétences généralistes de programmation auxquelles un tutoriel ne saurait se substituer (comme la génération de données synthétiques décrite dans la suite.)

DÉFI EXPLICABILITÉ

PRÉREQUIS UTILE
"NICE-TO-HAVE"

Génération de données synthétiques

Pouvoir générer vos propres données afin d'éprouver les modèles.

Bien que les Concepteurs fournissent leurs propres données (« Données du Défi »), ces jeux de test ne seront pas forcément suffisants pour alimenter les méthodes explicatives réutilisées ou inventées par votre équipe.

RISQUE

Perdre du temps le jour J à vous documenter sur les librairies existantes de génération de données synthétiques.

NOTRE RECOMMANDATION

Dans votre boîte à outils, arriver le jour J avec les capacités suivantes, par ordre de difficulté (et d'utilité) croissant :

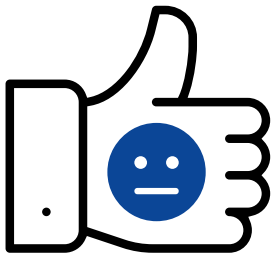
Démultiplier un jeu de données de départ en y apportant quelques modifications, aléatoires ou non.

Créer from scratch des données possédant des propriétés statistiques attendues

ceci n'est pas strictement nécessaire.

DÉFI EXPLICABILITÉ

PRÉREQUIS SUPERFLUS
"NEED-NOT-HAVE"



Comprendre comment les modèles ont été construits.

Le Tech Sprint vise à expliquer les modèles de risque de crédit basés sur du ML. Il ne présuppose pas de comprendre comment ils ont été construits.

Comprendre et maîtriser le cadre réglementaire et juridique des modèles de risque de crédit.

Le Tech Sprint vise à éclairer les enjeux réglementaires à la lumière de ce qui est faisable techniquement en explicabilité du ML, et pas l'inverse.

Les modèles proposés aux Défis sont supposés être réalistes et représentatifs des systèmes utilisés par les établissements de crédit français.

Des explications, pour qui, pour quoi ?

POUR QUI ?

les concepteurs des modèles proposés

mais aussi :

- les équipes en charge de leur surveillance et de leur bon fonctionnement,
- les agents de conformité,
- les consommateurs et clients finaux,
- les auditeurs internes et externes...

POUR QUOI ?

Informer les diverses parties prenantes (industrie, autorités de régulation et supervision, consommateurs) de :

- la réalité opérationnelle des modèles de ML en risque de crédit
- des difficultés associées à leur compréhension, qui pourront être mises en perspective de la réglementation existante voire même influer sur l'évolution réglementaire [5].

[5] À noter que le projet de réglementation de l'IA publié par la Commission européenne le 26 avril dernier place explicitement ces types de modèles (et non pas le secteur bancaire dans son ensemble) parmi les applications de l'IA à haut risque : l'enjeu est donc de taille !

DÉFI **EXPLICABILITÉ**

OBJECTIFS GÉNÉRAUX

OBJECTIF PRINCIPAL
Interprétabilité | Comprendre le fonctionnement des modèles

BONUS (si le temps le permet)
Mesurer l'équité algorithmique

NON-OBJECTIFS
Mesurer la performance des modèles
Mesurer le bien-fondé des prédictions

2

EXPLICABILITÉ

l'explicabilité recouvre généralement les notions de :
interprétabilité : lie la transparence et englobe intelligibilité et justification,
accountability : responsabilité
fairness (ou équité) : étude des biais discriminatoires et autres enjeux éthiques.

DÉFI EXPLICABILITÉ

OBJECTIF PRINCIPAL

Faire comprendre le fonctionnement des modèles proposés

L'explication algorithmique consiste à faire comprendre le fonctionnement de l'algorithme et à expliquer pourquoi une prédiction est produite (ou dans le cas de modèles décisionnels, pourquoi une décision est prise).

Quelles lignes de questionnement adopter ?

