



Montre-moi donc ton modèle !

Les graphiques d'interprétabilité du machine-learning
expliqués

Christophe Regouby
18 mars 2021

AIRBUS



Le contexte : la réglementation !

- France

Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décision ...

« Art. R. 311-3-1-2.-**L'administration** communique à la personne faisant l'objet d'une **décision individuelle** prise sur le fondement d'un traitement algorithme, à la demande de celle-ci, **sous une forme intelligible** et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

« 1° Le degré et le mode de contribution du traitement algorithme à la prise de décision ;

« 2° **Les données** traitées et leurs sources ;

« 3° **Les paramètres** de traitement et, le cas échéant, **leur pondération**, appliqués à la situation de **l'intéressé** ;

« 4° **Les opérations** effectuées par le traitement ; ».

- Europe : RGPD / GDPR

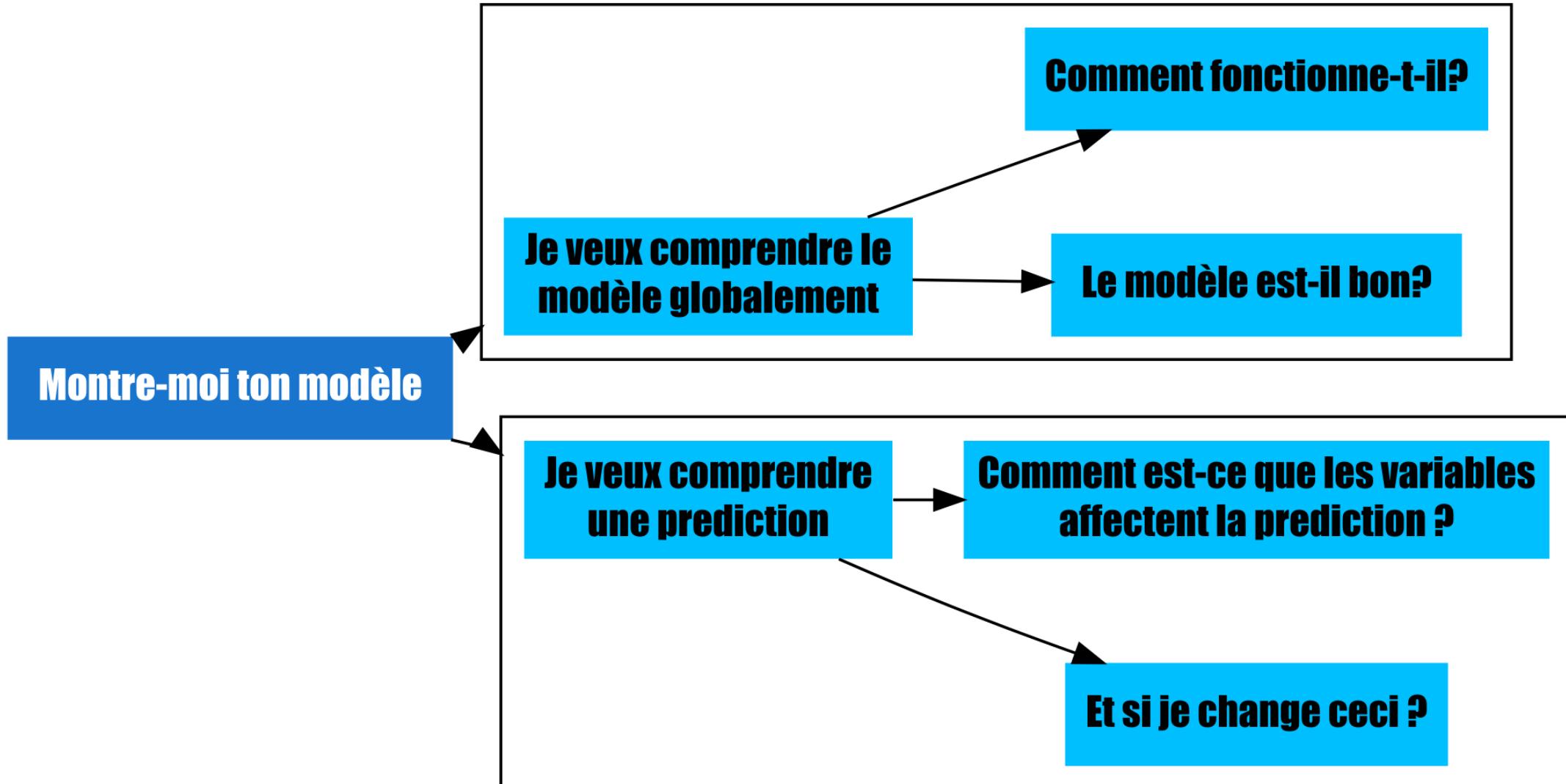
« Article 15.1.h) : La personne concernée a le droit d'obtenir du responsable du traitement la confirmation que des données à caractère personnel la concernant sont ou ne sont pas traitées et, lorsqu'elles le sont, l'accès auxdites données à caractère personnel ainsi que les informations suivantes:

l'existence d'une prise de **décision automatisée**, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, **des informations utiles concernant la logique sous-jacente**, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée.

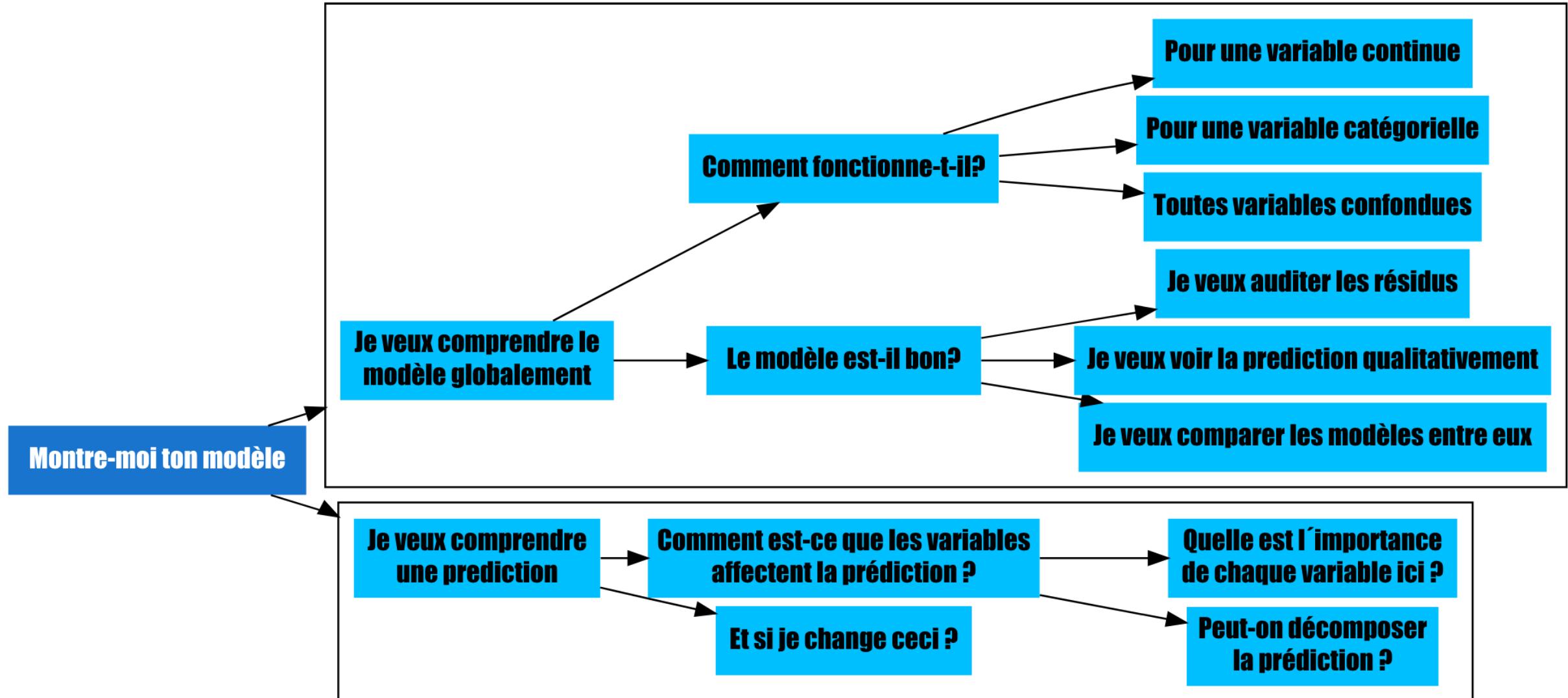
- toutes les réglementations réunies

<https://github.com/ModelOriented/MAIR>

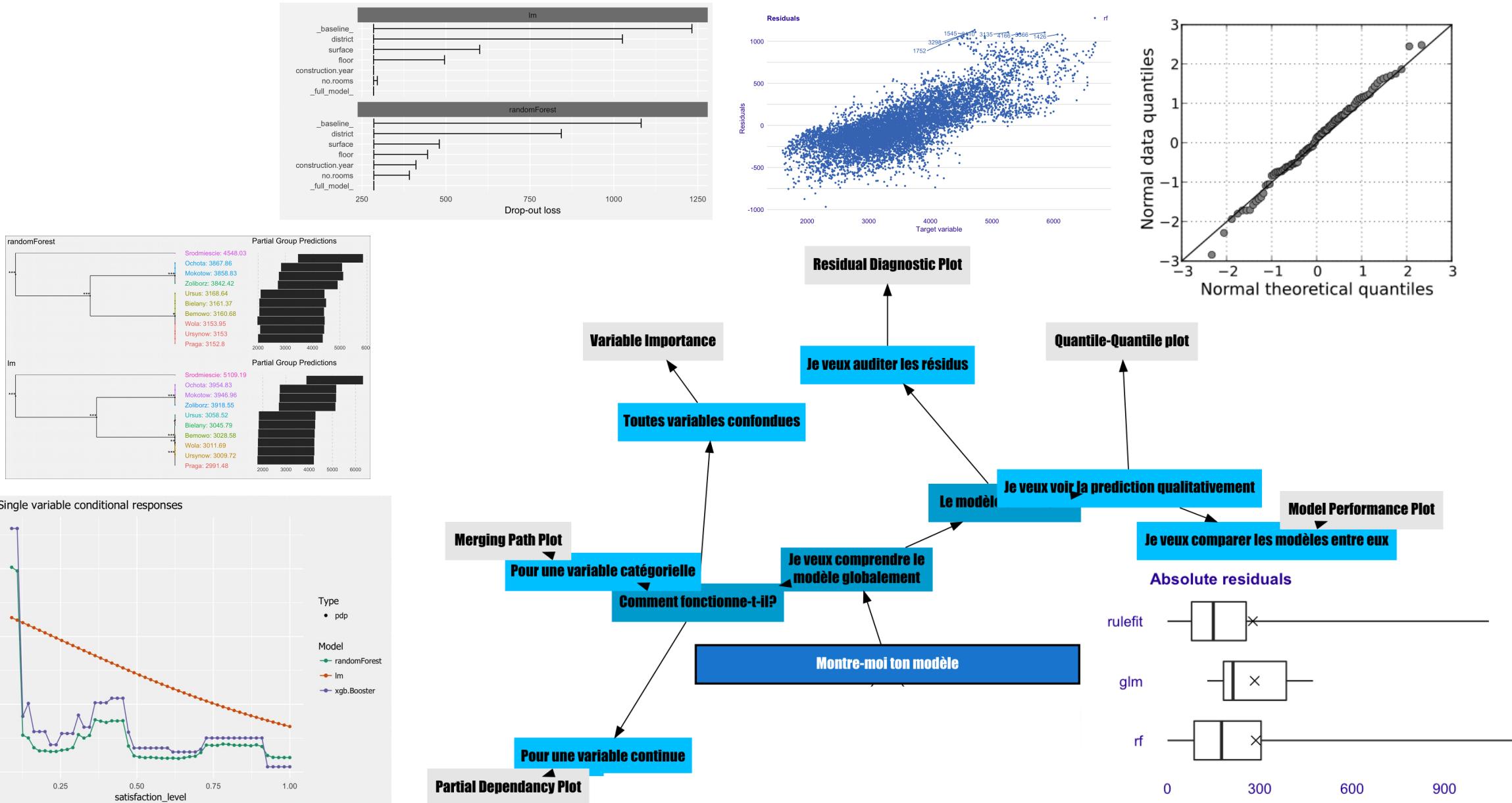
Montre-moi ton modèle : la problématique



