100% ACTUAIRES & DATA SCIENCE
Jeudi 17 novembre 2022

Institut des ACTUAIRES

Explicabilité des modèles de machine learning, quelques explications et mise en pratique

Arthur MAILLART Consultant expert chez Detralytics

Fabien VINAS Group Head of Data Analytics & AI chez Allianz Trade

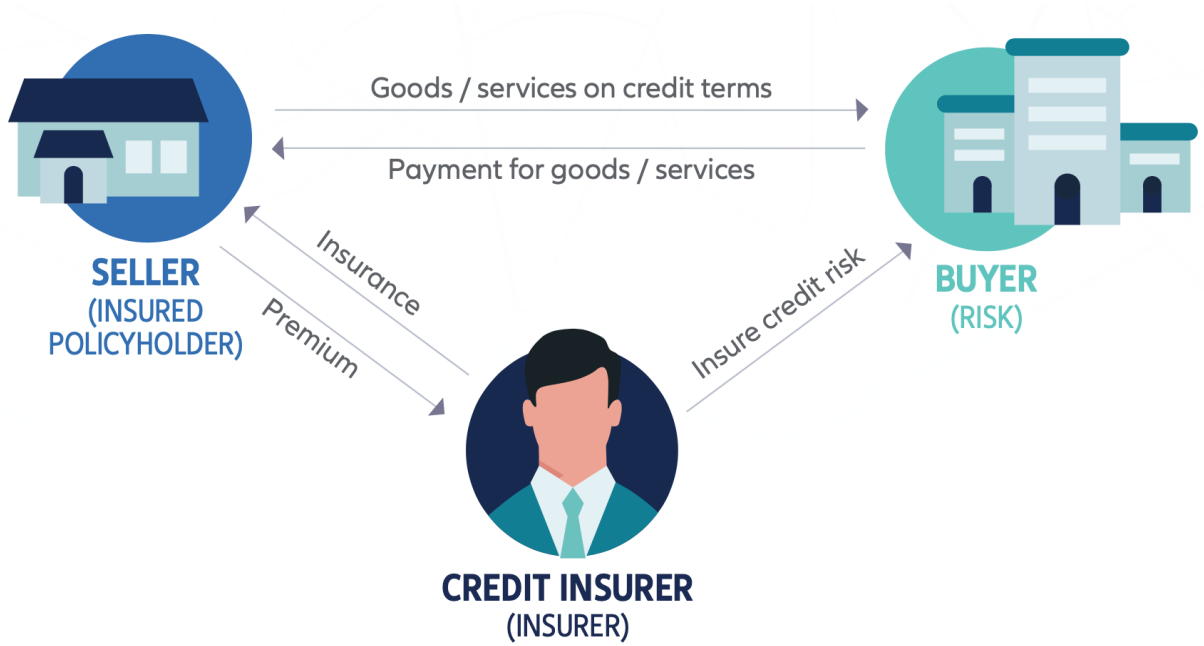
Christian ROBERT Directeur scientifique chez Detralytics

L'Assurance Crédit

L'Assurance Crédit couvre le risque de **défaut de paiement dans le cadre de transactions commerciales entre entreprises**.

Les clients d'Allianz Trade sont ainsi **uniquement des entreprises**.

Le risque assuré (défaut de paiement) porte ainsi essentiellement sur les clients de nos clients (leurs "acheteurs").

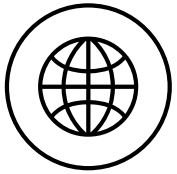


Allianz | Allianz Trade

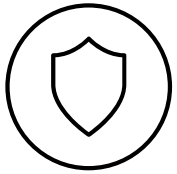
anciennement...

EULER HERMES

Allianz Trade en chiffres

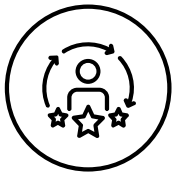


34%
Global market share *.



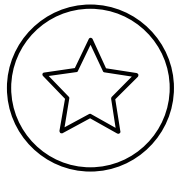
€931 billion
Commercial transactions protected glo

1,000
Claims indemnified per week.



62,000+
Corporate customers.

22,000
Credit limit requests received per day.



AA
Credit Rating (Standard and

83 million
Companies monitored across all sectors in 160+

90%
Credit limit requests processed in 48h.

5,500+
Employees worldwide in 50+ countries with 80 nationalities.

**on addressable global TCI market, excluding ECAs*

Le contexte

- L'étude porte sur l'automatisation d'une tâche pour **la souscription risque chez Allianz Trade**. Cette dernière peut être décrite comme un problème de classification binaire.

- Le premier modèle XGBoost n'a pas été retenu car les explications graphiques fournies avec SHAP aux experts métiers ont été jugées insuffisantes du point de vue de l'interprétabilité.

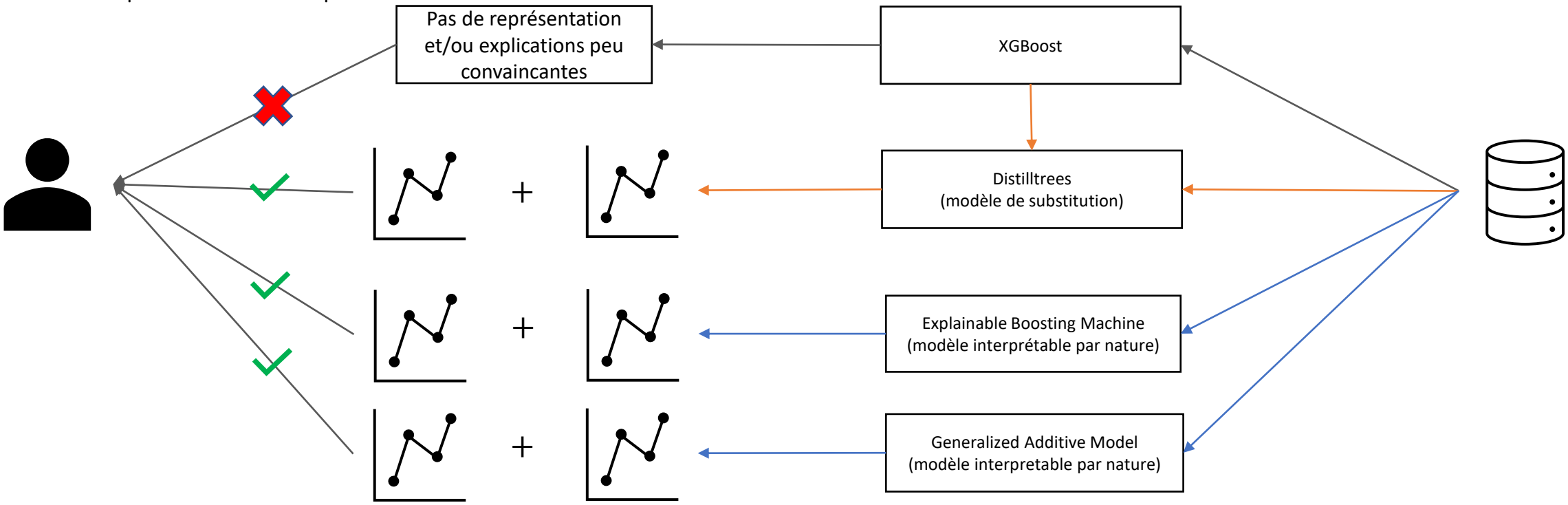
Deux nouvelles approches sont donc étudiées:

- un modèle interprétable par nature
- un modèle de substitution qui permet d'expliquer le modèle complexe

Le cadre de l'étude (1/2)

L'objectif ici n'est pas de discuter de la notion d'interprétabilité en elle même.

Nous partons du **postulat** que le fonctionnement d'un modèle additif est suffisamment simple pour être appris et maitrisé par des non-experts.



Le cadre de l'étude (2/2)

Nous évaluons les performances en AUC de plusieurs modèles additifs:

- Generalized Additive Model,
- Explainable Boosting Machine,
- Distilltrees,

que nous comparons à notre modèle de référence XGBoost.

Ainsi nous évaluons la perte de performance à laquelle il faut consentir pour obtenir un modèle interprétable.