Selon le document de reflexion de l'ACPR*, les lignes de questionnement auxquelles vise à répondre une explication sont les suivantes :

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

"Une bonne explication est une explication adaptée à son destinataire"

LES ATTENDUS

VOUS VOUS INTÉRESSEREZ
- **à l'interprétabilité**
- **à l'analyse des biais***

**si le temps le permet*

VOUS NE VOUS INTÉRESSEREZ PAS AUX ASPECTS :
- éthiques
- d'accountability

EXPLICATION DE **NIVEAU 1**

OBSERVATION

Elle répond à la question :

"Que fait l'algorithme?"
- angle technique-

"A quoi sert l'algorithme?"
- angle fonctionnel-

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 ?"

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

par la génération de l'algorithme lui-même de justification obtenues par apprentissage

1 2 3 4

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,

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, 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 utilisées,
- et un **examen** de l'ensemble des paramètres du modèle

MÉCANISME D'EXPLICABILITÉ

LES 4 NIVEAUX D'EXPLICATION*

GOUVERNANCE DES ALGORITHMES D'INTELLIGENCE ARTIFICIELLE DANS LE SECTEUR FINANCIER

Laurent Dupont, Olivier Fliche, Su Yang
Pôle Fintech-Innovation, ACPR
Juin 2020

CONSULTER →

DÉFI EXPLICABILITÉ

MÉCANISME D'EXPLICABILITÉ RECHERCHÉE

Devoir englober les concepts d'intelligibilité, de justification, d'approximation et de réplication

Intelligibilité
comprendre comment fonctionne un modèle

Justification
déterminer si les prédictions sont en phase avec des normes ou standards, explicites ou non, internes ou externes à l'organisation.
Ex: critères d'équité algorithmique et d'absence de biais discriminatoires.

Approximation et Réplication
pouvoir reproduire un modèle de façon approximative ou exacte afin d'en mieux saisir la construction.

Ces définitions correspondent en gros aux 4 niveaux d'explication proposés par le document de l'ACPR sur la gouvernance de l'IA*.

OBJET DE L'EXPLICATION

Donner à comprendre autant que possible :

1. **Le modèle lui-même**
2. **Les données sur lesquelles il a été entraîné**
   A savoir : volumétrie, caractéristiques statistiques, anomalies, points ou sous-populations d'intérêt, etc.
3. **Son processus de construction**
   C'est l'esprit du reverse engineering : inférer non seulement la classe d'algorithme de ML, mais aussi ses hyperparamètres et autres éléments de configuration, toute particularité du modèle donné, mais idéalement aussi le langage de programmation dans lequel il a été implémenté...

MÉCANISME D'EXPLICABILITÉ

LES 4 NIVEAUX D'EXPLICATION*

GOUVERNANCE DES ALGORITHMES D'INTELLIGENCE ARTIFICIELLE DANS LE SECTEUR FINANCIER

Laurent Dupont, Olivier Fliche, Su Yang
Pôle Fintech-Innovation, ACPR
Juin 2020

CONSULTER →

LE BONUS

Si votre équipe est parvenue à remplir l'objectif principal du Tech Sprint et il qu'il vous reste du temps

Mesurer l'équité algorithmique (ou fairness)

Sur certains modèles, vous pourrez :

définir les biais de nature problématique ou pas (biais de classification ou de prédiction, ou biais statistiques non souhaités déjà présents dans les données) ;

caractériser et quantifier ces biais, par des métriques ou des méthodes explicatives appropriées ;

déterminer dans quelle mesure les biais présents dans les données sont reflétés, voire renforcés, par les modèles de Machine Learning.

La priorité de cet objectif est moindre.

L'équité algorithmique pourrait en revanche faire l'objet d'un événement futur [6].

