Interprétabilité des Modèles de Tarification en Actuariat

Application à l'Assurance Automobile

Franklin FEUKAM KOUHOUE

Actuaire Associé

27 mars 2025

1 INTRODUCTION

De grands défis sociétaux

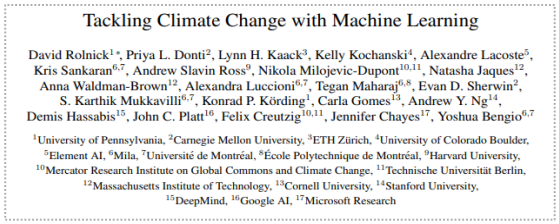
Changement Climatique

- Augmentation de la **fréquence** des évènements climatiques **extrêmes** (inondations, tempêtes, grêle).

Hyperpersonnalisation des services

- Attente des assurés : une offre tarifaire et de garanties adaptées à **leurs besoins** et à leur **profil de risque réel**.

Des réflexions émergentes


Tackling Climate Change with Machine Learning

David Rolnick[1]*, Priya L. Donti[2], Lynn H. Kaack[3], Kelly Kochanski[4], Alexandre Lacoste[5], Kris Sankaran[6,7], Andrew Slavin Ross[9], Nikola Milojevic-Dupont[10,11], Natasha Jaques[12], Anna Waldman-Brown[12], Alexandra Luccioni[6,7], Tegan Maharaj[6,8], Evan D. Sherwin[2], S. Karthik Mukkavilli[6,7], Konrad P. Körding[1], Carla Gomes[13], Andrew Y. Ng[14], Demis Hassabis[15], John C. Platt[16], Felix Creutzig[10,11], Jennifer Chayes[17], Yoshua Bengio[6,7]

[1]University of Pennsylvania, [2]Carnegie Mellon University, [3]ETH Zürich, [4]University of Colorado Boulder, [5]Element AI, [6]Mila, [7]Université de Montréal, [8]École Polytechnique de Montréal, [9]Harvard University, [10]Mercator Research Institute on Global Commons and Climate Change, [11]Technische Universität Berlin, [12]Massachusetts Institute of Technology, [13]Cornell University, [14]Stanford University, [15]DeepMind, [16]Google AI, [17]Microsoft Research


SOA Research Institute

The Insurance Market in the Era of Digital Transitions

Relationships Between Insurers, Big Tech, and Insurtech

AUTHORS Arthur Charpentier, L'Université du Québec à Montréal
Raphaël Suire, Nantes University, IAE—Graduate school of Management, France

SPONSOR Actuarial Innovation and Technology Strategic Research Program Steering Committee

Point d'Attention

Le Manque d'Interprétabilité : un obstacle


Original Research Article

Artificial intelligence and personalization of insurance: Failure or delayed ignition?

Arthur Charpentier[1] and Xavier Vamparys[2]

BIG DATA & SOCIETY

Big Data & Society
January-March: 1–13
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241291817
journals.sagepub.com/home/bds

Sage

1 • PLAN DE LA PRESENTATION

2. Motivation de la recherche d'Interprétabilité

3. Méthodes d'Interprétabilité explorées

4. Résultats du cas d'application

2 | Motivation de la recherche d'interprétabilité

2 • MOTIVATION DE LA RECHERCHE D'INTERPRETABILITE

Pour L'ACTUAIRE

L'Actuaire est le concepteur et le **garant technique** du modèle d'évaluation du risque (modèle de calcul de la **prime d'assurance**)

Les Raisons

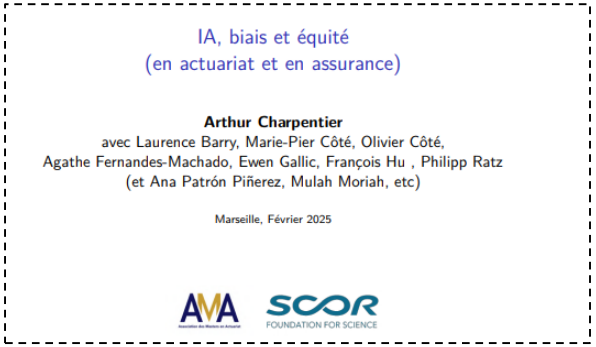
Comprendre le fonctionnement du Modèle

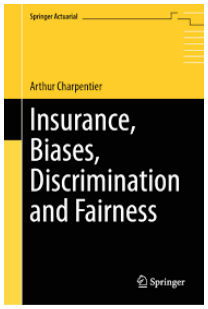
- **Pédagogie interne:** *« Ce que l'on conçoit bien s'énonce clairement, Et les mots pour le dire arrivent aisément. »* **Nicolas Boileau** (1636 - 1711).
- **Mieux maîtriser** le risque (assurance indicielle).