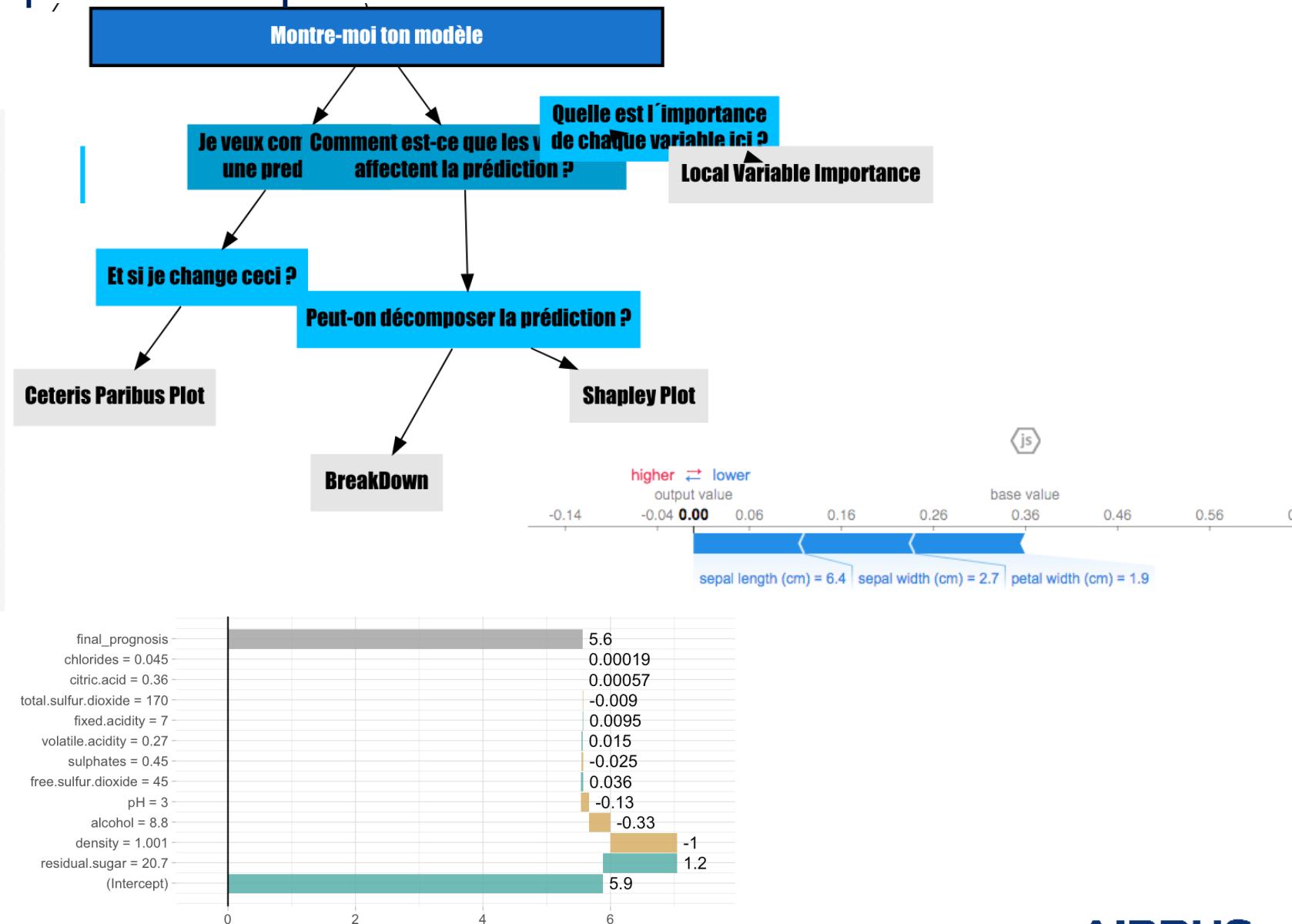
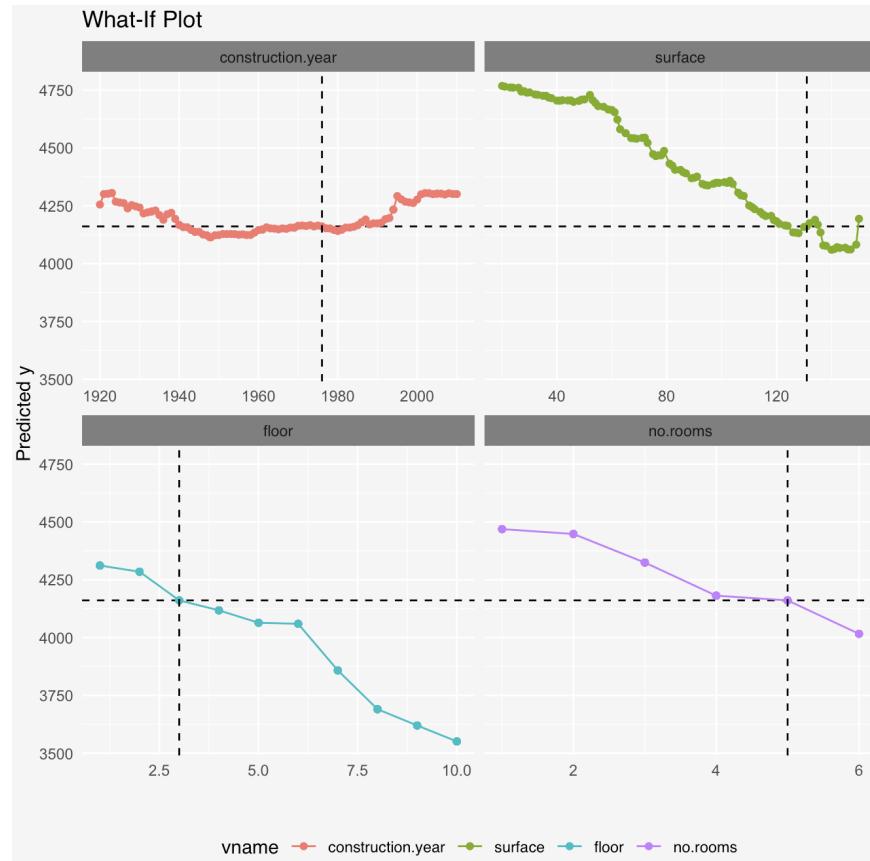
Montre-moi ton modèle : la problématique



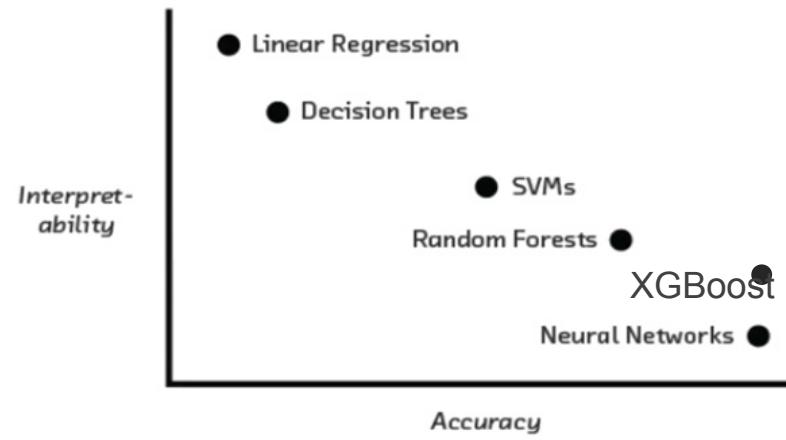
Montre-moi ton modèle : la problématique



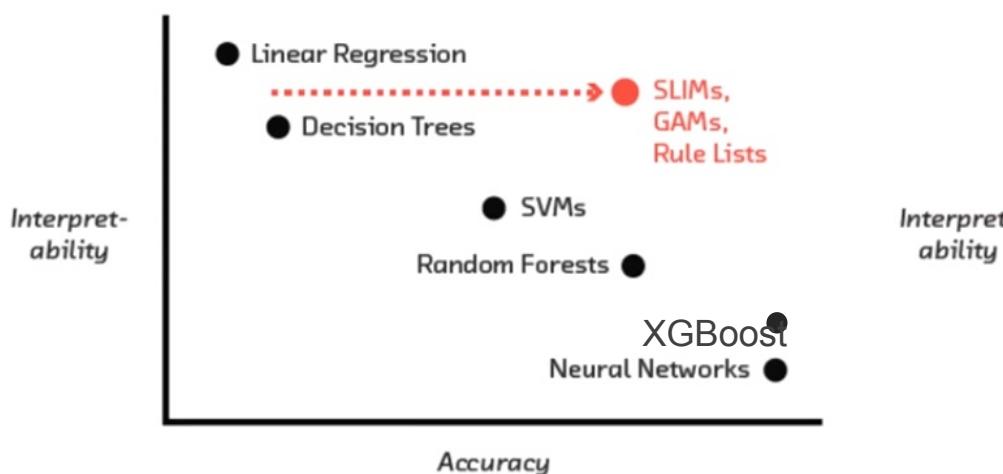
Montre-moi ton modèle : la problématique