Remarque:

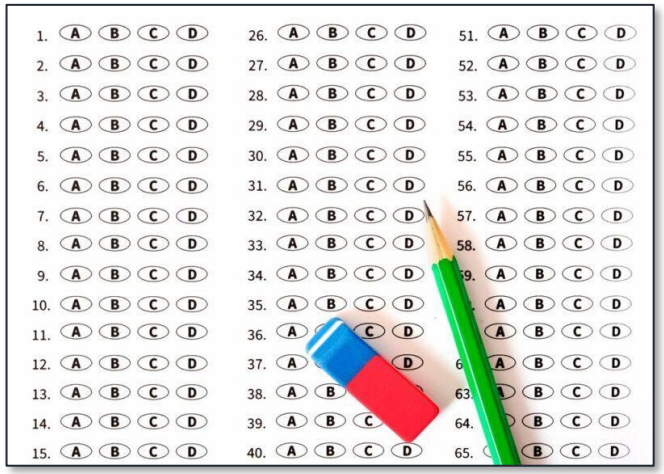
L'approche Distilltrees est initialement prévue pour fournir une explication pour les prédictions d'un ensemble d'arbres. Néanmoins, l'explication fournie étant elle-même un modèle, elle peut fournir des prédictions de bonne qualité. C'est pourquoi, dans ce cas particulier, nous mesurerons à la fois la fidélité des prédictions à celles du modèle de référence XGBoost et à la fois la performance en prédiction du modèle (explication).

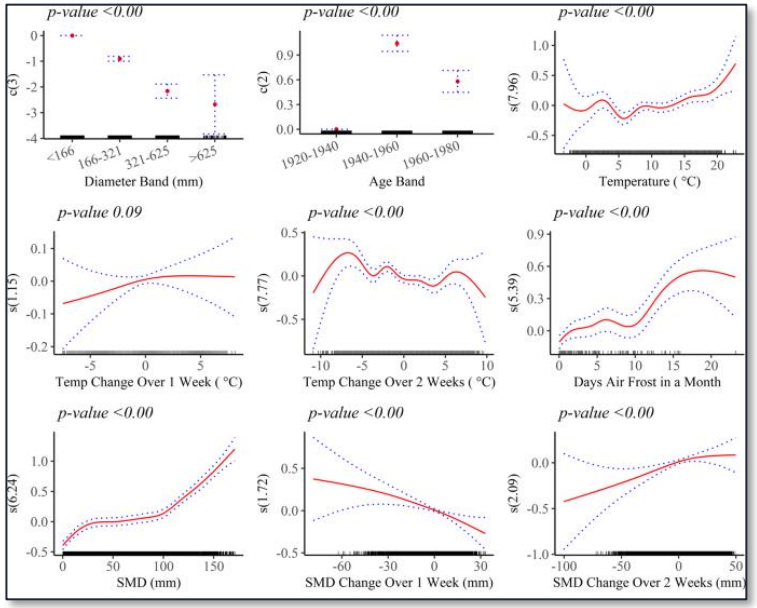
Generalized Additive Models

Un modèle interprétable par nature.

De la score card aux Generalized Additive Models

Un modèle linéaire généralisé où la variable de réponse s'écrit comme une combinaison linéaire de fonctions inconnues des covariables.





$$g(E[y]) = \beta_0 + \sum f_j(x_j)$$

Les fonctions de bases sont décomposées dans des bases de fonctions

$$f_j(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x_j)$$

Une méthode d'estimation basée sur la déviance du GLM et une pénalité pour régularité

$$D(\beta) + \sum_j \lambda_j \int f_j''(x)^2 dx$$

$$\hat{\beta} = \text{argmin}_\beta \{D(\beta) + \sum_j \lambda_j \beta^T S_j \beta\}$$

- Risque d'overfitting
- Pour des bases de données peu importantes (difficilement parallélisable)

Explainable Boosting Machine

Produire un modèle interprétable par nature

Explainable Boosting Machines: utilisation du Gradient Boosting pour calibrer un GAM

README.md

InterpretML - Alpha Release

license MIT | python 3.6 | 3.7 | 3.8 | pypi v0.2.7 | build passing | coverage 89% | code quality: python A | maintained yes

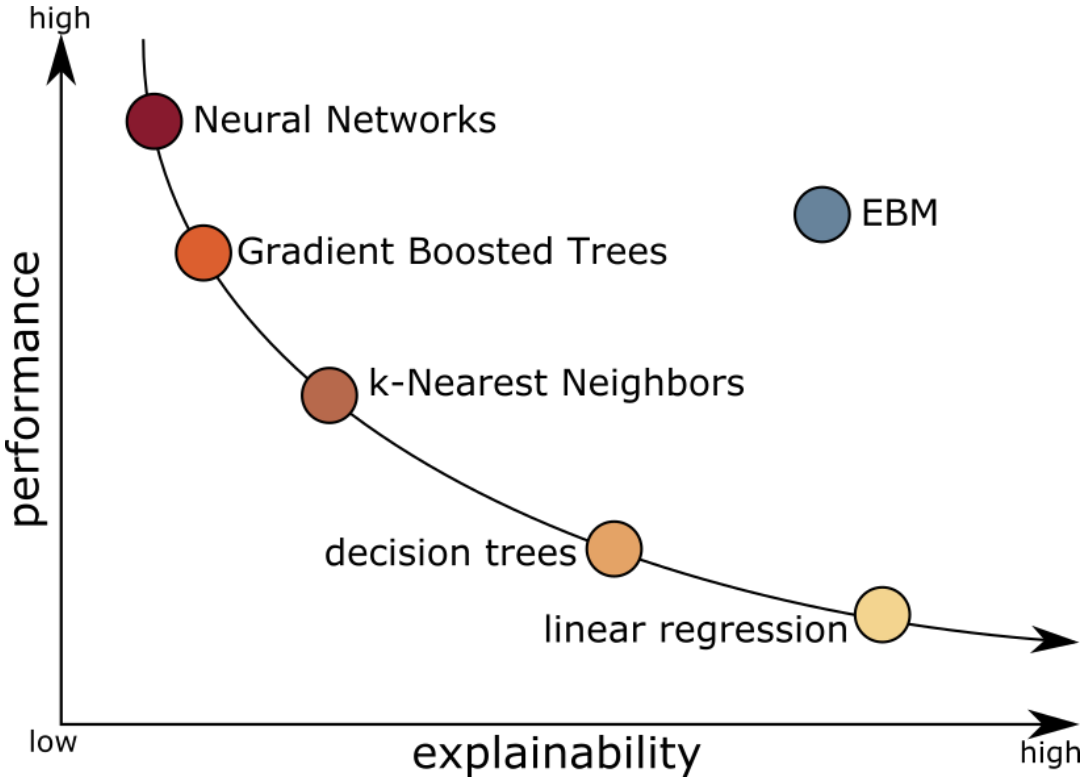
> In the beginning machines learned in darkness, and data scientists struggled in the void to explain them.
>
> Let there be light.

InterpretML is an open-source package that incorporates state-of-the-art machine learning interpretability techniques under one roof. With this package, you can train interpretable glassbox models and explain blackbox systems. InterpretML helps you understand your model's global behavior, or understand the reasons behind individual predictions.