Valider et contrôler les éventuelles risques Modèle

- Détecter des éventuelles **biais d'équité** ou **biais de discrimination tarifaire** dus à l'utilisation des variables sensibles, telles que le **genre** ou **l'origine ethnique**, pour segmenter les assurés.
- Détecter les autres éventuelles **incohérences** et **zones d'instabilité** dans le fonctionnement du modèle.

Pour aller plus loin





2 • MOTIVATION DE LA RECHERCHE D'INTERPRETABILITE

Pour L'ASSUREUR

L'Assureur (ou Compagnie d'assurance) porte la responsabilité globale des décisions à l'issue du processus de tarification.

Les Raisons

S'assurer de la conformité réglementaire

- Exigence de **transparence** de l'ACPR.
- Le règlement européen sur l'IA (AI Act)

Améliorer les stratégies de gestion du risque

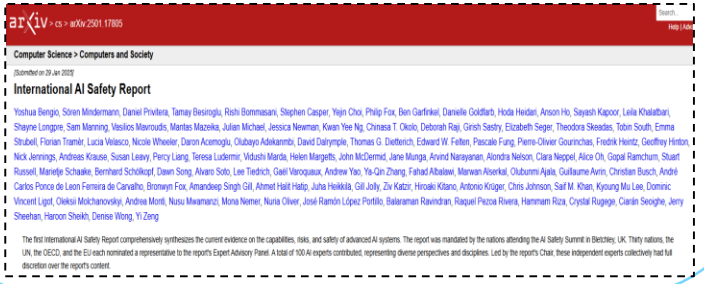
- Comprendre **les raisons d'une hausse de la prime** de l'assuré. Puis fournir à l'assuré des recommandations pour améliorer son profil de risque.
- Identifier des **nouvelles corrélations** (nouveaux facteurs de risques) peu intuitives.

Identifier les profils de risques émergents

Pour aller plus loin


Topics
European Parliament
Topics › Digital › Artificial intelligence › EU AI Act: first regulation on artificial intelligence
EU AI Act: first regulation on artificial intelligence
The use of artificial intelligence in the EU is regulated by the AI Act, the world's first comprehensive AI law. Find out how it protects you.


Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier
Ce document de réflexion s'inscrit dans le cadre des travaux menés par l'ACPR sur l'intelligence artificielle (IA) depuis 2018. En mars 2019, après un premier rapport et une première consultation publique, l'ACPR a lancé des travaux exploratoires avec quelques acteurs du secteur financier afin d'éclairer les enjeux d'explicabilité et de gouvernance de l'IA – au sens essentiellement de *Machine Learning* (ML). Composés d'entretiens et d'ateliers techniques, ils couvraient trois domaines : la lutte contre le blanchiment et le financement du terrorisme (LCB-FT), les modèles internes et en particulier le *scoring* de crédit, et la protection de la clientèle. Deux axes d'étude en sont ressortis : ceux de l'évaluation et de la gouvernance des algorithmes d'IA.


arXiv > cs > arXiv:2501.17805
Computer Science > Computers and Society
International AI Safety Report
Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, Yi Zeng
The first International AI Safety Report comprehensively synthesizes the current evidence on the capabilities, risks, and safety of advanced AI systems. The report was mandated by the nations attending the AI Safety Summit in Bletchley, UK. Thirty nations, the UN, the OECD, and the EU each nominated a representative to the report's Expert Advisory Panel. A total of 100 AI experts contributed, representing diverse perspectives and disciplines. Led by the report's Chair, these independent experts collectively had full discretion over the report's content.

3

Méthodes d'interprétabilité explorées

3 • METHODES D'INTERPRETABILITE EXPLOREES

Dans la littérature, nous distinguons deux grandes approches d'interprétabilité:

- **Interprétabilité dite basée sur le modèle (IBM)**
- **Interprétabilité dite post-hoc**

IBM

- **Parcimonie**
- **Modularité**
- **Simulabilité**

Post hoc

- **Méthodes globales vs Méthodes locales**
- **Méthodes spécifiques au modèle vs Méthodes indépendantes du modèles**
- **Méthodes dépendantes des données vs Méthodes indépendantes des données**

Remarque

Dans le mémoire nous nous sommes focalisés sur l'approche d'interprétabilité dite ***post hoc,*** *et sur les outils* qui sont **indépendants des données** et **indépendants du modèle à expliquer.**