[6] *Ce défi secondaire peut d'ailleurs s'avérer encore plus difficile à relever que celui d'explicabilité, c'est pourquoi il est secondaire. En effet, peu de variables sensibles auront été collectées (voire aucune selon les modèles) et donc il ne sera pas question de détecter des biais à caractère discriminatoire, mais des biais statistiques à connotation neutre. De plus, la variable-cible ne sera pas toujours disponible dans les jeux de test (et par définition elle ne sera pas disponible dans les données générées par les participants), donc les métriques d'équité pertinentes seront dans la catégorie de la parité démographique plutôt que celle des mesures de taux d'erreur.*

DÉFI EXPLICABILITÉ

NON-OBJECTIFS DU TECH SPRINT

mesurer la performance des modèles

Il ne s'agira pas d'évaluer la performance des modèles prédictifs (par exemple si les estimations de probabilité de défaut à un an sont proches de la réalité une fois l'année écoulée).

mesurer le bien-fondé des prédictions

Il ne s'agira pas non plus de déterminer le bien-fondé des prédictions que les modèles produisent (par exemple si une décision de refus de crédit sur la base d'une prédiction est contestable).

A NOTER

Etant donné ces objectifs, **les valeurs de la variable-cible** sont fournies dans certains jeux de données de test, mais uniquement afin d'évaluer l'équité des prédictions et non leur bien-fondé ou la performance des modèles.

DÉFI EXPLICABILITÉ

RESTITUTION & LIVRABLES

Restitution | 5 minutes pour convaincre

Votre équipe fournira le lendemain du défi une restitution orale (5 minutes max.) devant le Jury et l'ensemble de l'audience et des invités du Tech Sprint.

Cette présentation peut inclure : des **slides** (remis la veille à la fin des travaux) **et/ou une démo** (qui devra donc être rigoureusement préparée et minutée).

Livrables | Adaptez votre explication !

Consommateur, équipe en charge d'utiliser ou de surveiller l'algorithme, auditeur interne ou externe etc., les profils de ses destinataires sont multiples : **Adaptez toujours votre explication à celui à qui elle s'adresse !**

3

A votre code !

IMPORTANT !
Vos livrables seront à remettre
le **30 juin à 18h00** au plus tard

LIVRABLES

Une bonne explication est une explication adaptée à son destinataire.

Vous êtes libres de ne proposer qu'un type d'explications de combiner plusieurs types. Le nombre relatif d'explications à fournir pour chaque modèle n'est pas prescrit, il est fonction de l'approche que votre équipe aura retenue et pourra varier entre les cas – certes extrêmes – d'une seule explication globale et d'une (voire plusieurs !) explications par point de donnée.

Le Rapport d'explicabilité
Ce modèle vise à fournir une description objective, selon un ensemble de caractéristiques. Il peut être rempli une seule fois, ou autant de fois qu'un type d'explication donné a été utilisé dans vos travaux.

Format : modèle PDF éditable remis à votre équipe

Une Annexe (optionnelle au rapport)
Elle est destinée à fournir des informations complémentaires sur les travaux menés, la méthodologie employée et les accomplissements (si le modèle fourni n'est pas suffisant).
Elle peut inclure notamment inclure toute sorte de visualisation, graphique ou autre.

Format : Libre. PDF recommandé

Votre support de présentation
Il sera utilisé le lendemain lors de la présentation devant le Jury. Votre équipe pourra également réaliser une démo interactive

Format : version PDF de vos slides / démo interactive (si souhaité)

DÉFI EXPLICABILITÉ

CRITÈRES D'APPRÉCIATION

4

Eclairez le jury !

CRITÈRES D'APPRÉCIATION

Le règlement du Tech Sprint précise que l'appréciation du jury portera sur les points suivants :

1. **les accomplissements techniques** contenus dans les travaux menés par l'Équipe d'Analystes

2. **le caractère innovant des travaux**
   - au plan scientifique
   - au plan méthodologique

3. **le caractère clair, pédagogique, utile et juste de la restitution** tant du point de vue des métiers, des auditeurs, que des personnes en charge du suivi ou de la maintenance des systèmes

4. **la qualité des explications elles-mêmes** qui pourra s'évaluer via :
   - d'une part leur caractère compréhensible (simplicité et pertinence vis-à-vis des besoins du destinataire)
   - d'autre part leur fidélité et complétude vis-à-vis du modèle

5. **la contribution des travaux aux enjeux métier** du risque de crédit :
   - compréhension de l'estimation du risque de défaut opérée par divers modèles prédictifs
   - assistance à l'interprétation de leurs sorties par des experts métier
   - aide à leur maintenance, etc.