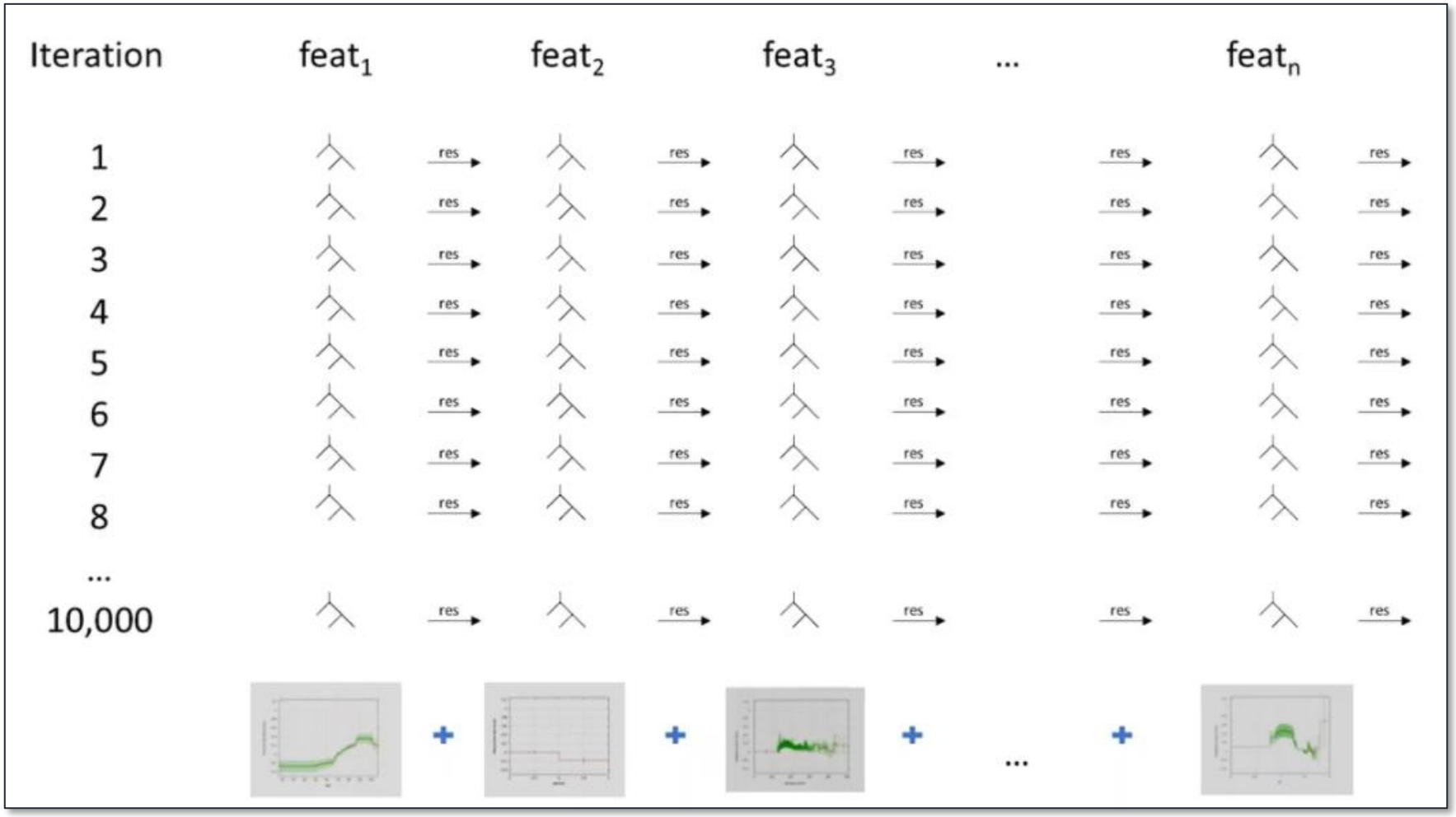
Interpretability is essential for:

- Model debugging - Why did my model make this mistake?
- Feature Engineering - How can I improve my model?
- Detecting fairness issues - Does my model discriminate?
- Human-AI cooperation - How can I understand and trust the model's decisions?
- Regulatory compliance - Does my model satisfy legal requirements?
- High-risk applications - Healthcare, finance, judicial, ...