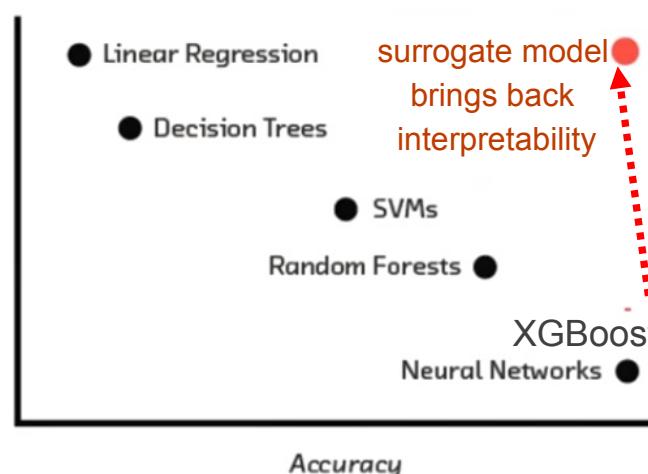
Exemple pour des données tabulaires



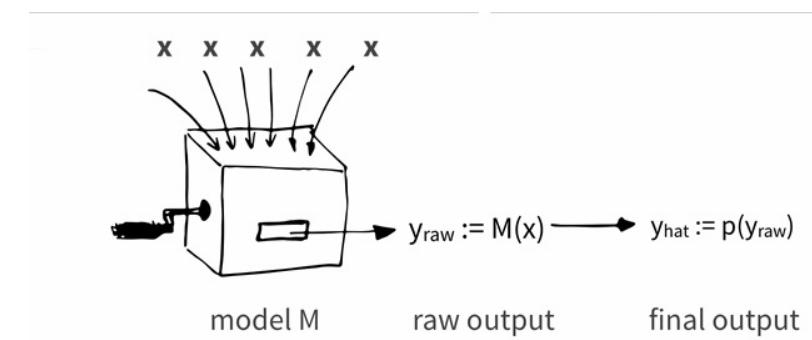
Utiliser des modèles moins célèbres



Construire un «modèle de remplacement»



Conditional Expected Responses



La complexité du problème: les différents formats de donnée

- Donnée tabulaire : (1 ligne par observation, 1 colonne par prédicteur) Classification
- Prédicteurs numériques
 - Prédicteurs categoriels
-
- Images data Régression linéaire
 - Text data
 - Audio data
 - Video data
 - Graph data

Prérequis: montre-moi tes données & construit 3 modèles

Description: df[,6] [6 × 6]

	prix_m2 <dbl>	année_construction <dbl>	surface <dbl>	étage <int>	nb_chambre <dbl>	quartier <fctr>
1	5897	1953	25	3	1	Saint-Georges
2	1818	1992	143	9	5	Patte D'oie
3	3643	1937	56	1	2	Capitole
4	3517	1995	93	7	3	Les Chalets
5	3013	1992	144	6	5	Saint-Cyprien
6	5795	1926	61	6	2	Saint-Georges

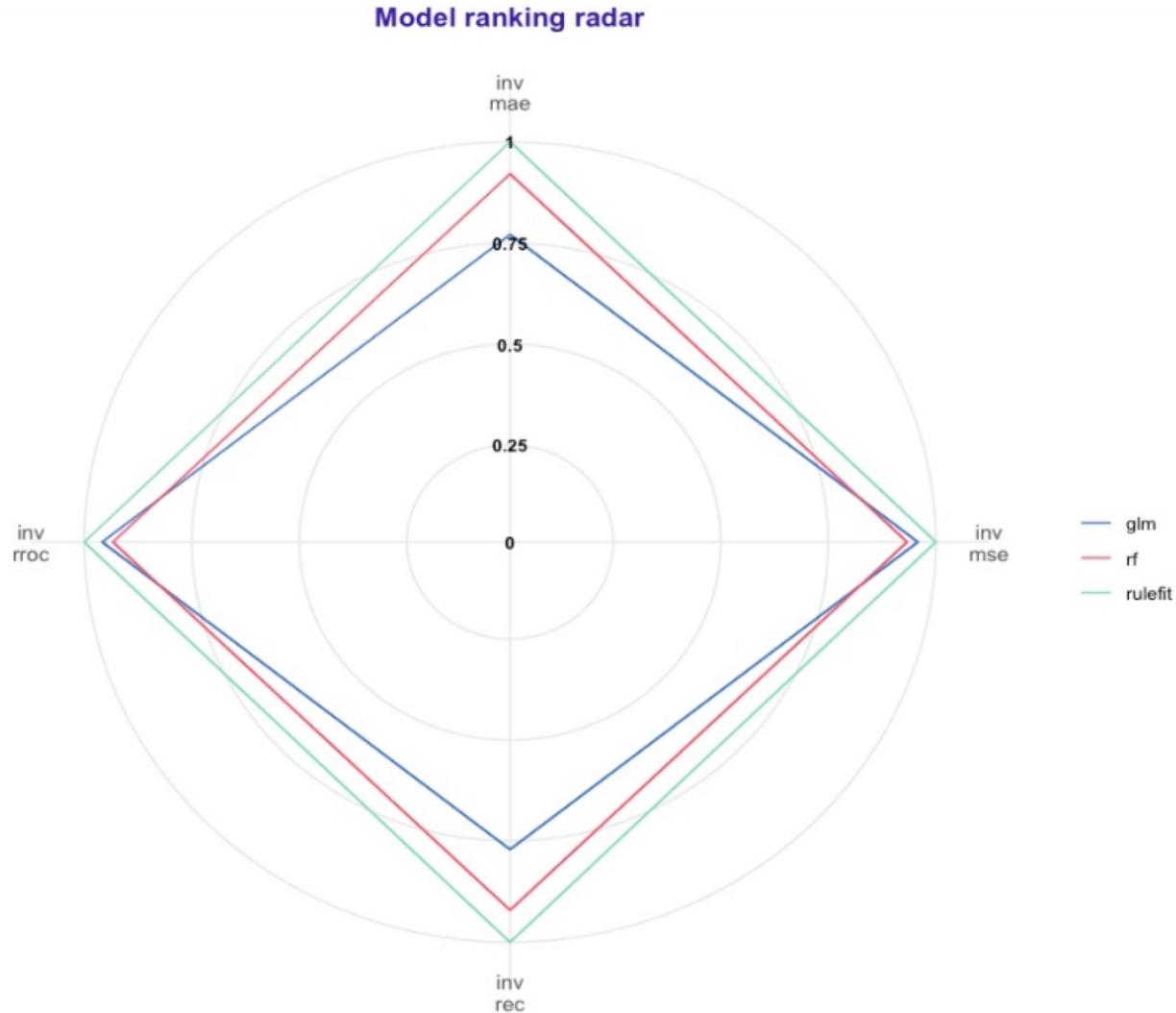
```
## Random forest    rf_model ← ranger(prix_m2 ~ ., data = appartements, num.trees = 300,
#                                     importance = "impurity", seed = 2323)
rf_predict ← predict(rf_model, appartements_test)
sqrt(mean((rf_predict$predictions - appartements_test$prix_m2)^2))
```
[1] 286.8595

Linear model glm_model ← glm(prix_m2 ~ ., data = appartements)
glm_predict ← predict(glm_model, appartements_test)
sqrt(mean((glm_predict - appartements_test$prix_m2)^2))
```
[1] 283.0865

## RuleFit model     rulefit_model ← h2o.rulefit(x=predictor,
#                                         training_frame = appartements_h2o, model_type = "rules_and_linear",
#                                         max_num_rules = 100, min_rule_length = 2,
#                                         max_rule_length = 7
# )
rulefit_predict ← h2o.predict(rulefit_model, newdata = appartements_test_h2o)
sqrt(mean((rulefit_predict - appartements_test_h2o$prix_m2)^2))
[1] 274.001
```

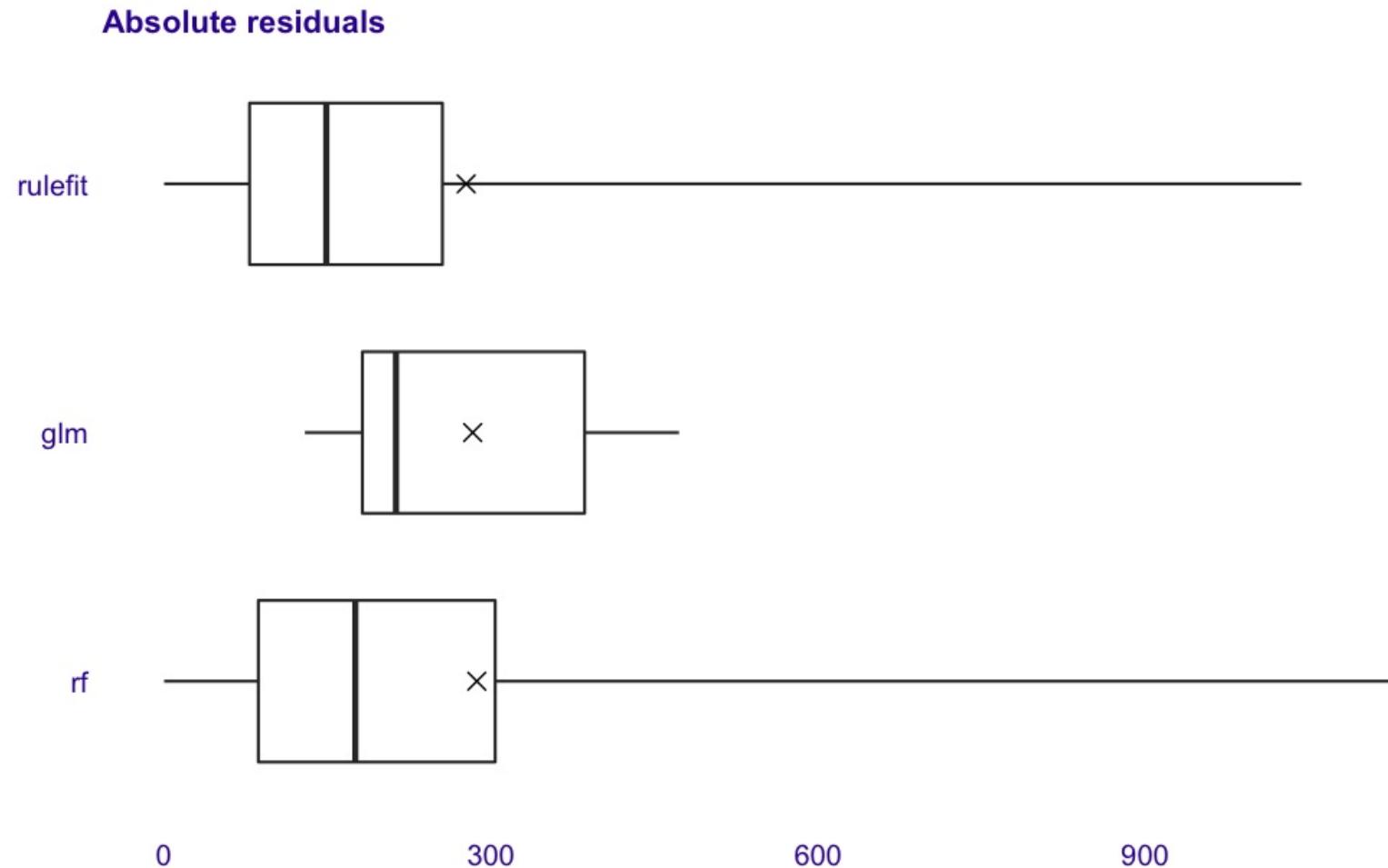
Performance des modèles : Comparaison des modèles entre eux 1/2

Radar de classement des modèles (model ranking radar)