6. **la contribution des travaux à l'éclairage des enjeux réglementaires** en termes de maîtrise des risques, de gouvernance et de protection de la clientèle.

Bonne chance à toutes et à tous !

DÉFI EXPLICABILITÉ

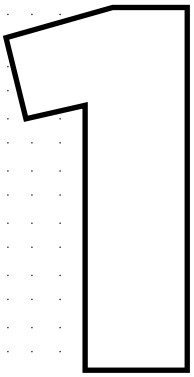
ANNEXES
LIENS & RÉFÉRENCES

La lecture de ces références est totalement facultative :)

tech SPRINT
30.06 › 01.07
ACPR | Banque de France

ANNEXE 1

MODÈLES DE RISQUE DE CRÉDIT



MACHINE LEARNING EXPLAINABILITY IN FINANCE: AN APPLICATION TO DEFAULT RISK ANALYSIS

—

Ce rapport de la Banque d'Angleterre porte spécifiquement sur les modèles d'estimation de probabilité de défaut qui font l'objet du Tech Sprint.

44 pages, août 2019 (anglais)

CONSULTER →

INTERPRETABLE CREDIT APPLICATION PREDICTIONS WITH COUNTERFACTUAL EXPLANATIONS

—

article sur l'explicabilité des modèles de risque de crédit qui ajoute aux approches standard :

• d'une part la production de contre-factuelles positives (expliquant les décisions d'octroi de prêt plutôt que de refus),
• d'autre part des méthodes de pondération des variables prédictives permettant de réduire la longueur des explications produites.

CONSULTER →

MACHINE LEARNING: CONSIDERATIONS FOR FAIRLY AND TRANSPARENTLY EXPANDING ACCESS TO CREDIT.

—

- concerne les modèles de risque de crédit, avec la double perspective d'explicabilité et d'équité algorithmiques
- aborde des points rarement soulevés, encore moins dans ce scénario spécifique.
- sont définies, au-delà de la simple dichotomie entre explications locales et globales, des nuances de localité : la «data locality» et la «value locality»

29 pages, juillet 2020 (anglais)

CONSULTER →

Cette annexe propose quelques références du domaine de la modélisation des risques de crédit, tant sous les aspects **techniques** (mathématique et informatique entre autres), **métiers** (au pluriel : conformité, risques, etc.), que **réglementaires**.

ANNEXE 2

MÉTHODES EXPLICATIVES DES MODÈLES DE ML

Cette annexe propose des références de la littérature de recherche (appliquée). Elle inclut quelques livres en ligne et des papiers de recherche pour les méthodes les plus récentes.

Utilité :
pour les Analystes :
Avoir un aperçu de l'ensemble des méthodes standard existantes (éventuellement de les réutiliser telles quelles, voire de les adapter à leurs besoins) ;

pour les membres du Jury :
Pouvoir évaluer l'innovation et la créativité des méthodes utilisées par les Analystes

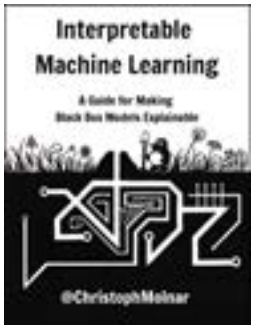
LIVRES EN LIGNE

ARTICLES

MÉTHODES D'EXPLICATION

MÉTHODES EXPLICATIVES INDÉPENDANTES DU MODÈLE

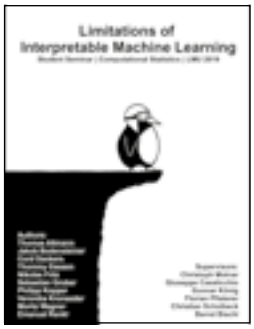
LIVRES EN LIGNE



INTERPRETABLE MACHINE LEARNING
A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE