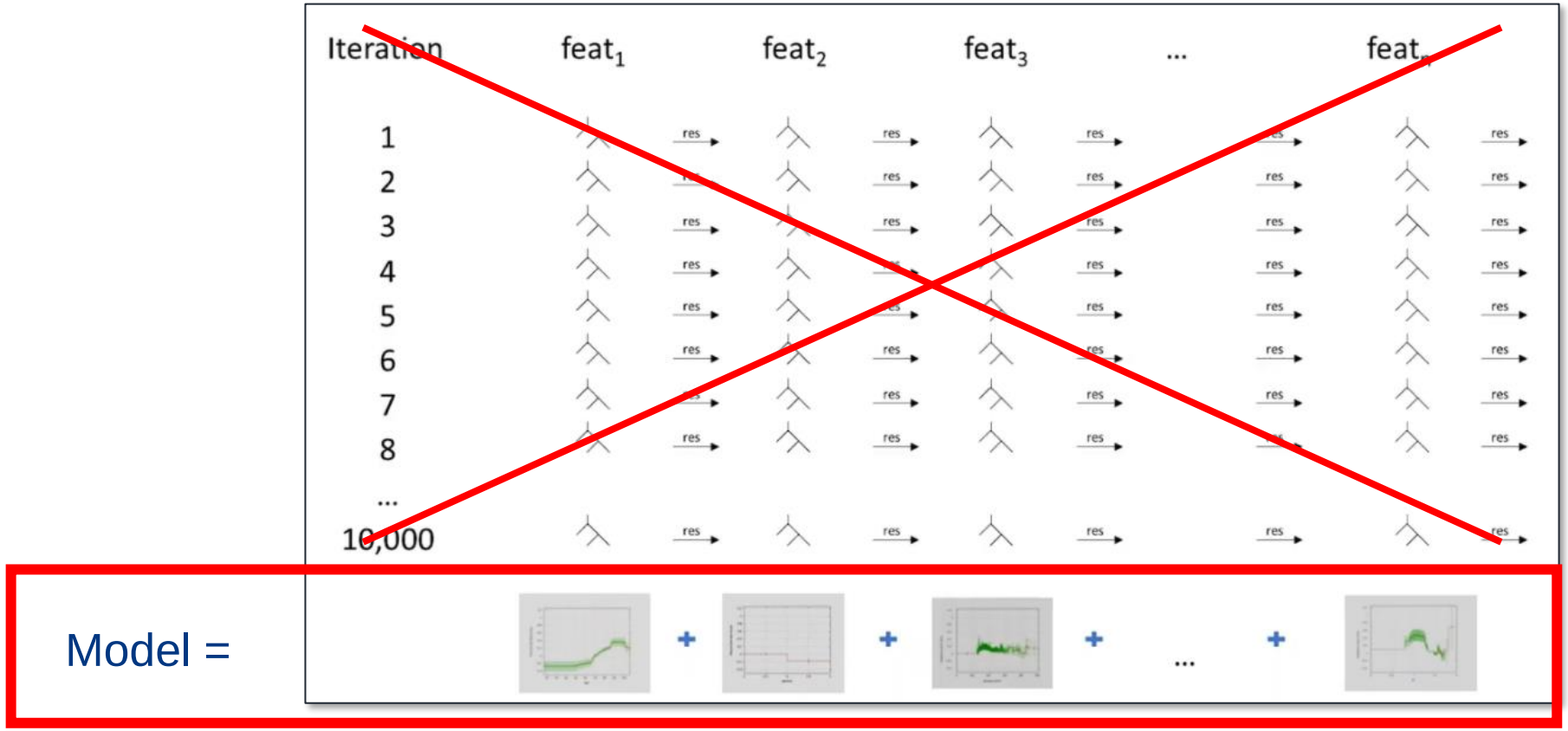
Microsoft Research



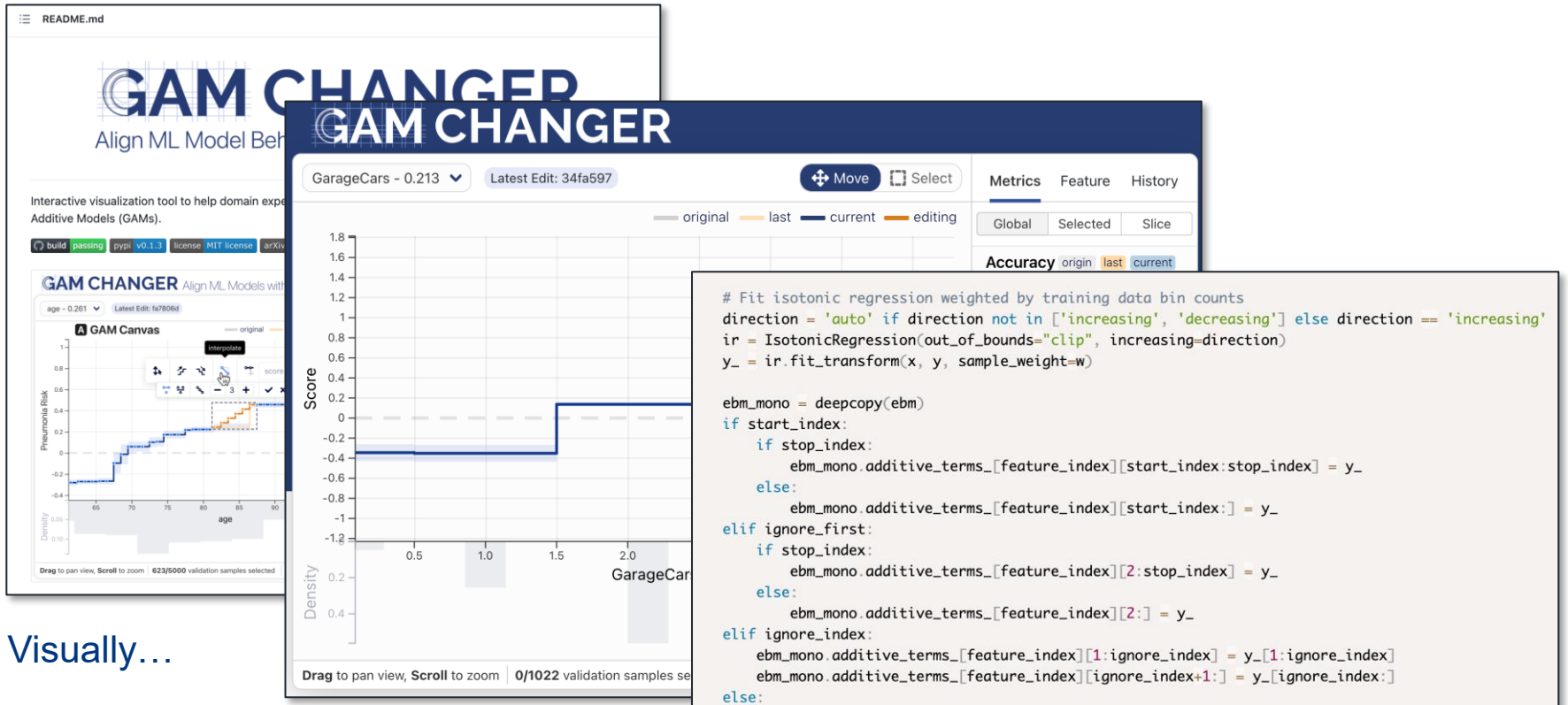
Explainable Boosting Machines: utilisation du Gradient Boosting pour calibrer un GAM



Explainable Boosting Machines: utilisation du Gradient Boosting pour calibrer un GAM



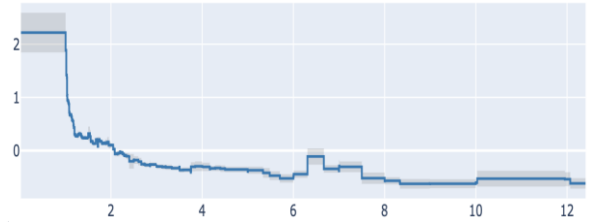
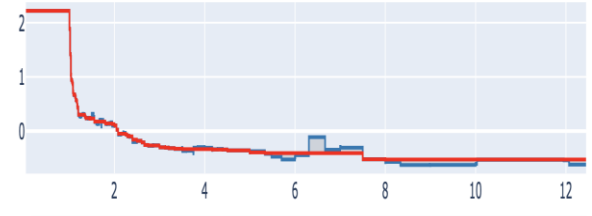
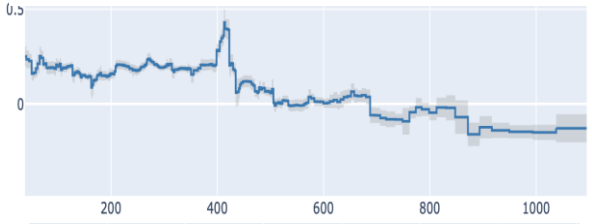
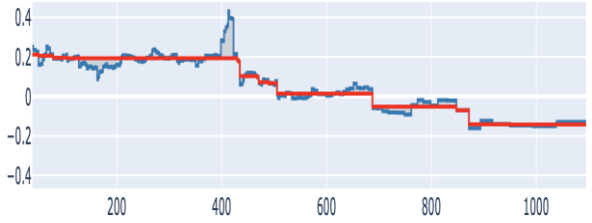
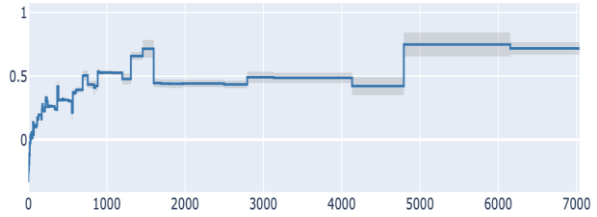
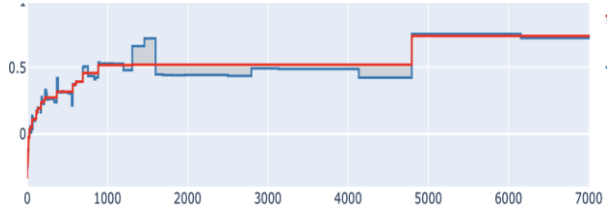
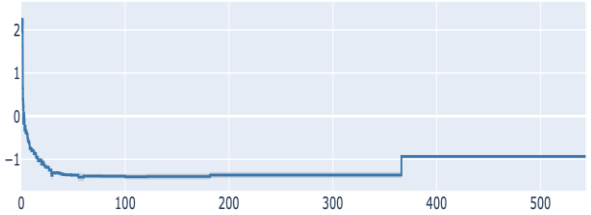
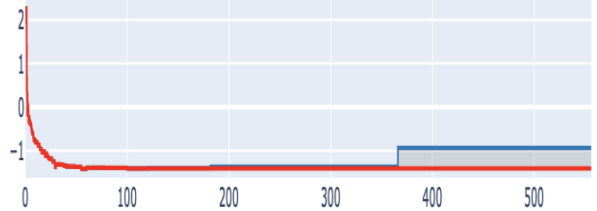
En pratique, les shape plots issus d'EBM méritent d'être lissés, voire édités



Visually…

… or via some code

En pratique, les shape plots issus d'EBM méritent d'être lissés, voire édités

Variable	Shape plot before smoothing	Shape plot after Isotonic regression
Ratio A		
Delay B		
Size factor C		
Ratio D		

Une performance prédictive qui vaut souvent celle du gradient boosting

Public datasets

Classification Performance (AUROC)					
Model	heart-disease (303, 13)	breast-cancer (569, 30)	telecom-churn (7043, 19)	adult-income (32561, 14)	credit-fraud (284807, 30)
EBM	**0.916**	**0.995**	**0.851**	**0.928**	0.975
LightGBM	0.864	0.992	0.835	**0.928**	0.685
Logistic Regression	0.895	**0.995**	0.804	0.907	0.979
Random Forest	0.89	0.992	0.824	0.903	0.95
XGBoost	0.87	**0.995**	0.85	0.922	**0.981**

Figure 3: Classification performance for models across datasets (rows, columns).

AZ Trade projects

Classification Performance (AUC)

Model	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
EBM	0.860	0.844	0.800	0.761	0.963
LightGBM	**0.869**	0.853	0.794	0.761	**0.973**
Logistic Regression	0.799	0.711	0.699	0.741	0.949
Random Forest	0.859	0.832	0.755	0.746	0.947
XGBoost	0.868	**0.854**	**0.804**	**0.772**	**0.973**

Distilltrees

Fournir une explication du modèle XGBoost

Idées générales de Distilltrees

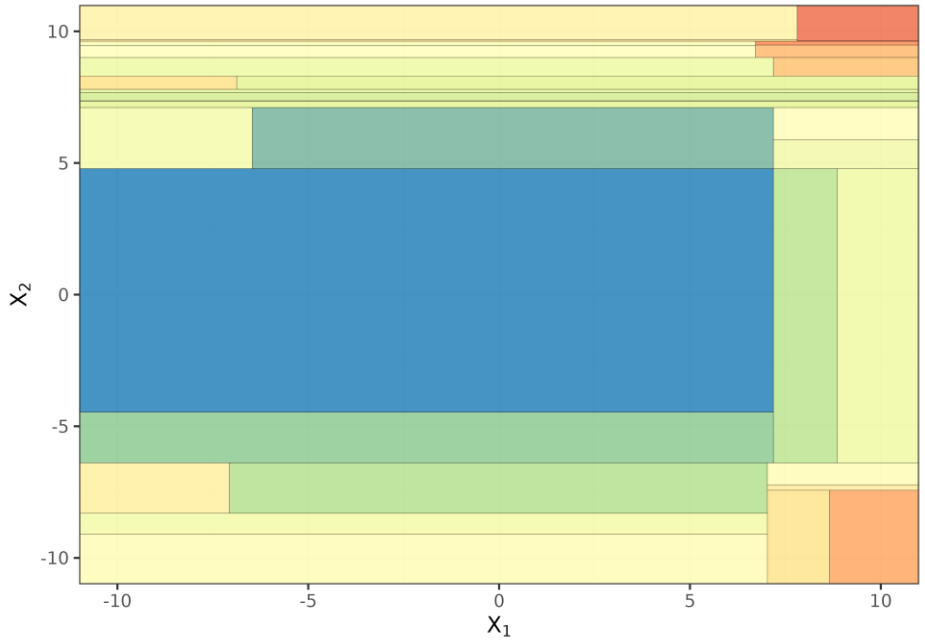
Avec Distilltrees, nous voulons encoder la connaissance de modèles additifs d'ensemble d'arbres (e.g. GBM, XGBoost, Random Forest,...).