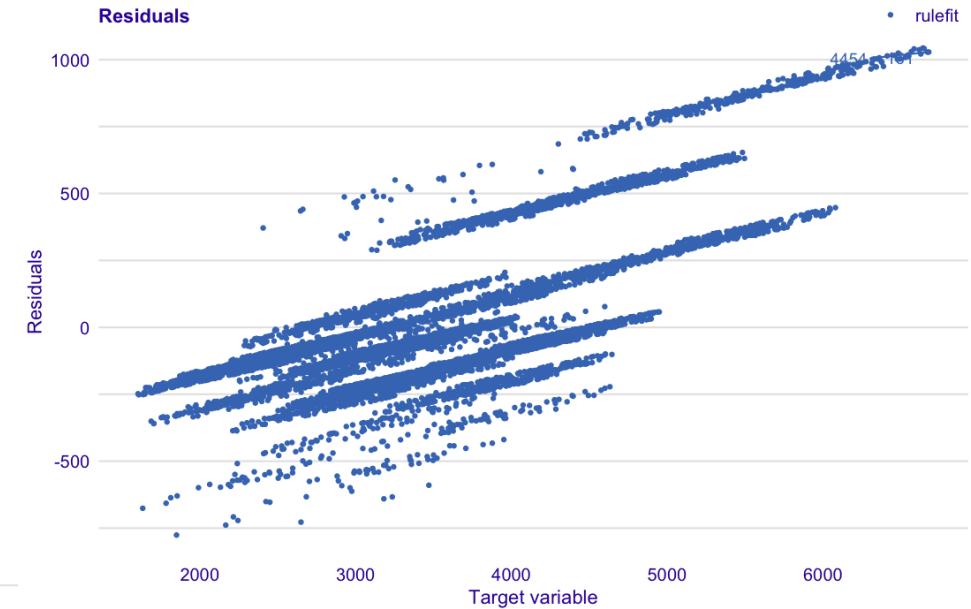
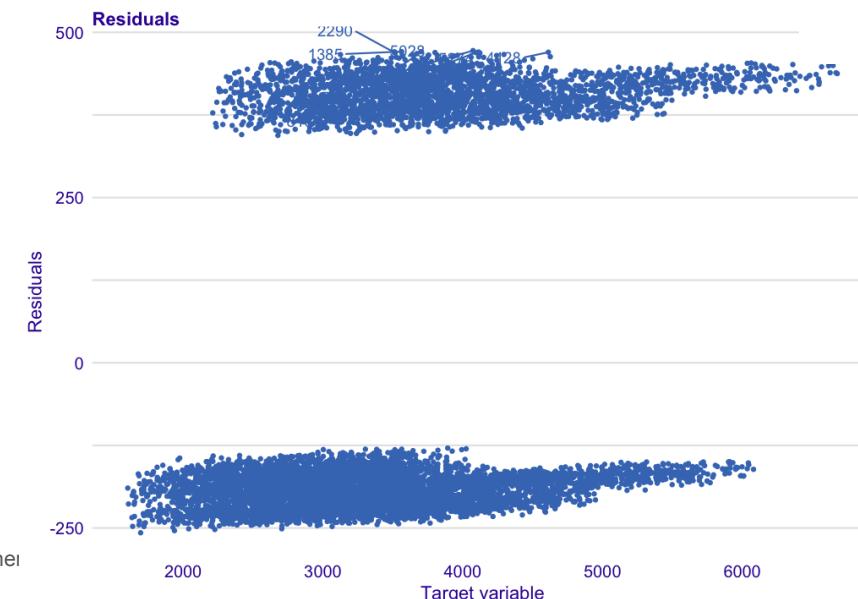
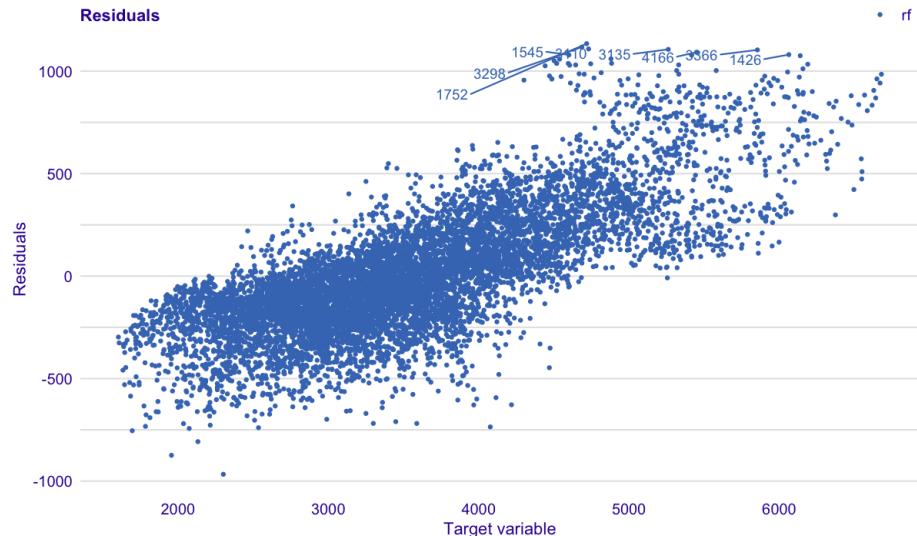


Performance des modèles : Comparaison des modèles entre eux 2/2

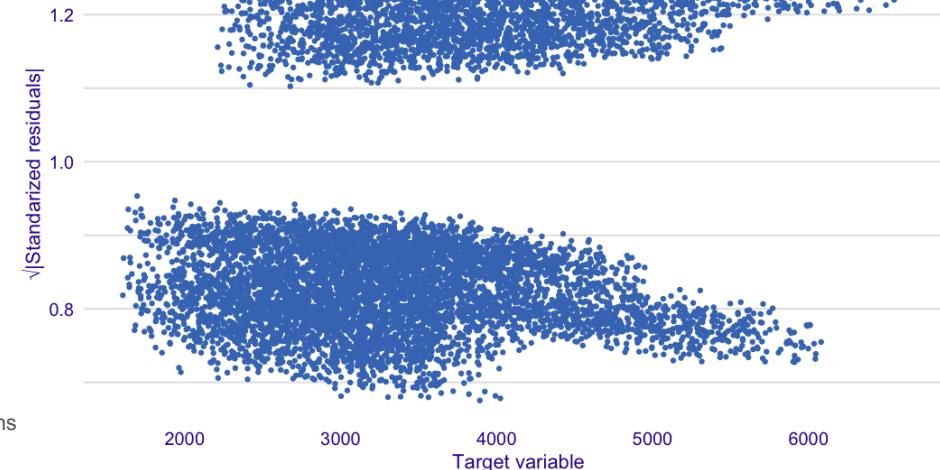
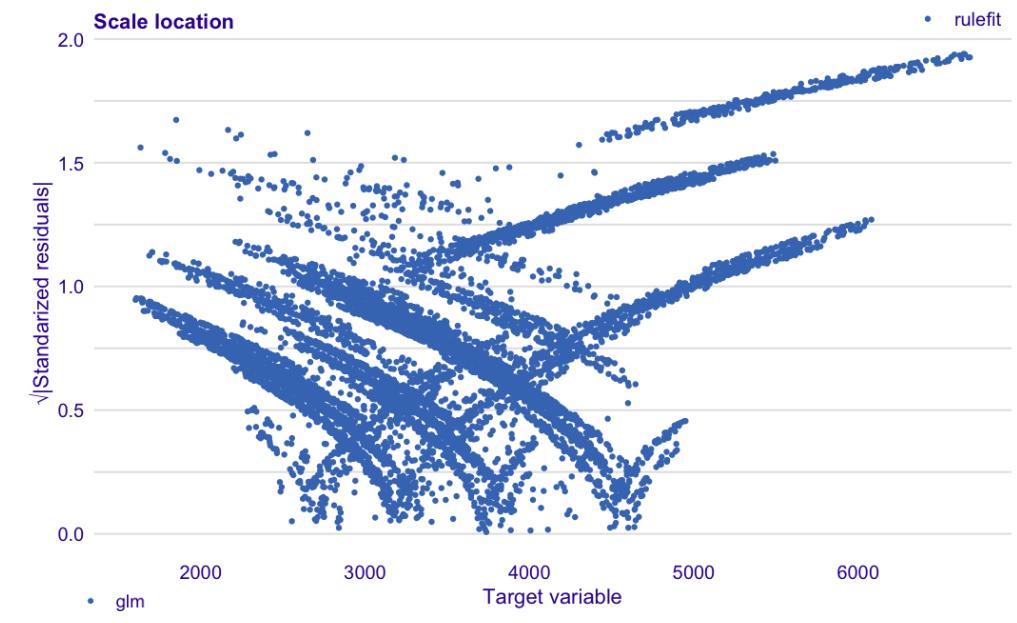
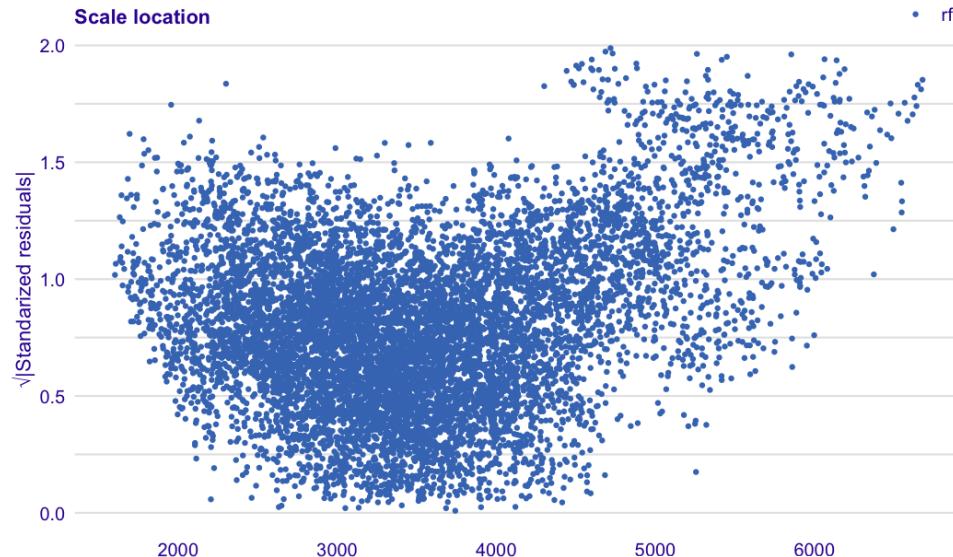
Boîte à moustache des résidus en valeur absolue (Boxplot of Absolute Residuals)