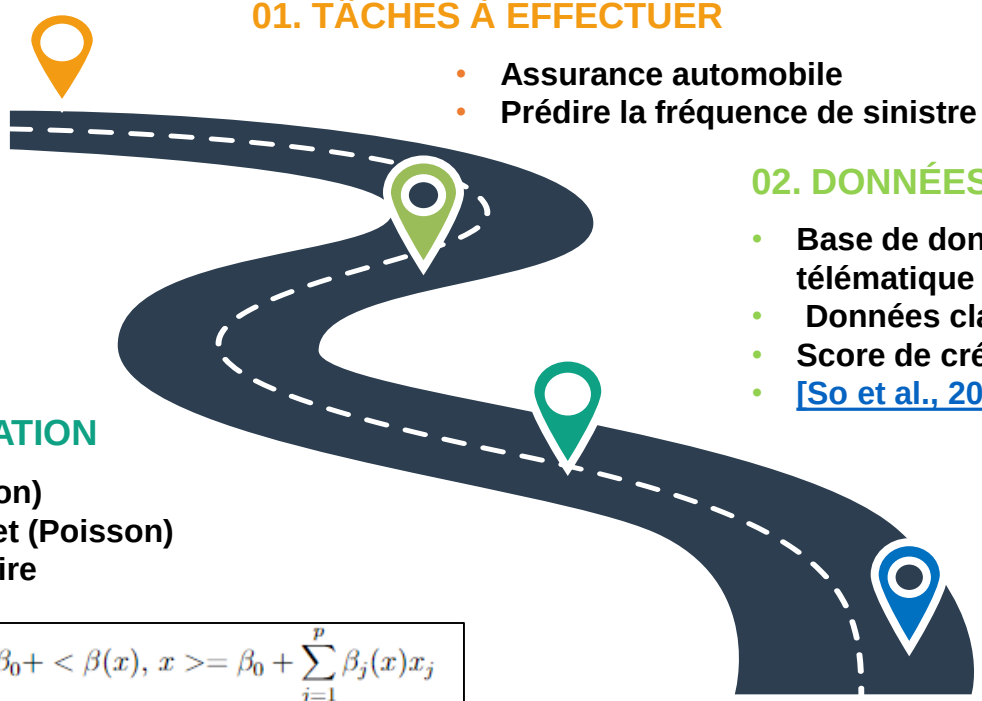
Institut des ACTUAIRES

4 | RÉSULTATS DU CAS D'APPLICATION

Institut

Etapes phares de la mise en oeuvre

Nous présentons ici les principales étapes de la mise en œuvre de ces méthodes d'interprétabilité dans notre contexte de tarification automobile.



01. TÂCHES À EFFECTUER

- **Assurance automobile**
- **Prédire la fréquence de sinistre**

02. DONNÉES DISPONIBLES

- **Base de données synthétiques pour la télématique des conducteurs.**
- **Données classiques et télématiques**
- **Score de crédit**
- **[So et al., 2021], *Université du Connecticut***

03. MODÉLISATION

- **GLM (Poisson)**
- **LocalGLMnet (Poisson)**
- **Forêt aléatoire**

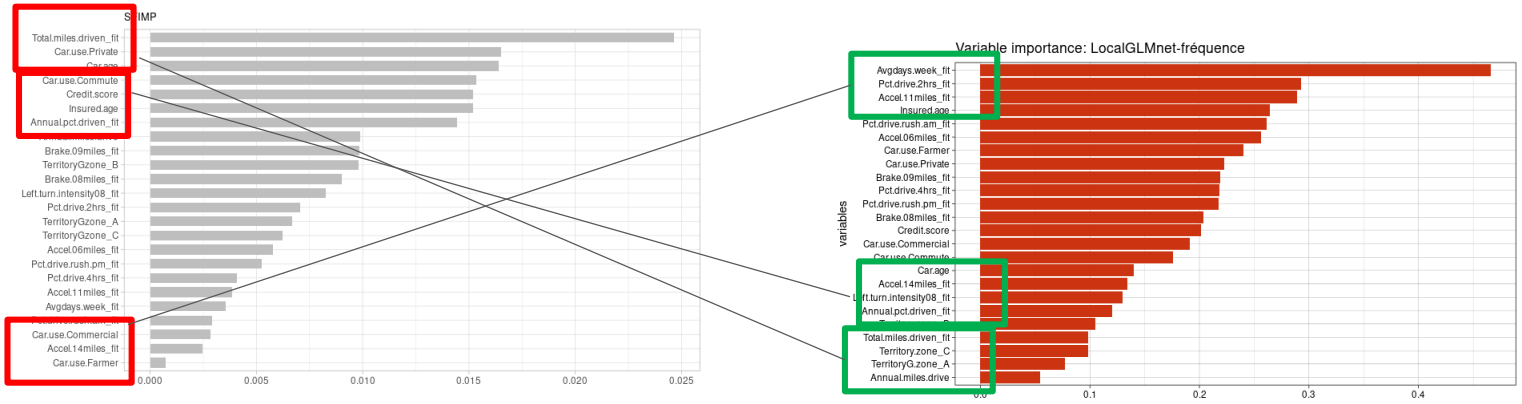
$$x \longmapsto m(\mu(x)) = \beta_0 + < \beta(x), x >= \beta_0 + \sum_{j=1}^{p} \beta_j(x) x_j$$

04. INTERPRÉTATION

- **GLM Poisson**
- **LocalGLMnet Poisson**

Importance des caractéristiques

Un premier élément que nous regardons lorsque nous souhaitons interpréter un modèle est l'importance globale des caractéristiques.