Le livre de référence de Christophe Molnar

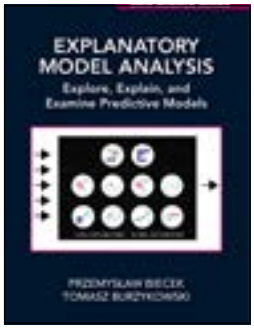
CONSULTER →



LIMITATIONS OF INTERPRETABLE MACHINE LEARNING METHODS

Ce livre, très récent et plus avancé, explique les limites du machine learning interprétable

CONSULTER →



EXPLANATORY MODEL ANALYSIS
EXPLORE, EXPLAIN AND EXAMINE PREDICTIVE MODELS

L'ouvrage assez complet de Przemyslaw Biecek et Tomasz Burzykowski, avec des exemples en R et Python.

CONSULTER →

ARTICLES

Les articles suivants proposent une analyse des méthodes explicatives et de ce qui constitue une bonne explication.

EXPLANATION IN HUMAN-AI SYSTEMS:
A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI

CONSULTER →

EXPLAINING EXPLANATIONS:
An Overview of Interpretability of Machine Learning

CONSULTER →

MÉTHODES D'EXPLICATION

DOCUMENTATION DES JEUX DE DONNÉES

MODEL CARDS FOR MODEL REPORTING

CONSULTER →

AEQUITAS:
a bias and fairness audit toolkit

CONSULTER →

MÉTHODES EXPLICATIVES INDÉPENDANTES DU MODÈLE

Les approches d'explicabilité sont souvent classées en deux catégories :

L'explicabilité globale tend à fournir une explication sur le comportement global du modèle,

l'explication locale donne une explication pour une prédiction précise.

Explications locales post-hoc

PROTODASH
Gurumoorthy et al., 2019

CONSULTER →

CONTRASTIVE EXPLANATIONS METHOD
Dhurandhar et al., 2018

CONSULTER →

CONTRASTIVE EXPLANATIONS METHOD WITH MONOTONIC ATTRIBUTE FUNCTIONS
Luss et al., 2019

CONSULTER →

LIME
Ribeiro et al. 2016, Github

CONSULTER →

SHAP
Lundberg, et al. 2017, Github

CONSULTER →

Explications globales post-hoc

PROFWEIGHT
Dhurandhar et al., 2018

CONSULTER →

Explications locales conjointes à la modélisation

TEACHING AI TO EXPLAIN ITS DECISIONS
Hind et al., 2019

CONSULTER →

Explications globales conjointes à la modélisation

BOOLEAN DECISION RULES VIA COLUMN GENERATION (LIGHT EDITION)
Dash et al., 2018

CONSULTER →

GENERALIZED LINEAR RULE MODELS
Wei et al., 2019

CONSULTER →

Métriques d'explicabilité

FAITHFULNESS
Alvarez-Melis et Jaakkola, 2018

CONSULTER →

MONOTONICITY
Luss et al., 2019

CONSULTER →

COMPLETENESS
Sundarajan et al., 2017

CONSULTER →

SENSITIVITY-N
Ancona et al., 2018

CONSULTER →

Métriques d'importance des variables

1 | Métriques les plus courantes

PERMUTATION FEATURE IMPORTANCE
Christoph Molnar, 2021

CONSULTER →

VARIABLE IMPORTANCE
Max Kuhn, 2019

CONSULTER →

2 | Métriques spécifiques à un modèle

MODÈLES LINÉAIRES
la valeur absolue de la t-statistique de chaque paramètre du modèle peut être utilisée

FORÊTS ALÉATOIRES
ce papier de recherche mentionne de nombreuses métriques applicables dans ce cas (*Cross Entropy-Information Gain*, mesure de Gini, *Mean Squared Error*, différence de précision après permutation, la méthode PIMP, *Recurrent relative variable importance*, *Recursive feature elimination*, etc.)