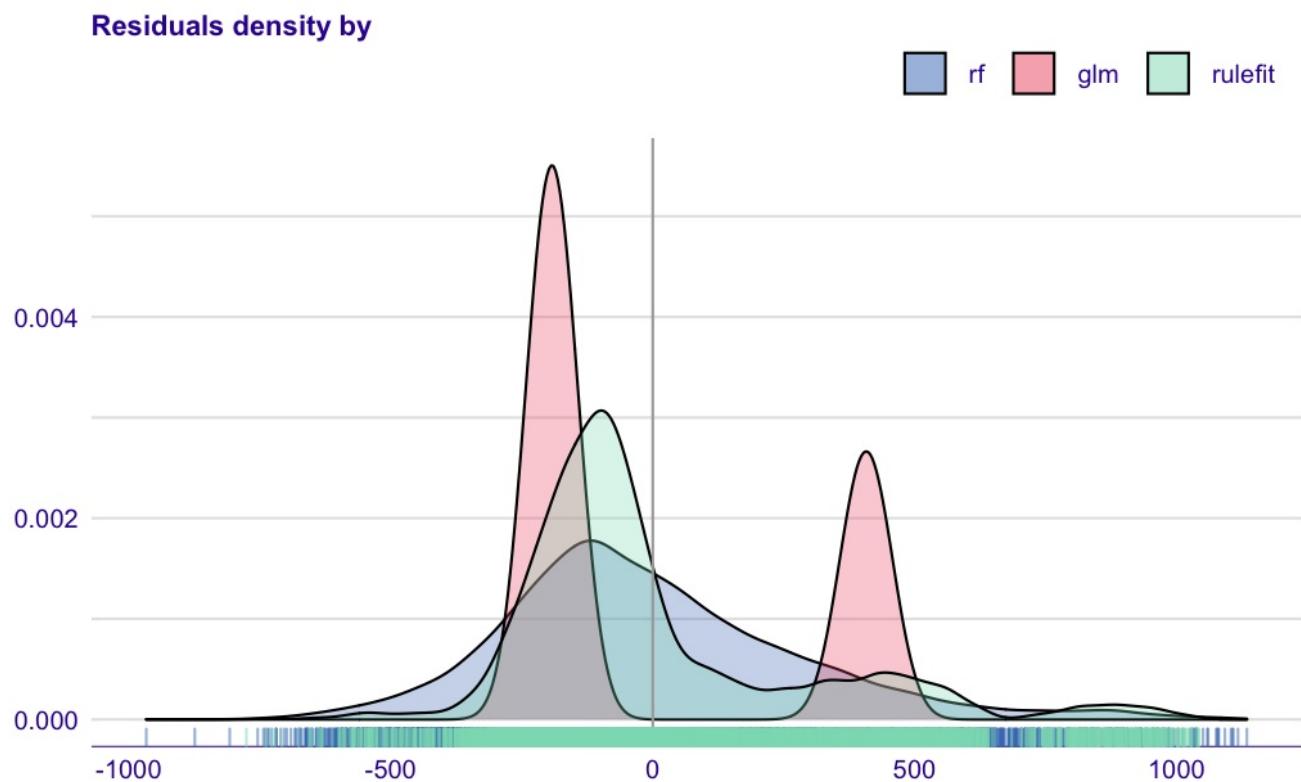
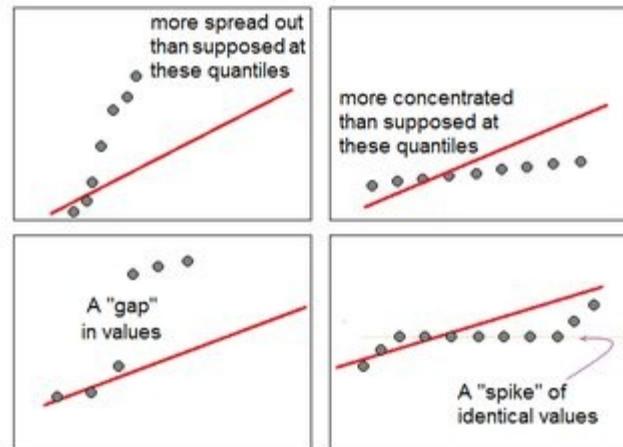
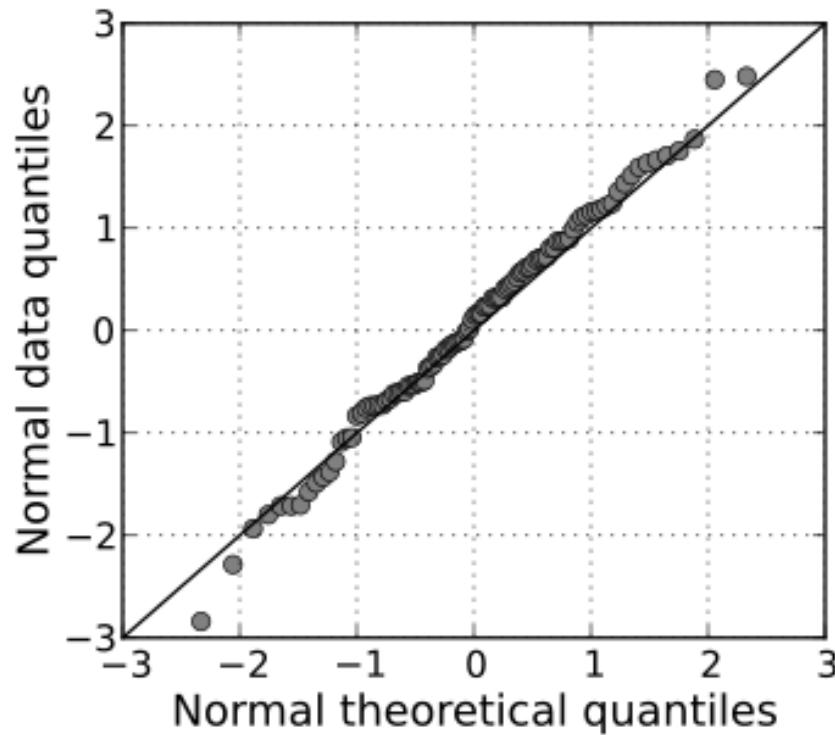
Diagnostique sur les résidus 1/3: nuage de point de diagnostique des résidus (residuals diagnostic plot)



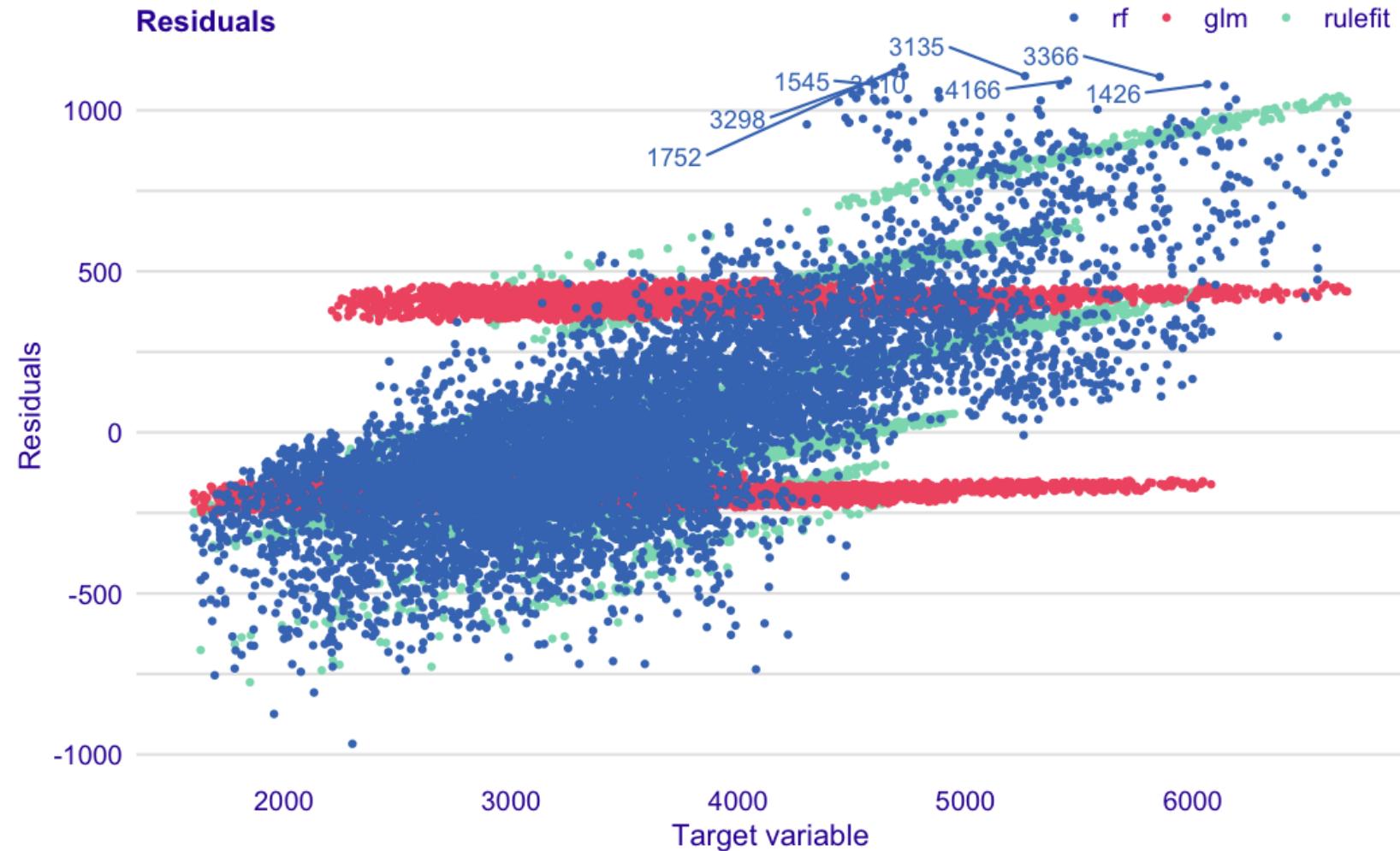
Diagnostique sur les résidus 2/3: nuage de point de diagnostique des résidus (scale location plot)



Diagnostique sur les résidus 3/3 : Quantile-Quantile plot (Q-Q plot) et plot de densité des résidus (residual density plot)