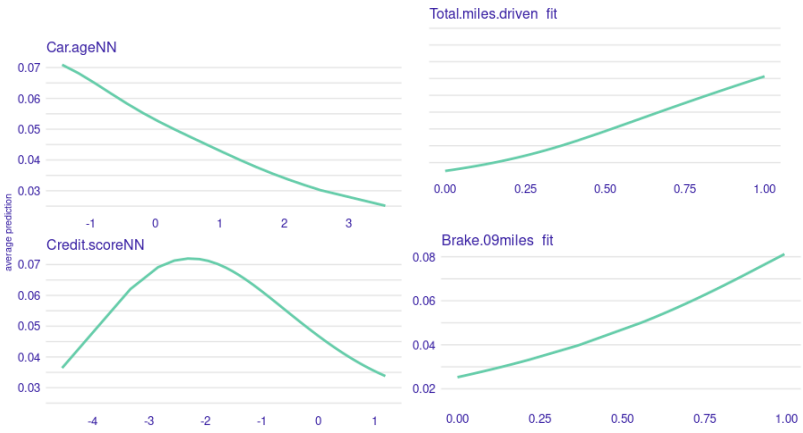
Observation

Le fait que les résultats des deux approches ne coïncident pas nécessairement est du au fait qu'ils ne reposent pas sur le même principe de fonctionnement sous-jacent.

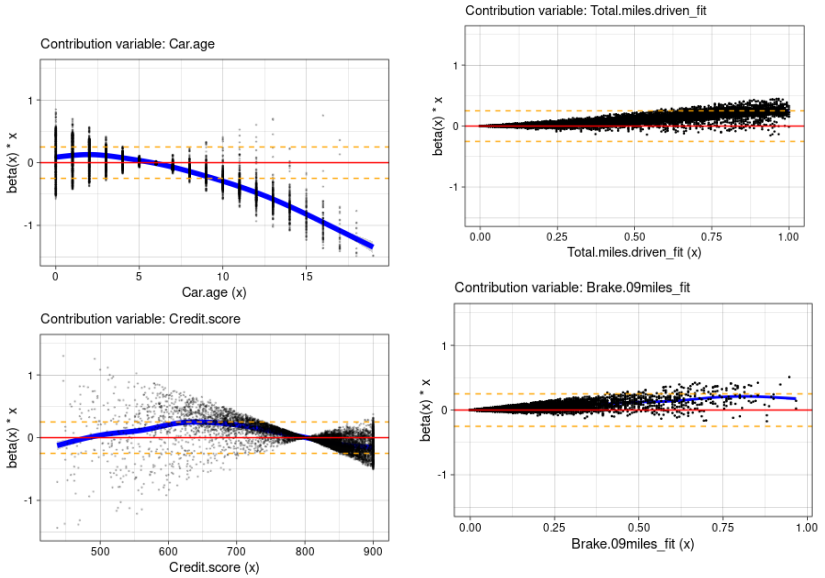
Effets des caractéristiques

Un deuxième élément que nous regardons lorsque nous souhaitons interpréter un modèle est l'effet global des caractéristiques.

ALE (Post hoc)

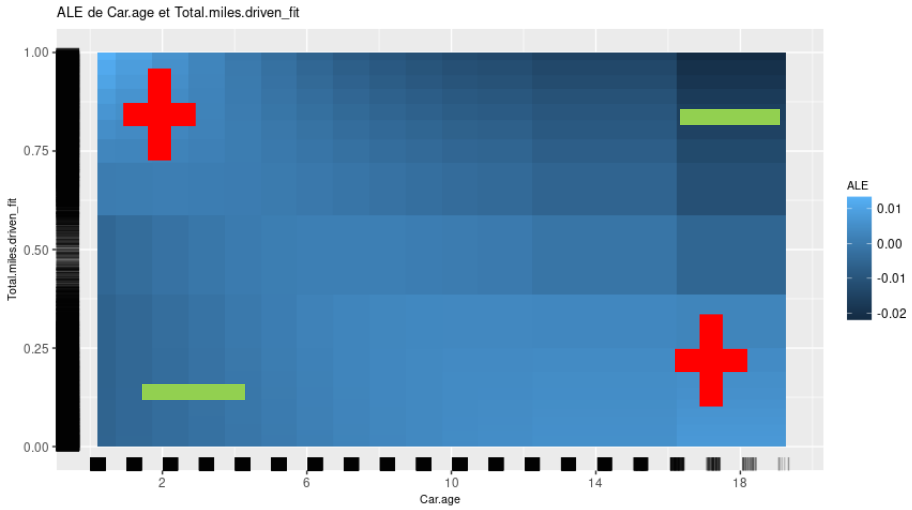


Contribution (IBM)