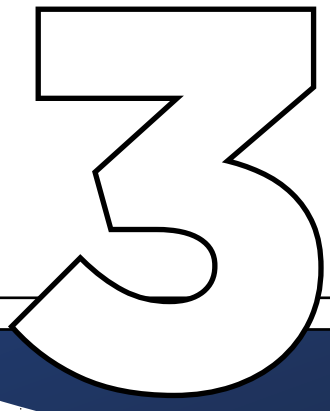
CONSULTER →

ANNEXE 3

TUTORIELS & MODÈLES

Cette annexe propose des liens vers :

- **des tutoriels incluant code (souvent sous forme de notebooks)**
- **des exemples de cas d'usage**
- **mais aussi des modèles d'analyse d'explicabilité**



TUTORIELS COMPLETS

INTERPRETING MACHINE LEARNING MODELS WITH THE IML PACKAGE
Code R

CONSULTER →

INTERPRETABLE MACHINE LEARNING WITH PYTHON
Notebooks Python

CONSULTER →

MODEL INTERPRETABILITY WITH DALEX
Librairie DALEX en R

CONSULTER →

INTERPRETABLE MACHINE LEARNING USING COUNTERFACTUALS
Librairie Alibi en Python

CONSULTER →

MACHINE LEARNING EXPLAINABILITY
Tutoriel complet par Kaggle Learn

CONSULTER →

PARTIAL DEPENDENCE PLOTS IN R
utilisation d'un package R pour produire des PDP

CONSULTER →

VISUALIZING ML MODELS WITH LIME
utilisation de LIME sous R

CONSULTER →

Tutoriels en trois épisodes
«Explainable AI»

Episode #1
THE IMPORTANCE OF HUMAN INTERPRETABLE MACHINE LEARNING
A brief introduction into human interpretable machine learning and model interpretation

CONSULTER →

Episode #2
MODEL INTERPRETATION STRATEGIES
Learn about model interpretation techniques, limitations and advances

CONSULTER →

Episode #3
HANDS-ON MACHINE LEARNING MODEL INTERPRETATION
A comprehensive guide to interpreting machine learning models

CONSULTER →

MODÈLES D'ANALYSE D'EXPLICABILITÉ

MODEL CARDS

CONSULTER →

EXPLAINABILITY FACT SHEETS:
A Framework for Systematic Assessment of Explainable Approaches

CONSULTER →

ANNEXE 4

LIBRAIRIES, OUTILS & PLATEFORMES D'EXPLICABILITÉ

Cette section propose une liste d'outils, librairies et services (le plus possible en accès libre, voire open source), à la fois génériques et spécifiques à la finance.

Seules sont couvertes les ressources pertinentes pour le Tech Sprint sur l'explicabilité des modèles de risque de crédit

En particulier, sont exclus :

- les domaines du NLP et de la Computer Vision ;
- les outils et librairies dédiés à des méthodes spécifiques à ces domaines, par exemple les CNN ou réseaux neuronaux convolutifs ;
- les méthodes applicables uniquement en boîte blanche (par exemple qui nécessitent un modèle PyTorch ou TensorFlow).

EN PYTHON

EN R

AUTRES LANGAGES

DANS LE NAVIGATEUR

PRODUCTION LOCALE

Interprétabilité des modèles

AI EXPLAINABILITY 360
cette librairie propose par ailleurs un arbre de décision pour choisir la méthode explicative appropriée

CONSULTER →

ALEPYTHON
package Python pour produire des ALE

CONSULTER →

ALIBI
librairie Python pour l'inspection et l'interprétation de modèles de ML

CONSULTER →

LIME
implémentation de LIME en Python

CONSULTER →

ANCHOR
implémentation de Anchors en Python

CONSULTER →

CONTRASTIVE EXPLANATION (FOIL TREES) CONTRASTIVE
implémentation en Python de «Contrastive Explanations with Local Foil Trees»

CONSULTER →

DALEX
librairie modèle-agnostique correspondant au livre Explanatory Model Analysis

CONSULTER →

DICE
implémentation en Python de l'article Diverse Counterfactual Explanations

CONSULTER →

ELI5
librairie Python d'inspection et d'explication de classificateurs, y compris des fonctionnalités «boîte noire»