Pour cela, nous proposons l'approche suivante:

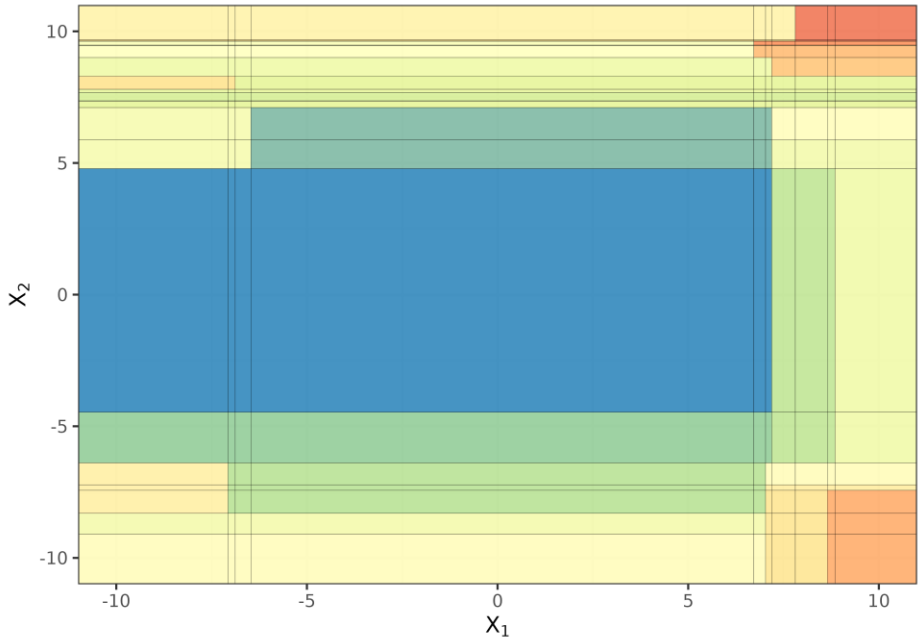
1. Récupérer chaque arbre de l'ensemble
2. Pour chacun de ces arbres identifier l'ensemble des seuils de coupe et créer une nouvelle partition définie par le croisement de ces seuils de coupe
3. Pour chacun des arbres, créer un nouveau jeu de données en faisant prédire l'arbre sur chaque élément de la nouvelle partition

Approximer un arbre par un modèle linéaire

Les variables X1 et X2 sont discrétisées sur base des seuils de coupe fournis par l'arbre



Partition et prédictions d'un arbre



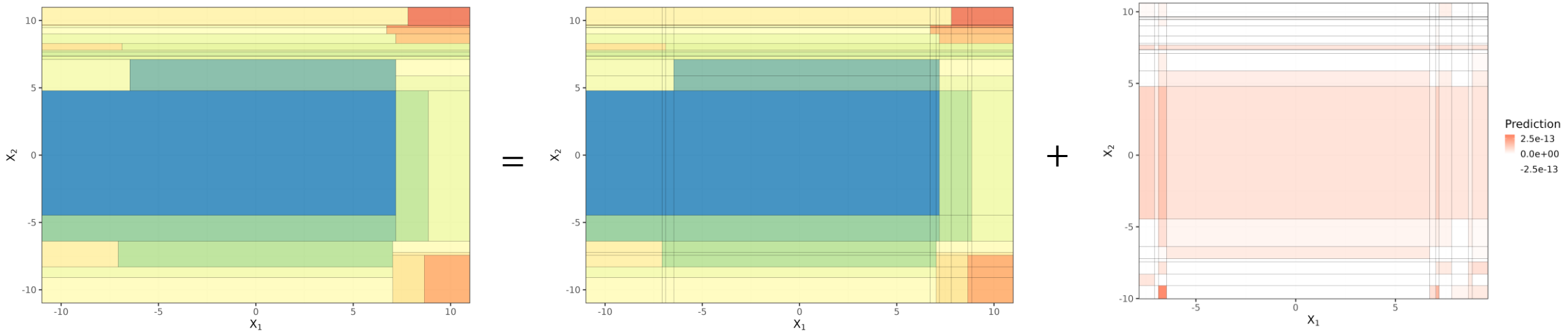
Partition induite par le modèle linéaire et prédictions répliquant celles de l'arbre

Idées générales de Distilltrees

Avec Distilltrees, nous voulons encoder la connaissance de modèles additifs d'ensemble d'arbres (e.g. GBM, XGBoost, Random Forest,...). Pour cela, nous proposons l'approche suivante:

1. Récupérer chaque arbre de l'ensemble
2. Pour chacun de ces arbres identifier l'ensemble des seuils de coupe et créer une nouvelle partition définie par le croisement de ces seuils de coupe
3. Pour chacun des arbres, créer un nouveau jeu de données en faisant prédire l'arbre sur chaque élément de la nouvelle partition
4. Entrainer un modèle linéaire sur ce nouveau jeu de données constitué de l'ensemble des couples (point moyen, réponse de l'arbre)