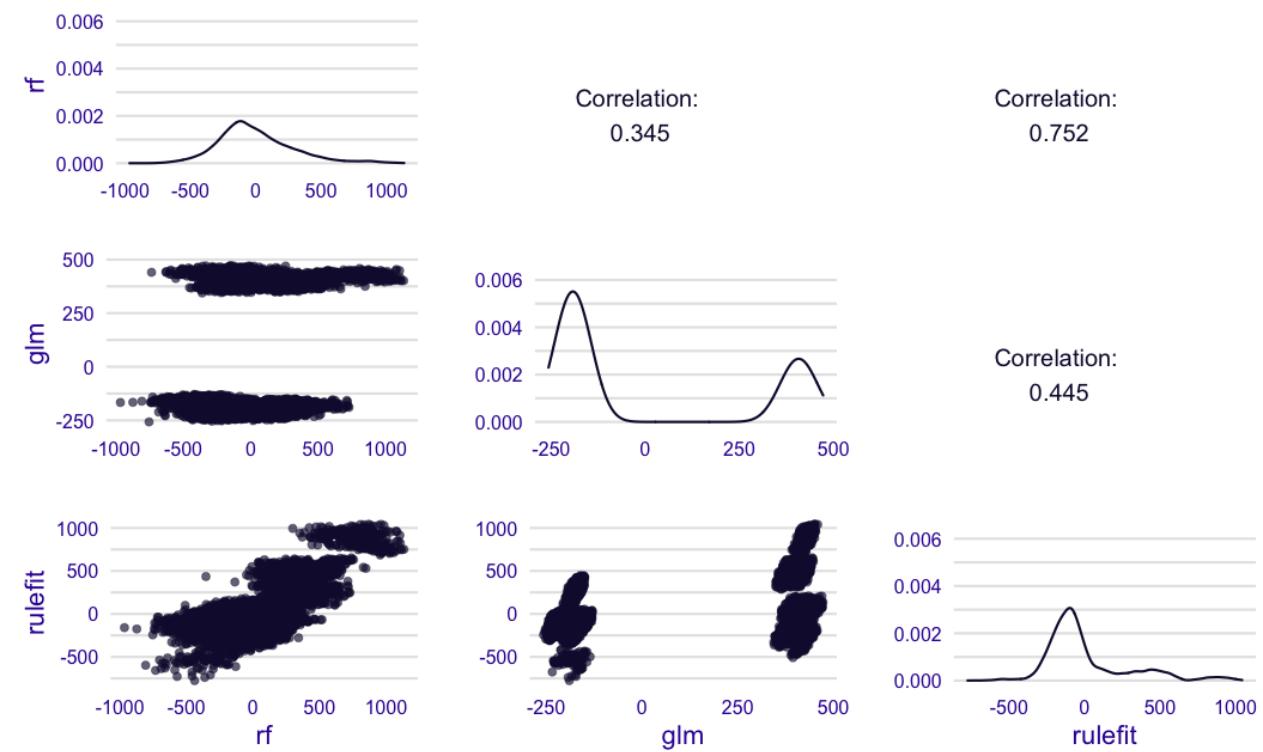
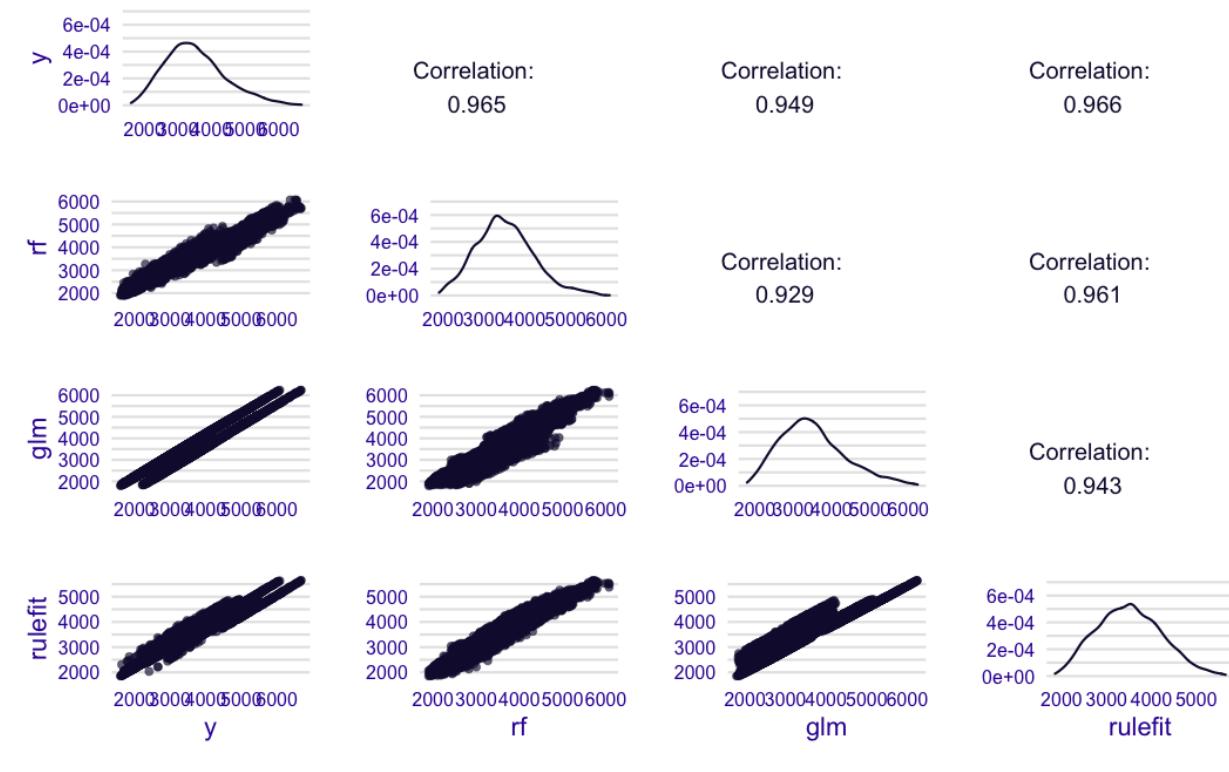
Comparaison des modèles entre eux 2/3 : nuage de point de diagnostique des résidus (residuals diagnostic plot)



Comparaison des modèles entre eux 3/3 :

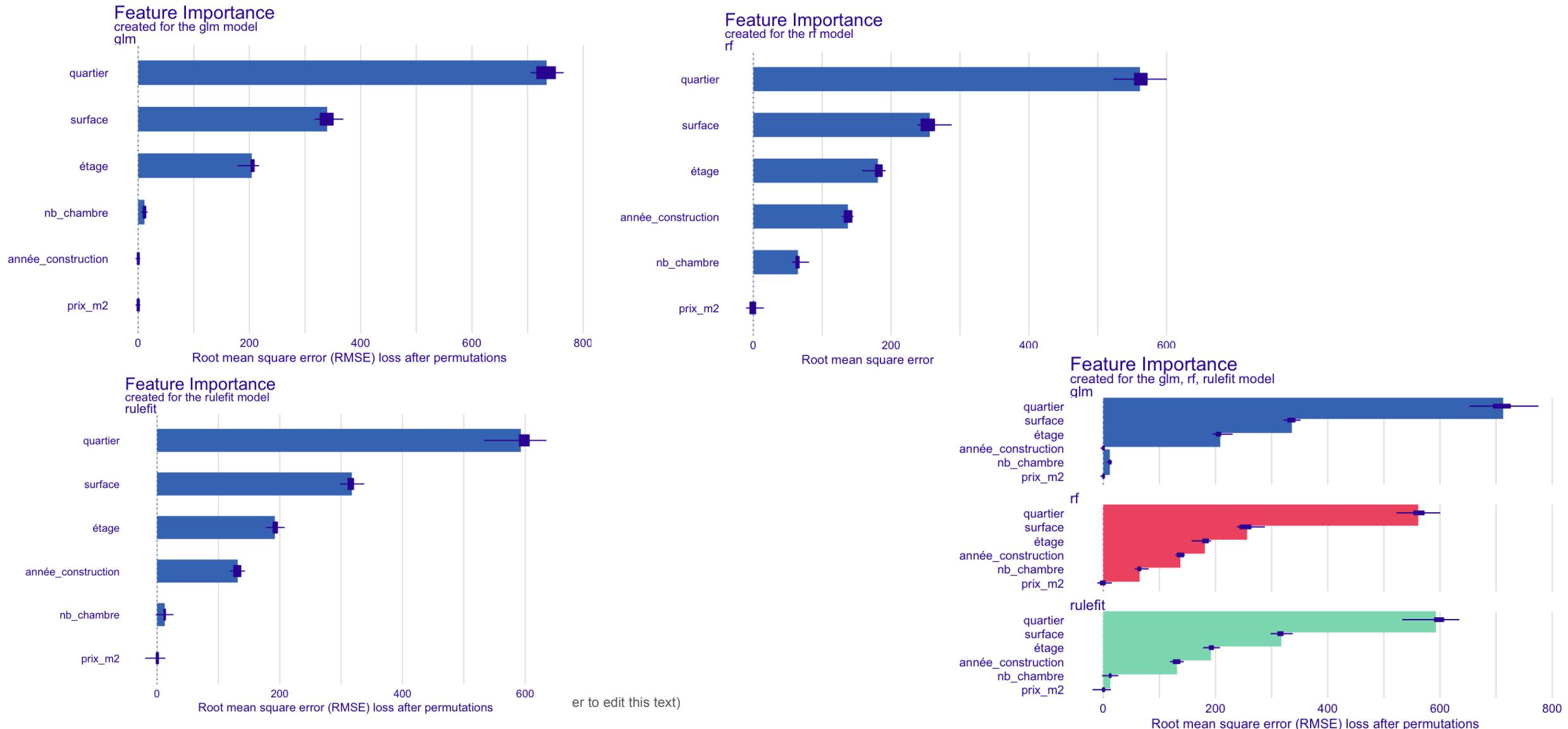
Diagramme de corrélation des prédition
(prediction correlation plot)

et de corrélation des residus
(residuals correlation plot)

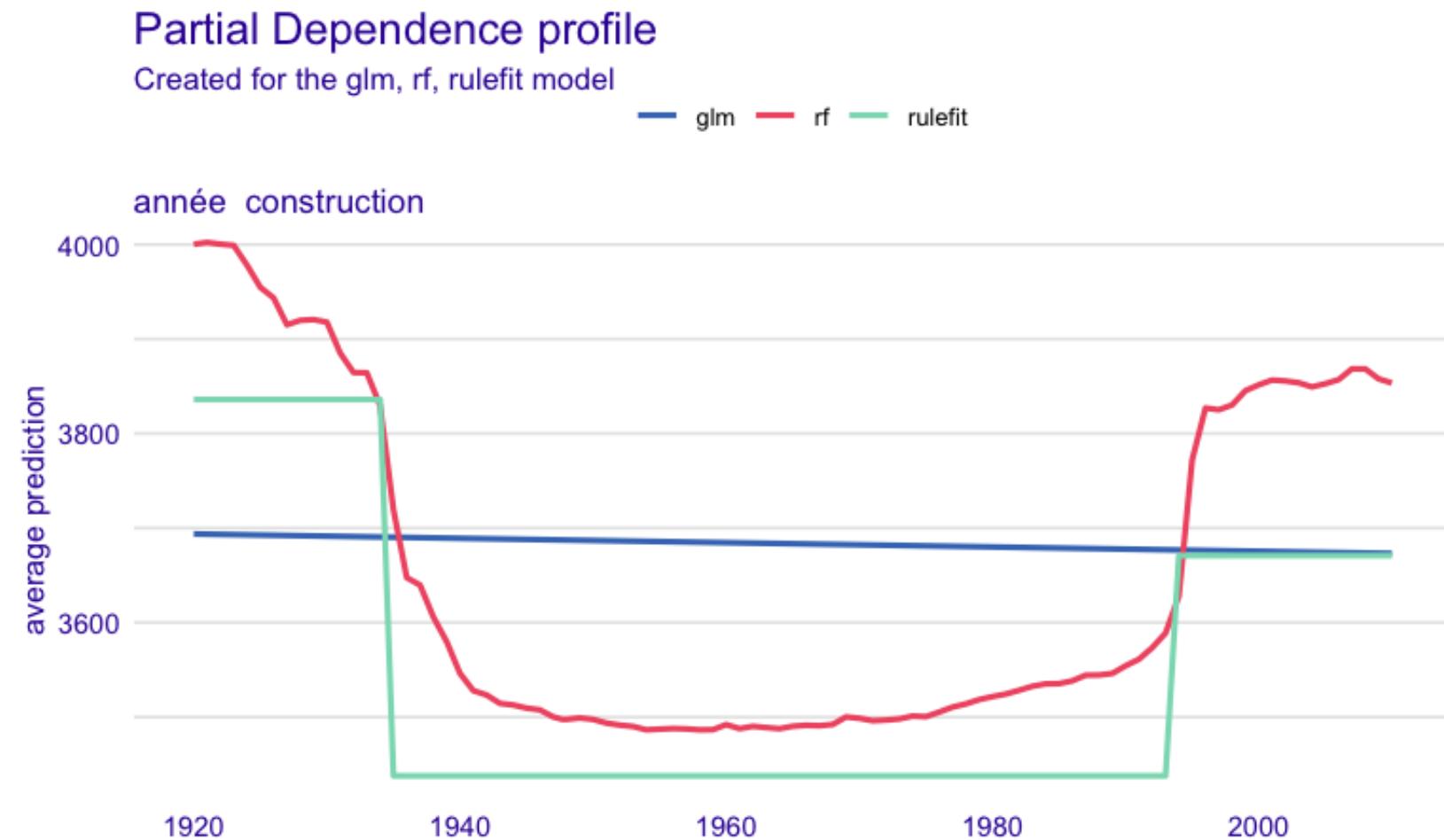


Compréhension globale du modèle:

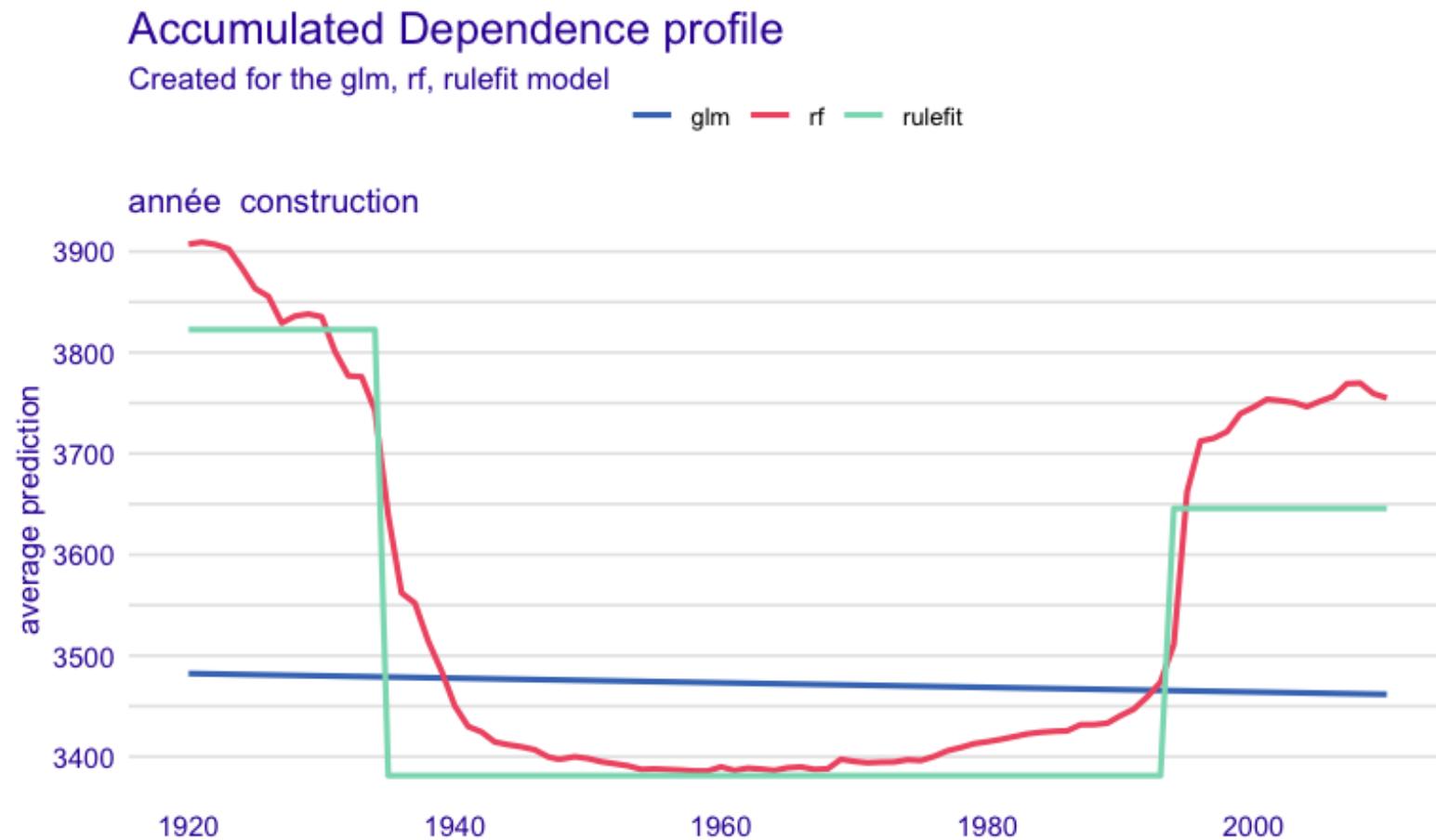
Graphique d'importance des variables (Global Variable Importance)



Compréhension du modèle: réponse à une variable continue : le Diagramme de dépendance partielle ou PDP Plot (Single variable, continuous)



Compréhension du modèle: réponse à une variable (Unique, continue) ALE plot



Compréhension du modèle - réponse à une variable : Individual conditional expectation plots (ICE plot)

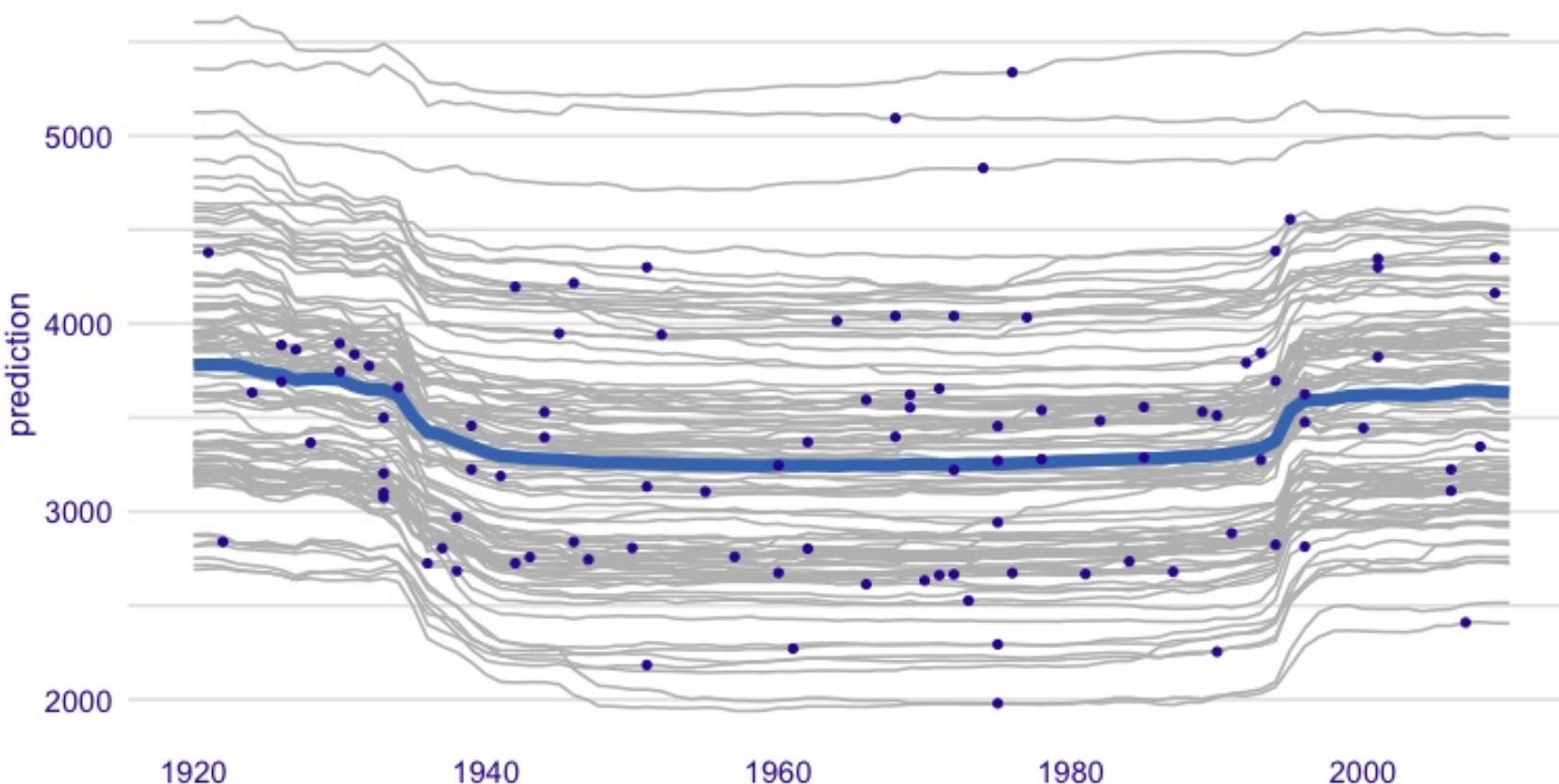
Compréhension du modèle: réponse à chaque variable continue : le Diagramme de dépendance partielle ou PDP Plot (Single variable, continuous)



Valider la structure du modèle localement : What-if plot avec les points de prediction

Ceteris Paribus profile

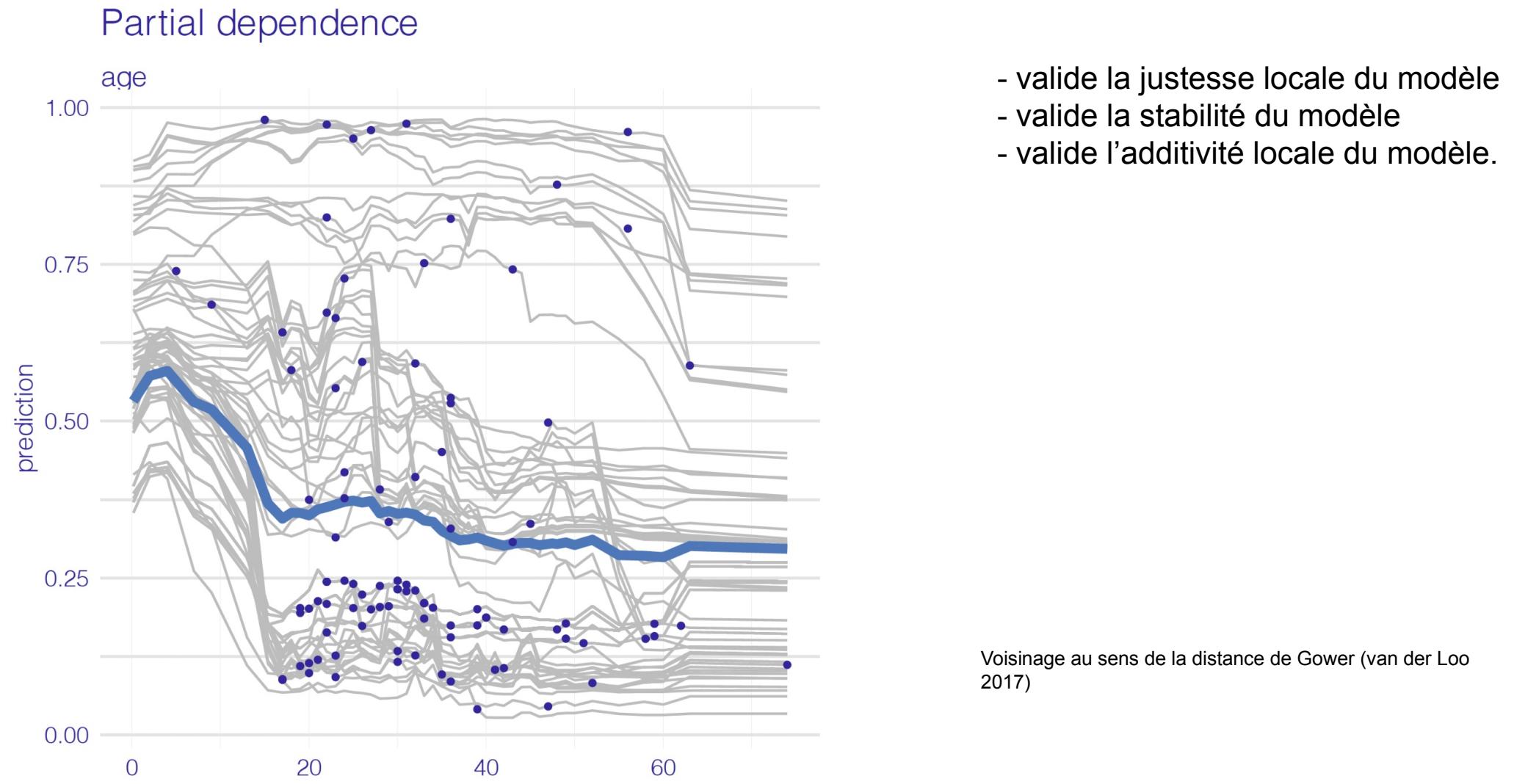
created for the rf model
année construction



- valide la justesse locale du modèle
- valide la stabilité du modèle
- valide l'additivité locale du modèle.