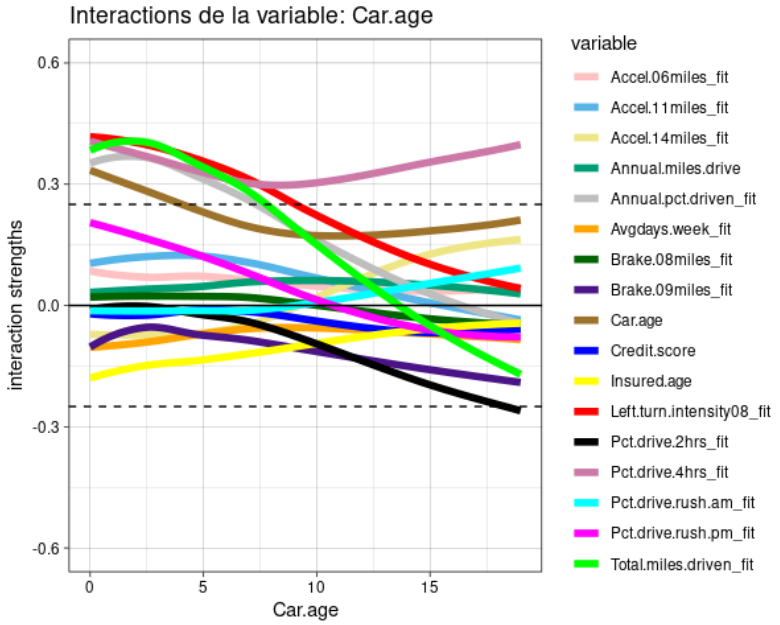


Interprétation globale: Interactions entre les caractéristiques (3/3)

ALE-2D *(Post hoc)*



Interactions (IBM)

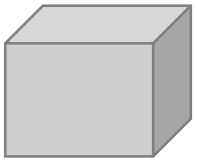


$$\nabla\beta_j(x) = \left(\frac{\partial}{\partial x_1}\beta_j(x), \cdots, \frac{\partial}{\partial x_p}\beta_j(x)\right)^\top \in \mathbb{R}^p$$

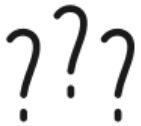
Interprétation locale: illustration (1/5)

Insured.age : **20 ans**	*Car.age* : **0 an** ; *Credit.score* : **580**	*Annual.miles.drive* : **12427.2**
Car.use : **Commute**	*TerritoryG* : **zone_C**	*Total.miles.driven_fit* : **0.92**
Annual.pct.driven_fit : **0.63**	*Pct.drive.rush.am_fit* : **0.11**	*Left.turn.intensity08_fit* : **0.96**
Pct.drive.rush.pm_fit : **0.95**	*Brake.08miles_fit* : **0.95**	*Brake.09miles_fit* : **0.83**
Pct.drive.4hrs_fit : **0.90**	*Pct.drive.2hrs_fit* : **0.98**	*Accel.14miles_fit* : **0.85**
Accel.11miles_fit : **0.87**	*Accel.06miles_fit* : **0.96**	*Avgdays.week_fit* : **1**

LocalGLMnet fréquence



Fréquence prédite= **2.77**
(La plus élevée du jeu de données test)



Client(e)

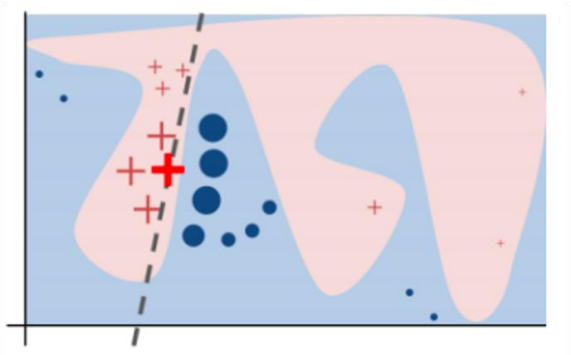
Régulateur

ACPR
BANQUE DE FRANCE

Assureur

Interprétation locale: illustration (2/5)