Approximer un arbre par un modèle linéaire

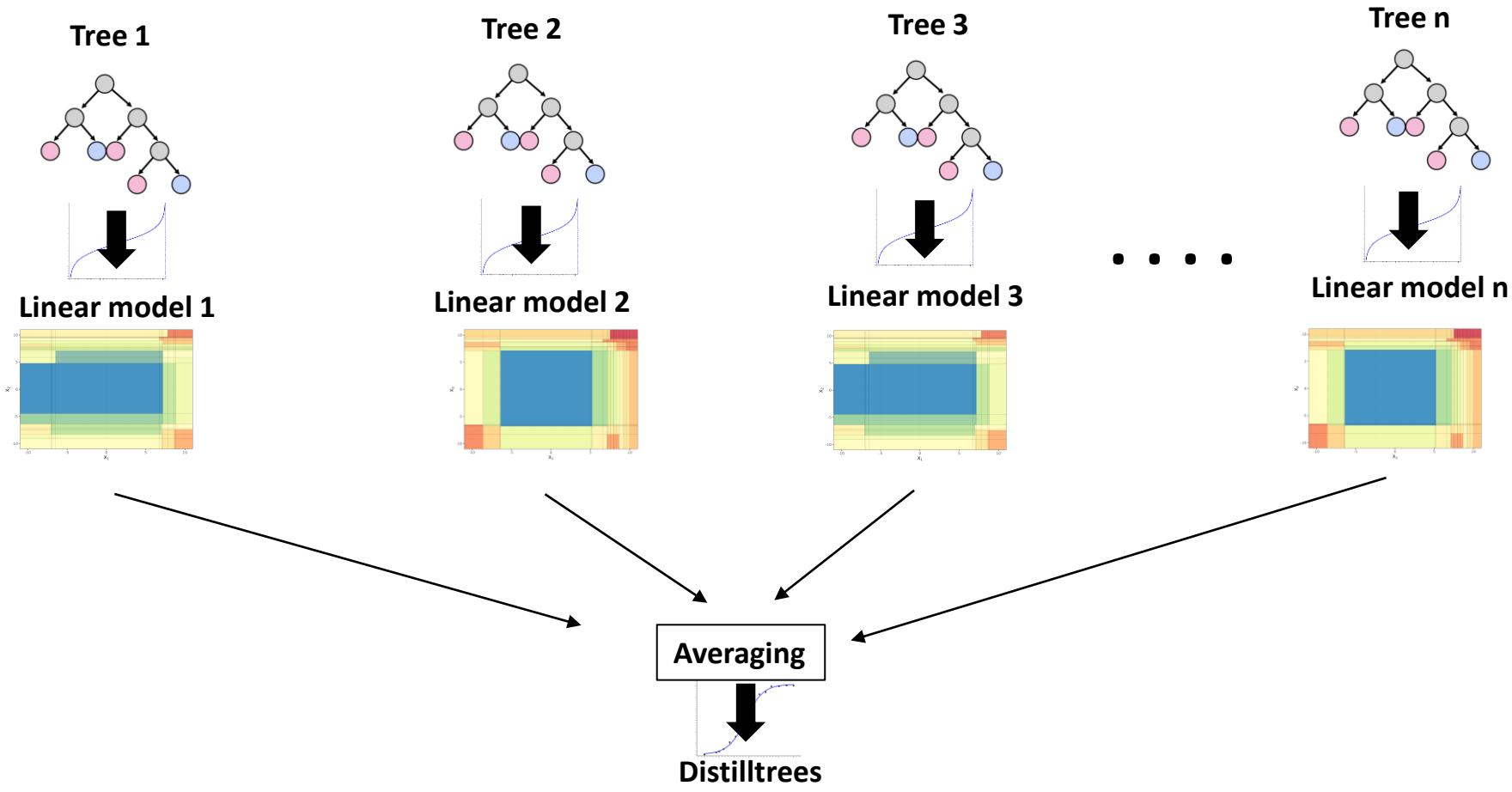


Réponse d'un arbre dans l'espace des covariables

Réponse du modèle linéaire dans l'espace des covariables

Erreur commise dans chaque rectangle

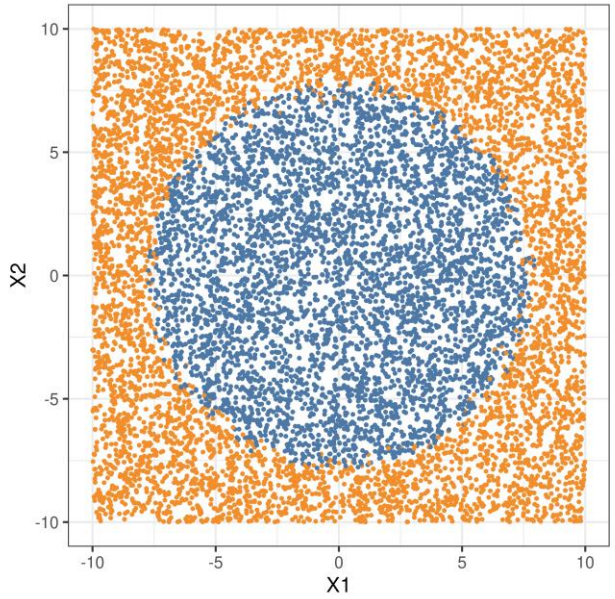
En combinant les prédictions des modèles linéaires entre elles, on peut approximer les prédictions de l'ensemble d'arbre. L'avantage par rapport à l'ensemble d'arbres est qu'on peut en extraire des graphiques explicatifs.



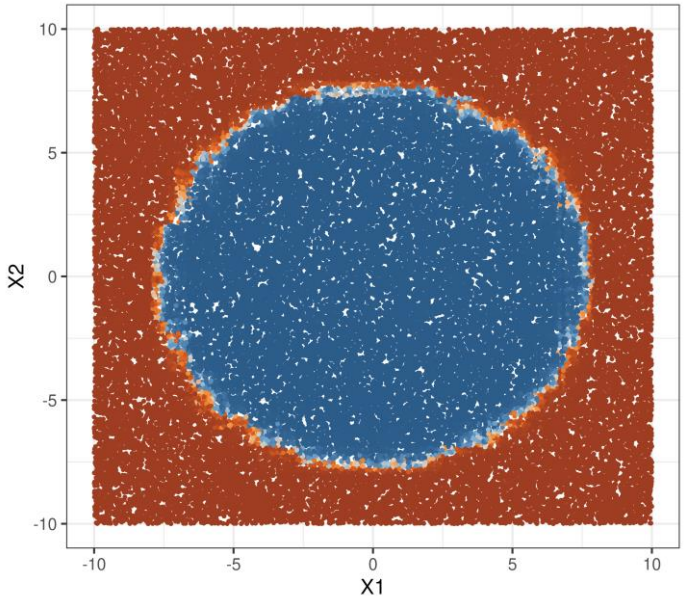
Exemple en 2D

Classification binaire

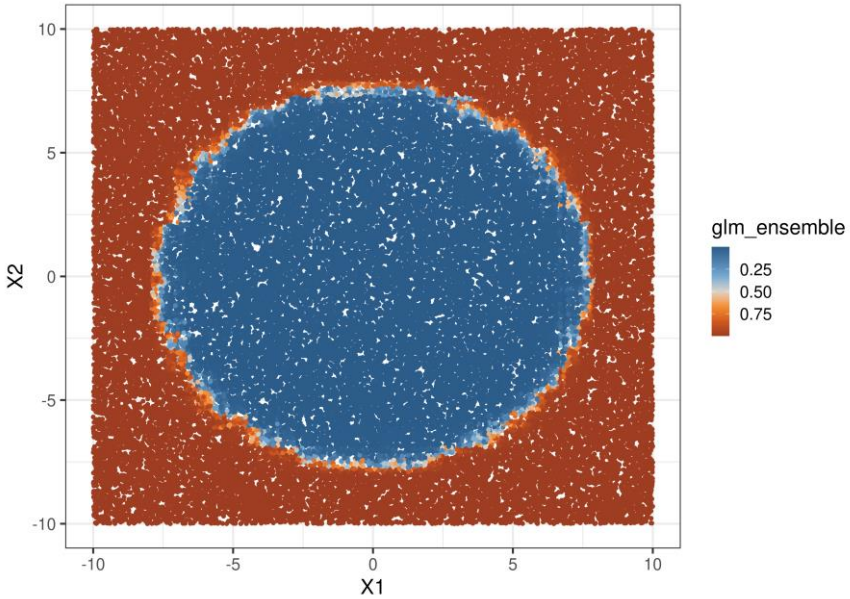
Nous illustrons l'approche en 2 dimension avec un exemple jouet simulé. Les données simulées construisent un cercle. A l'intérieur, les individus valent tous 0 et à l'extérieur, tous les individus valent 1.