CONSULTER →

FACET
librairie Python d'inspection et de simulation de modèles de ML

CONSULTER →

PYBREAKDOWN
implémentation en Python de la librairie breakDown

CONSULTER →

SAGE
implémentation de la méthode SAGE

CONSULTER →

SALIB
implémentation en Python de méthodes d'analyse de sensibilité

CONSULTER →

SHAP
Implémentation des articles de Lundberg

CONSULTER →

SHAPLEY
implémentation de calculs exacts et approchés des valeurs de Shapley

CONSULTER →

SHAPLEY FLOW
implémentation en Python de la méthode décrite dans Shapley Flow: A Graph-based Approach to Interpreting Model Predictions, technique plus avancée que les précédentes qui prend en compte l'ensemble du graphe de causalité dans la production de valeurs de SHAP)

CONSULTER →

SKATER
projet open source d'Oracle, en beta, d'interprétation de boîtes noires

CONSULTER →

XAI
développé par l'Institute for Ethical AI & ML

CONSULTER →

Analyse des jeux de données

LOFO-IMPORTANCE
implémentation en Python de la méthode LOFO - Leave One Feature Out

CONSULTER →

Analyse des biais de modèles

AI FAIRNESS 360

CONSULTER →

AEQUITAS

CONSULTER →

BLACKBOXAUDITING
audit de boîtes noires basé sur la mesure de *Disparate Impact*

CONSULTER →

FAIRML
boîte à outils Python d'audit de modèles «boîte noire», focalisée sur la fairness mais pas exclusivement

CONSULTER →

FAIRLEARN
librairie Python centrée sur la mesure et la remédiation de défauts d'équité de groupe

CONSULTER →

FAIRNESS_MEASURES_CODE
implémentation en Python des mesures de fairness décrites dans l'article Measuring discrimination in algorithmic decision making

CONSULTER →

PARITY-FAIRNESS
outil avec interface graphique d'investigation des biais de modèle

CONSULTER →

Interprétabilité des modèles

ALEPLOT
implémentation en Python de la méthode LOFO - Leave One Feature Out

CONSULTER →

DRWHYAI
collection d'outils exploratoires, d'analyse et de visualisation

CONSULTER →

DALEX
librairie modèle-agnostique correspondant au livre Explanatory Model Analysis

CONSULTER →

EXPLAINPREDICTION

CONSULTER →

FASTSHAP
implémentation en R d'une approximation des valeurs de Shapley

CONSULTER →

FEATUREIMPORTANCE
implémentation en R des métriques et visualisations décrites dans l'article Visualizing the Feature Importance for Black Box Models

CONSULTER →

FLASHLIGHT

CONSULTER →

IBREAKDOWN
successeur de la librairie breakDown

CONSULTER →

ICEBOX
package R implémentant les diagrammes d'Independent Conditional Explanation

CONSULTER →

IML
package R correspondant au livre *Interpretable Machine Learning* de Christoph Molnar

CONSULTER →

INGREDIENTS
librairie à la base de DALEX citée plus haut

CONSULTER →

LIME

CONSULTER →

LIVE
implémentation en R des méthodes décrites dans l'article «Why Should I Trust You?»: Explaining the Predictions of Any Classifie

CONSULTER →

PDP

CONSULTER →

SHAPFLEX
librairie de calcul de valeurs de Shapley prenant en compte les relations causales entre variables, implémentation de l'article Shapley Decomposition of R-Squared in Machine Learning Models

CONSULTER →

SHAPPER
port en R de la librairie Python shap

CONSULTER →

TREESHAP
implémentation en R de la méthode TreeShap, par ailleurs disponible en Python dans la librairie Alibi

CONSULTER →

Analyse des biais de modèles

AIF360
implémentation en R de la librairie AI Fairness 360

CONSULTER →

FAIRNESS
implémentation en R de nombreuses métriques de fairness issues des publications à l'état de l'art

CONSULTER →

FAIRMODELS
propose le calcul de métriques de fairness simples basées sur la matrice de confusion d'un attribut sensible

CONSULTER →

AUTRES LANGAGES

LIFT
librairie en Scala/Spark mais sa documentation github contient une bonne synthèse des différents types de métriques d'équité.