LIME *(Post hoc)*: illustration



Intuition: Construire un modèle de substitution locale interprétable

LIME *(Post hoc):* définition

explicateur linéaire parcimonieux

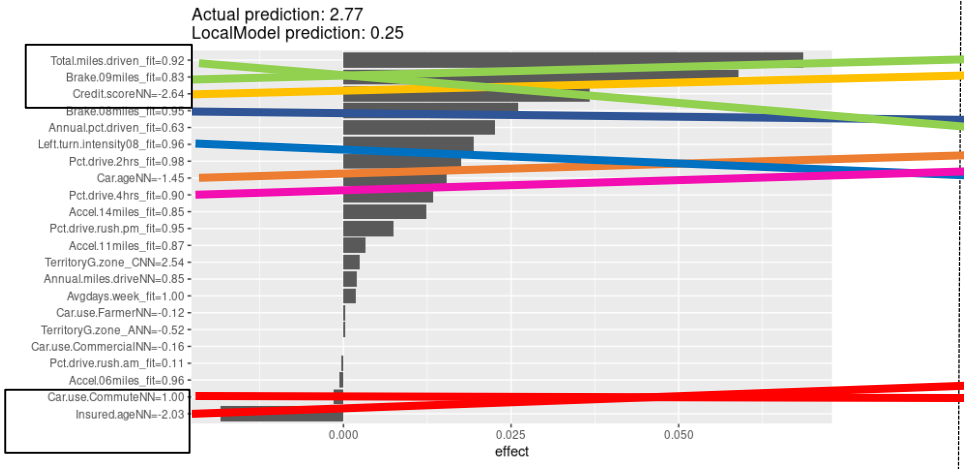
mesure de l'infidélité — **modèle interpretable** — **mesure de complexité**

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

une instance — **modèle à expliquer** — **mesure de proximité pour définir le voisinage local**

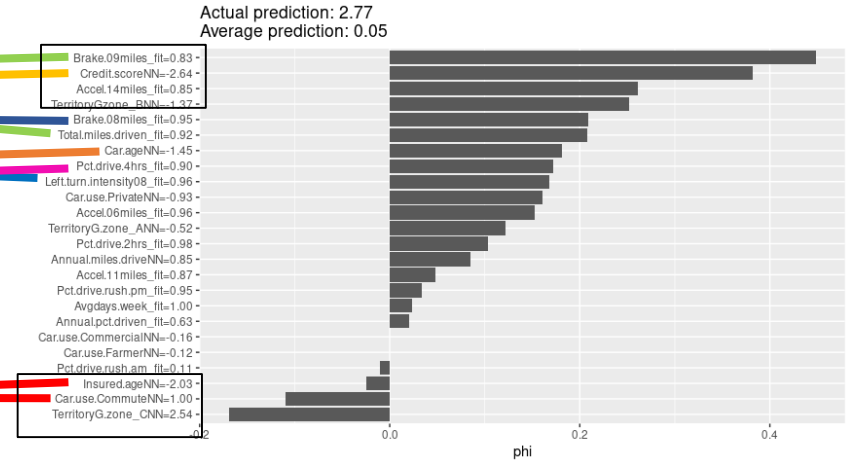
Interprétation locale: illustration (3/5)

LIME *(Post hoc)*



- **Total.miles.driven_fit, Brake.xxmiles** et **Credit.scoreNN** parmi celles qui contribuent le plus **positivement.**

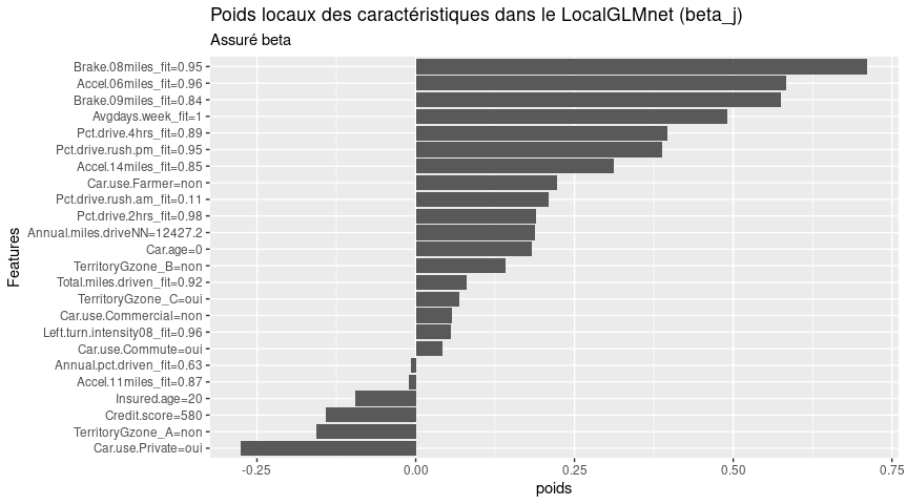
SHAP *(Post hoc)*



- **Insured.ageNN** contribue **négativement.**

Interprétation locale: illustration (4/5)