Données simulées



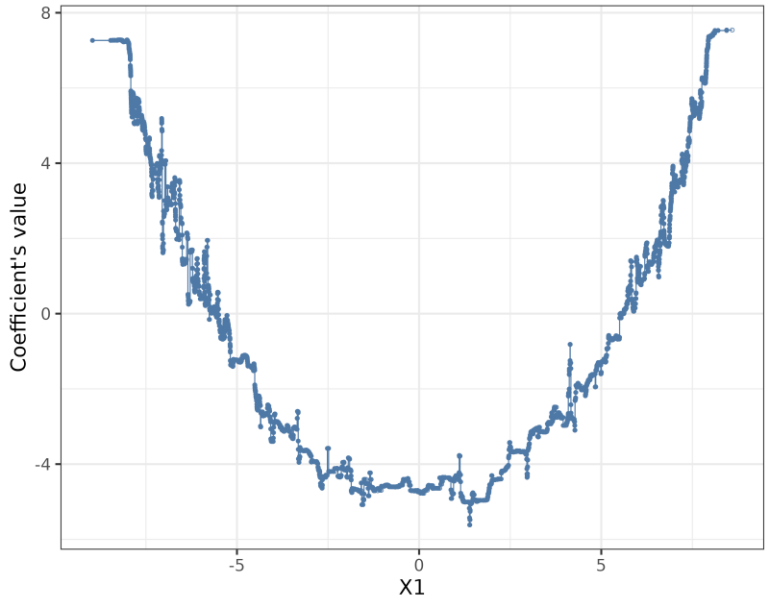
Réponse du modèle gbm



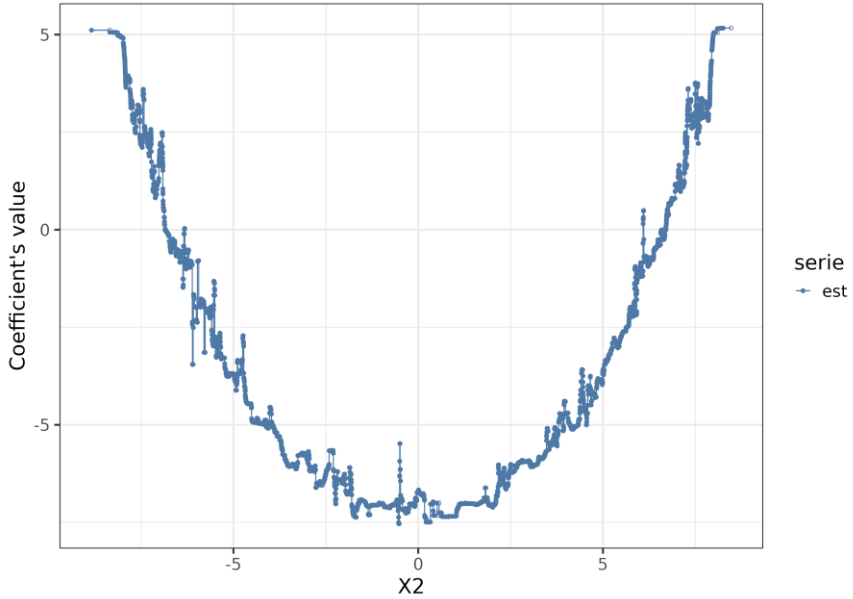
Réponse de l'ensemble de modèles linéaires

Explication fournie par le modèle

Score = 2.50 +



+



Pour passer du score à la prédiction, il suffit d'appliquer la fonction sigmoïde.

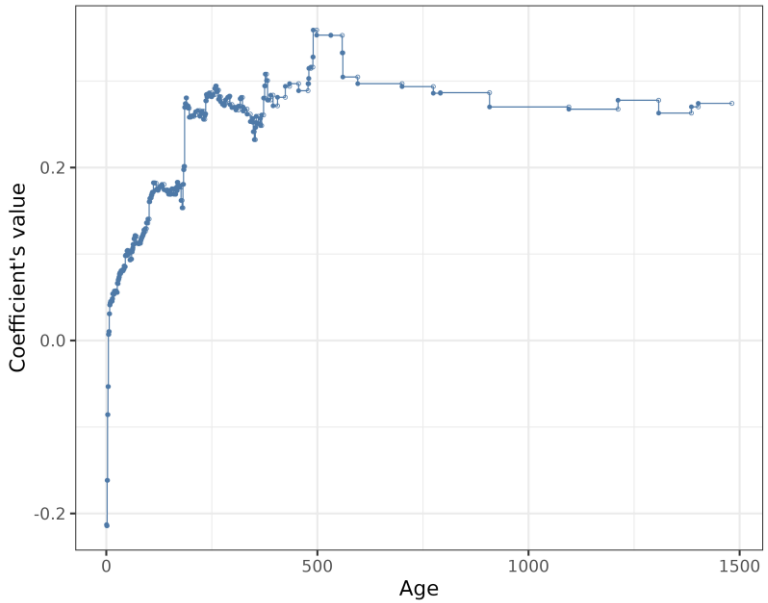
$$prediction = \frac{1}{1 + e^{-score}}$$

Cas d’usage

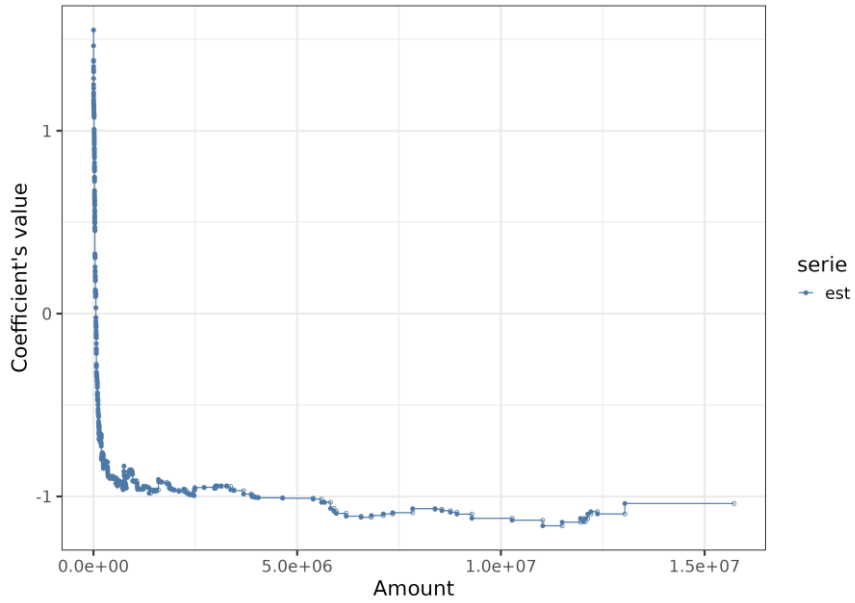
Classification binaire: automatisation d'une tâche pour la souscription risque chez Allianz Trade

Pour notre cas d'usage, distilltrees fournit entre autre les graphiques suivants pour expliquer les prédictions d'XGBoost:

Score = 1.17 +



+ … +



Explication du modèle linéaire

Fidélité des prédictions

La fidélité est la capacité d'un modèle à répliquer les prédictions d'un autre modèle. La prédiction du modèle expliqué est considérée comme la vérité et le modèle de substitution prédit cette nouvelle cible. Ici, nous proposons de mesurer la fidélité avec le R^2 score et l'AUC.

	Surrogate linear (17 variables)	Surrogate linear + all interactions (17 + 136 variables)
Train	93,81 %	99,73%
Test	94,44 %	99,77%

Fidélité mesurée avec le R^2 score

	Surrogate linear (17 variables)	Surrogate linear + all interactions (17 + 136 variables)
Train	98,42 %	99,93%
Test	98,74 %	99,95%

Fidélité mesurée avec l'AUC score

Comparaison des performances

Les performances, mesurées avec l'AUC, de tous les modèles testés dans cette étude sont résumées dans le tableau ci-dessous. Pour des raisons de lisibilité nous ne présentons que les modèles sans interactions.

	XGBoost	GAM (17 variables sans interactions)	EBM (17 variables sans interactions)	Surrogate linear (17 variables sans interactions)
Train	90,28 %	82,67 %	88,12%	87,95%
Test	89,37 %	81,83 %	87,69%	87,42%