CONSULTER →

DANS LE NAVIGATEUR

MANIFOLD
outil modèle-agnostique d'inspection visuelle de la performance de modèles

CONSULTER →

TENSORBOARD PROJECTO
visualisation de jeux de données par réduction de dimension

CONSULTER →

WHAT-IF TOOL
visualisation du comportement de modèles de ML

CONSULTER →

PRODUCTION LOCALE

Outils liés à l'explicabilité et l'interprétabilité, conçus, implémentés et open-sourcés par des acteurs du secteur financier français

SHAPASH
développé par la MAIF

CONSULTER →

SKOPE-RULES
développé par BPCE

CONSULTER →

ANNEXE 5

REVUE DE LITTÉRATURE

Cette section propose une liste d'articles issus de la recherche récente (2019-2021) en vue de :

illustrer l'état de l'art des méthodes explicatives

exposer les limites aux méthodes explicatives classiques

donner à connaître des réflexions issues d'acteurs français du secteur financier et d'institutions académiques françaises.

5

ÉTAT DE L'ART
DES MÉTHODES EXPLICATIVES

FACE :
Feasible and Actionable Counterfactual Explanations

CONSULTER →

EXPLAINABILITY FACT SHEETS:
A Framework for Systematic Assessment of Explainable Approaches

CONSULTER →

FAT FORENSICS:
A Python Toolbox for Algorithmic Fairness, Accountability and Transparency

CONSULTER →

bLIMEy:
Surrogate Prediction Explanations Beyond LIME

CONSULTER →

EXPLANATION IN ARTIFICIAL INTELLIGENCE:
Insights from the Social Sciences

CONSULTER →

RANDOMIZED ABLATION FEATURE IMPORTANCE

CONSULTER →

LIRME:
Locally Interpretable Ranking Model Explanation

CONSULTER →

UNDERSTANDING COMPLEX PREDICTIVE MODELS WITH GHOST VARIABLES

CONSULTER →

FEATURE IMPACT FOR PREDICTION EXPLANATION

CONSULTER →

EXPLOITING PATTERNS TO EXPLAIN INDIVIDUAL PREDICTIONS

CONSULTER →

L-SHAPLEY AND C-SHAPLEY:
Efficient Model Interpretation for Structured Data

CONSULTER →

ALGORITHMIC TRANSPARENCY VIA QUANTITATIVE INPUT INFLUENCE:
Theory and Experiments with Learning Systems

CONSULTER →

A UNIFIED APPROACH TO INTERPRETING MODEL PREDICTIONS

CONSULTER →

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX:
Automated Decisions and the GDPR

CONSULTER →

LES LIMITES
DES MÉTHODES EXPLICATIVES CLASSIQUES

THE BOUNCER PROBLEM:
Challenges to Remote Explainability

CONSULTER →

"WHY SHOULD YOU TRUST MY EXPLANATION?"
Understanding Uncertainty in LIME Explanations

CONSULTER →

IBREAKDOWN:
Uncertainty of Model Explanations for Non-additive Predictive Models

CONSULTER →

PRODUCTION LOCALE

Articles récents issus d'acteurs français du secteur financier et d'institutions académiques françaises

X-SHAP:
towards multiplicative explainability of Machine Learning

CONSULTER →

CONCEPT TREE:
High-Level Representation of Variables for More Interpretable Surrogate Decision Trees

CONSULTER →

CONFIDENT INTERPRETATIONS OF BLACK BOX CLASSIFIERS

CONSULTER →

A MULTI-LAYERED APPROACH FOR INTERACTIVE BLACK-BOX EXPLANATIONS

CONSULTER →

tech SPRINT

30.06 › 01.07

ACPR | Banque de France

DÉFI EXPLICABILITÉ

édition juin-juillet 2021

ACPR
BANQUE DE FRANCE

NOUS CONTACTER