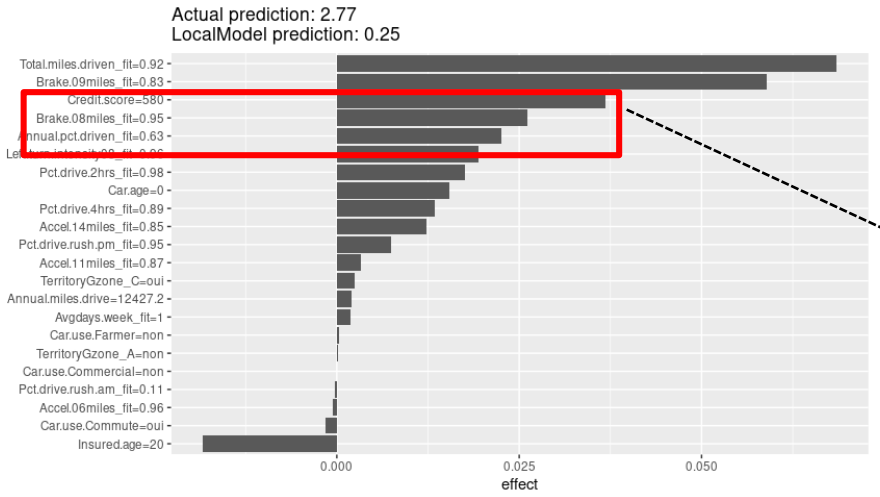
Interprétation Basée sur le Modèle (IBM)



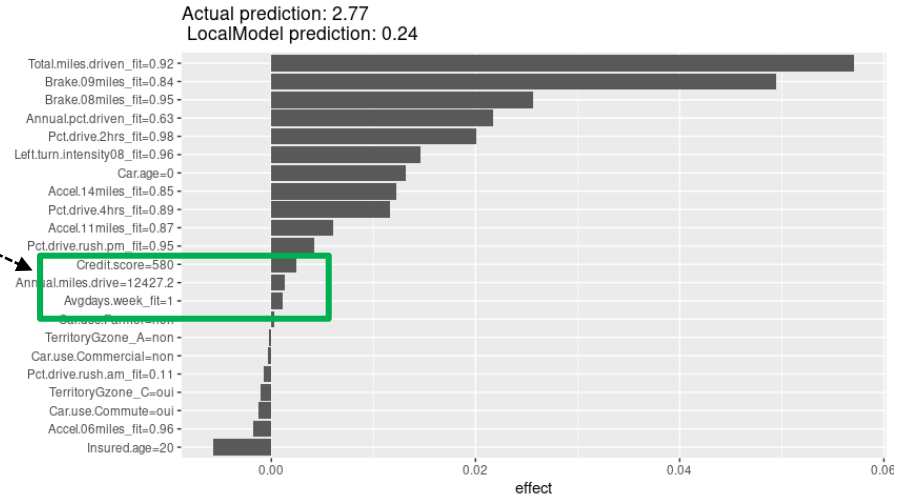
Les interprétations post-hoc sont globalement cohérentes avec l'interprétation basée sur le modèle, à quelques petites exceptions près.

Interprétation locale: limites de la méthode LIME (5/5)