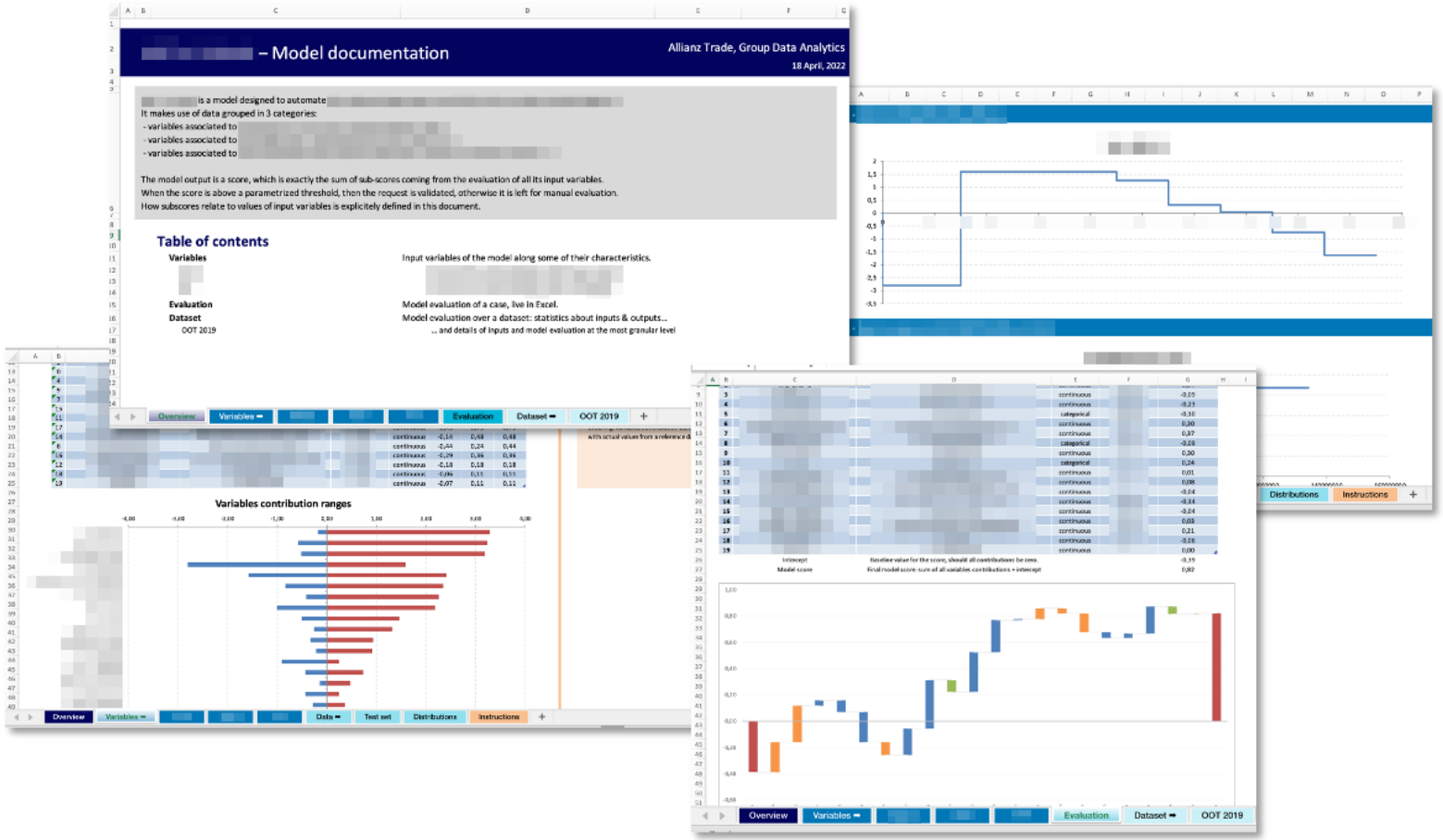
Performance mesurée avec l'AUC score

Intérêts de la méthode

- Cette méthode permet d'expliquer un modèle additif de prédictions d'arbres lorsque la fidélité est suffisamment élevée.
- Si la performance est proche de celle du modèle initial (XGBoost ici), il est possible de créer un modèle de prédiction explicable qui remplace notre modèle initial.
- L'approche peut se décliner pour la classification et la régression.
- L'approche peut s'appliquer sur une vaste classe de modèle: random forest, gradient boosting et variantes. De plus, même si nous ne l'avons pas testée spécifiquement dessus, cette technique peut s'appliquer sur une combinaison linéaire de modèle additifs de prédictions d'arbres (stacking).

Retour d’expérience

ebm2xls: Export exact d'un GAM dans un fichier Excel: permet aux non-techniciens de s'approprier le modèle dans un cadre technique maîtrisé



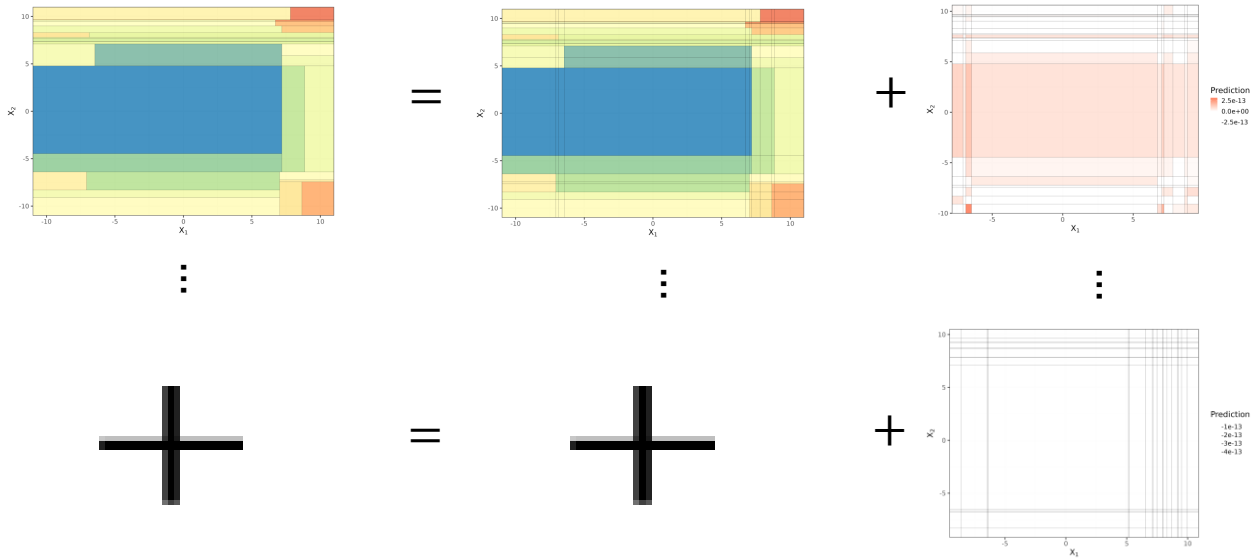
- Modèle exactement exporté dans Excel
- Prédiction exacte calculable dans Excel, sur un cas comme en masse sur un portefeuille
- Tests de sensibilité

- Documente automatiquement le modèle
- Rend les modèles tangibles, les démystifie auprès des collègues non-experts
- Facilite la revue experte et les échanges
- Facilite l'intégration technique (IT)

Annexes

Approximer un ensemble d'arbre par un modèle linéaire

En combinant les prédictions des modèles linéaires entre elles, on peut approximer les prédictions de l'ensemble d'arbre. L'avantage par rapport à l'ensemble d'arbres est qu'on peut en extraire des graphiques explicatifs.



Fidélité des prédictions

La fidélité est la capacité d'un modèle à répliquer les prédictions d'un autre modèle. La prédiction du modèle expliqué est considérée comme la vérité et le modèle de substitution prédit cette nouvelle cible. Ici, nous proposons de mesurer la fidélité avec le R^2 score et l'AUC.

	Surrogate linear (17 variables)	Surrogate linear + some interactions (17 + 17 variables)	Surrogate linear + all interactions (17 + 136 variables)
Train	93,81 %	95,99 %	99,73%
Test	94,44 %	96,35 %	99,77%

Fidélité mesurée avec le R^2 score

	Surrogate linear (17 variables)	Surrogate linear + some interactions (17 + 17 variables)	Surrogate linear + all interactions (17 + 136 variables)
Train	98,42 %	98,92 %	99,93%
Test	98,74 %	99,04 %	99,95%

Fidélité mesurée avec l'AUC score

Remarques:

Dans le cas où l'on souhaite mesurer la fidélité avec l'accuracy:

- Pour convertir les probabilités prédites en prédictions binaires, nous choisissons le seuil qui maximise l'accuracy. Avec ce seuil, nous obtenons une fidélité de 94,18%.
- Sous ce seuil on obtient la matrice de confusion suivante:

	Predicted: 0	Predicted: 1
Actual: 0	38 461	2249
Actual: 1	1931	29 203

Comparaison des performances

Les performances, mesurées avec l'AUC, de tous les modèles testés dans cette étude sont résumées dans le tableau ci-dessous. Pour des raisons de lisibilité nous ne présentons que les modèles sans interactions.

	XGBoost	GAM (17 variables sans interactions)	EBM (17 variables sans interactions)	Surrogate linear (17 variables sans interactions)
Train	90,28 %	82,67 %	88,12%	87,95%
Test	89,37 %	81,83 %	87,69%	87,42%

Performance mesurée avec l'AUC score

	Predicted: 0	Predicted: 1
Actual: 0	37 114	4606
Actual: 1	8488	21 636

Matrice de confusion XGB

- **Accuracy: 81,77%**
- **Recall: 71,82%**
- **Precision: 82,45%**

	Predicted: 0	Predicted: 1
Actual: 0	33722	7998
Actual: 1	6670	23454

Matrice de confusion Distilltrees

- **Accuracy: 79,58%**
- **Recall: 77,86%**
- **Precision: 74,57%**