Explications LIME *initiales*



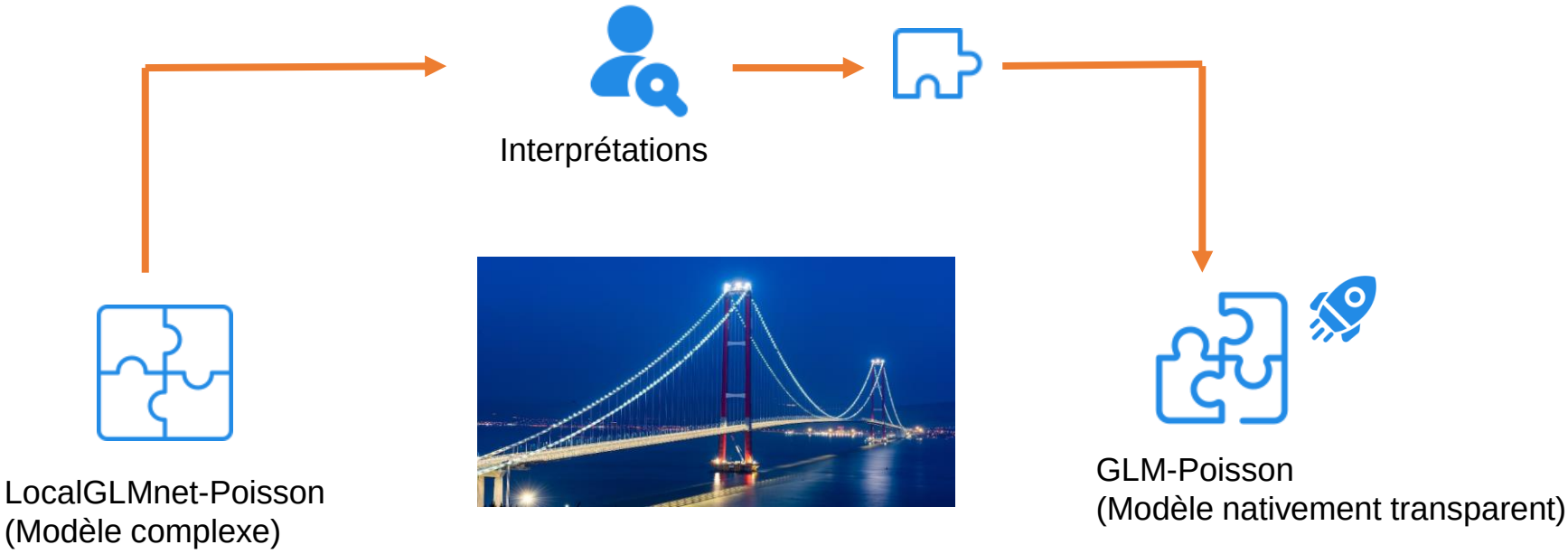
Explications LIME attaquées



$$e(x) = \begin{cases} f(x), \text{ si } is\_OOD(x) \leq 0.8 \\ \psi(x), \text{ si } is\_OOD(x) > 0.8 \end{cases}$$

En dehors du poids de la variable **Credit.score** qui a été modifié, les explications sont quasiment restées inchangées par ailleurs.

Ingénierie des caractéristiques: principe (1/2)



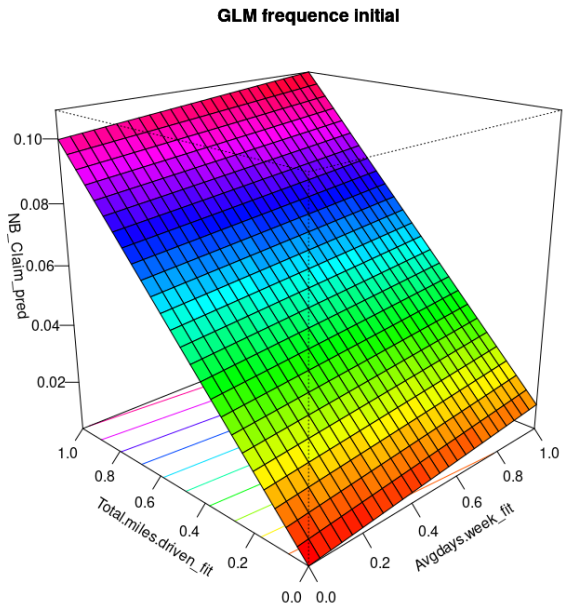
Observation

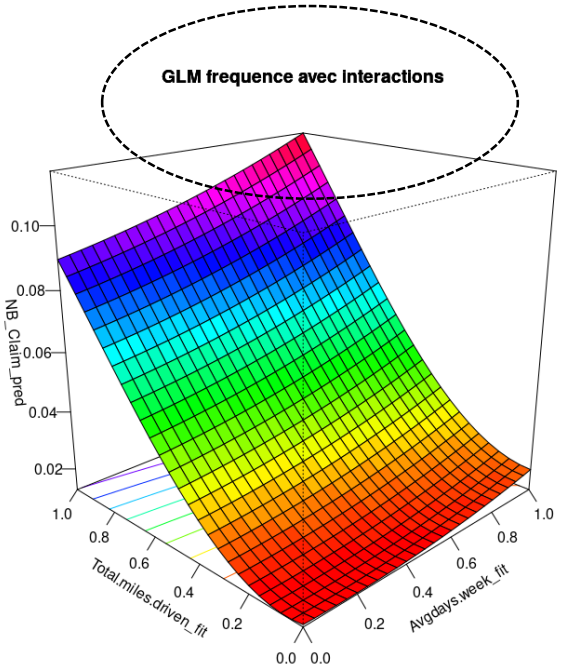
L'interprétabilité peut permettre de réaliser de l'ingénierie des caractéristiques

Ingénierie des caractéristiques: résultats (2/2)

Cas pratique

Sur ces graphiques nous illustrons un cas pratique où nous nous servons des résultats de l'interprétation de notre modèle hybride LocalGLMnet pour enrichir le modèle GLM initial.



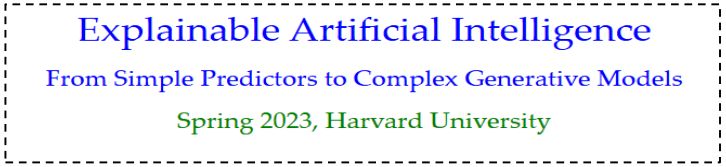


5 Conclusion

- L'interprétabilité **n'est pas un luxe technique**, mais une **exigence éthique** pour que la complexité algorithmique reste au service de l'intelligence humaine notamment dans des **domaines sensibles** comme celui des **assurances**.

- Mettre en place des explications par des **méthodes contrefactuelles.**

- La communauté scientifique se tourne davantage vers des **modèles hybrides**.



THANK YOU

QUESTION ?