Voisinage au sens de la distance de Gower (van der Loo 2017)

Valider la structure du modèle localement : What-if plot avec des points voisins...

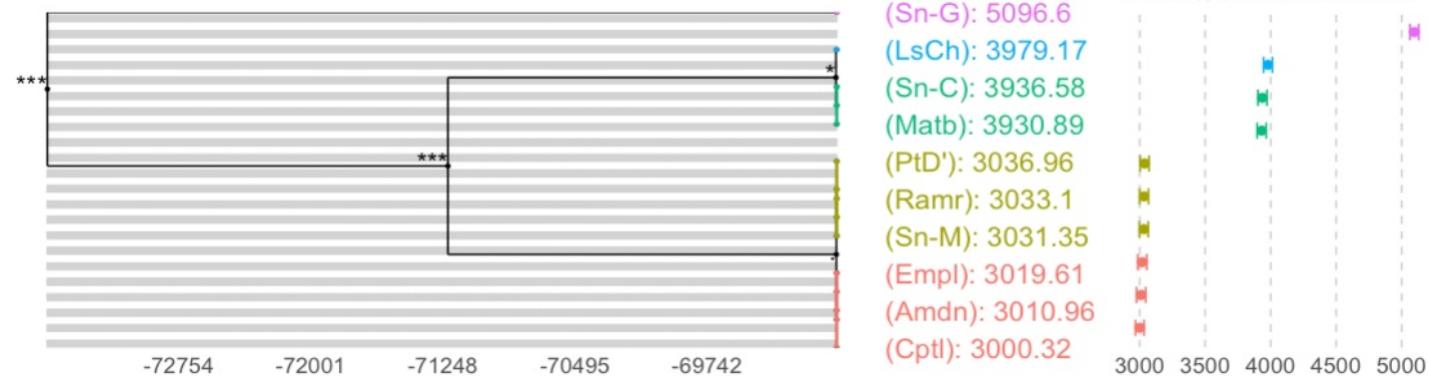


Compréhension du modèle: réponse à une variable (Unique, catégorielle)

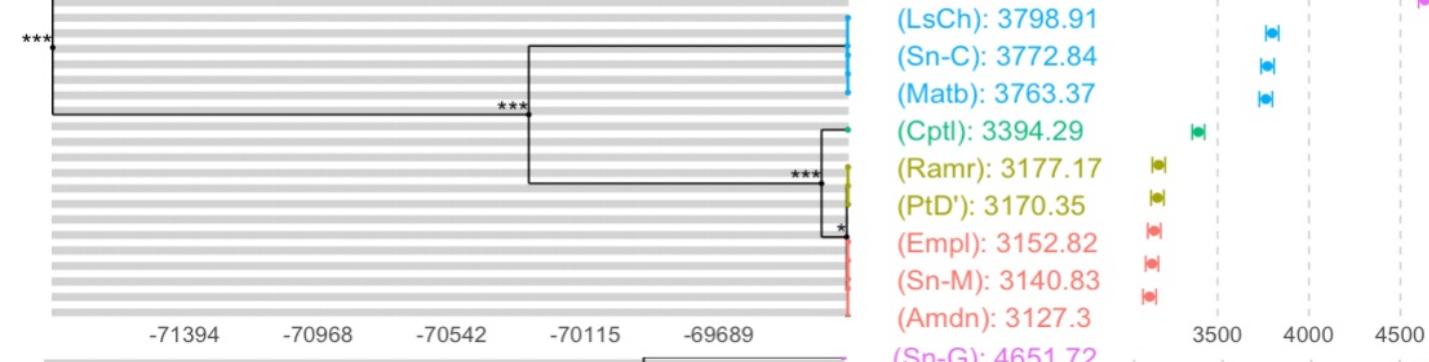
Merging Path Plot

Modèle linéaire

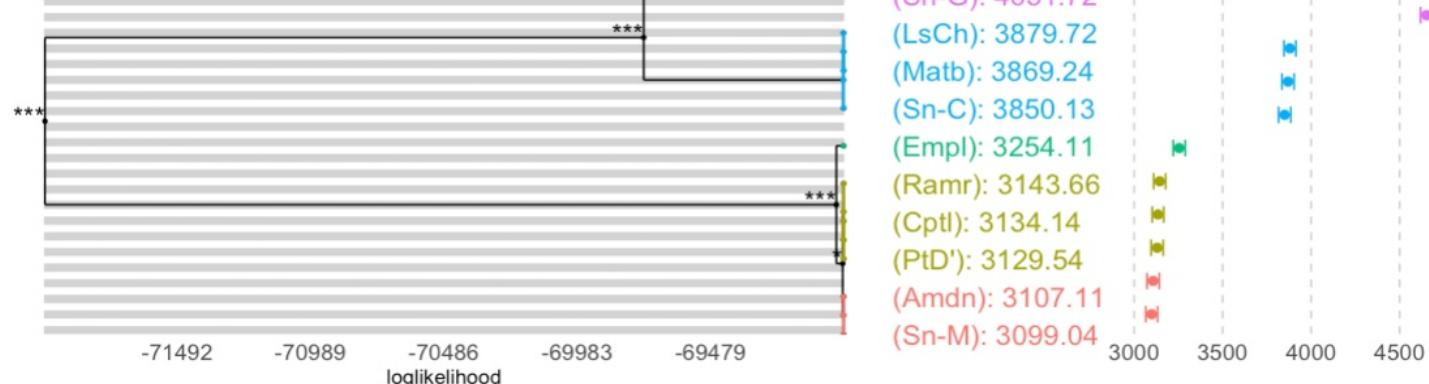
Factor Merger Tree



Modèle random-forest



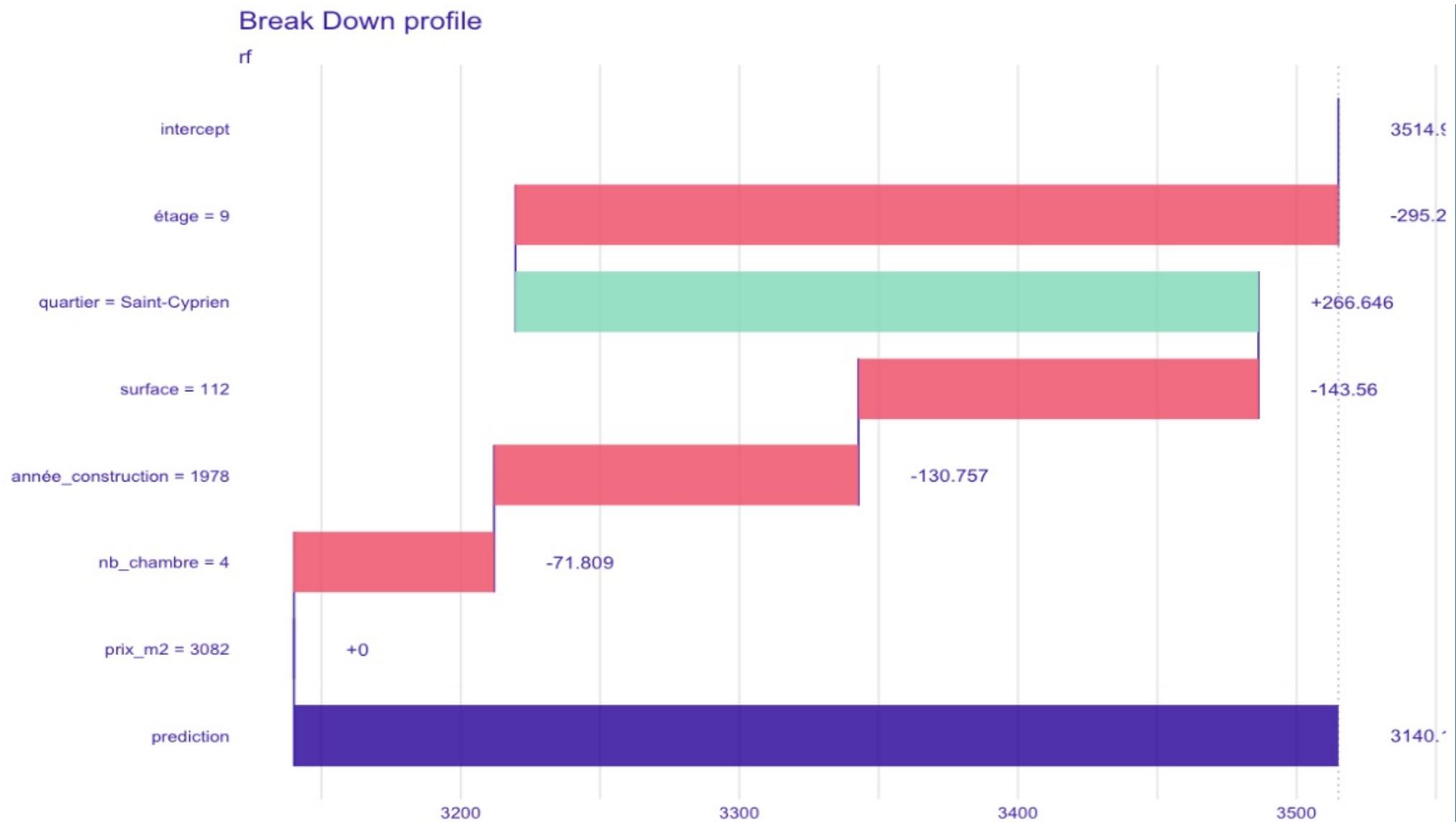
Modèle rulefit



Explication d'une prédition unique : Maintenant expliquez-moi...

Le prix de mon appartement construit en 1978, avec 4 chambres et de 22 m²,
situé au 9^{ème} étage dans le quartier Saint-Cyprien

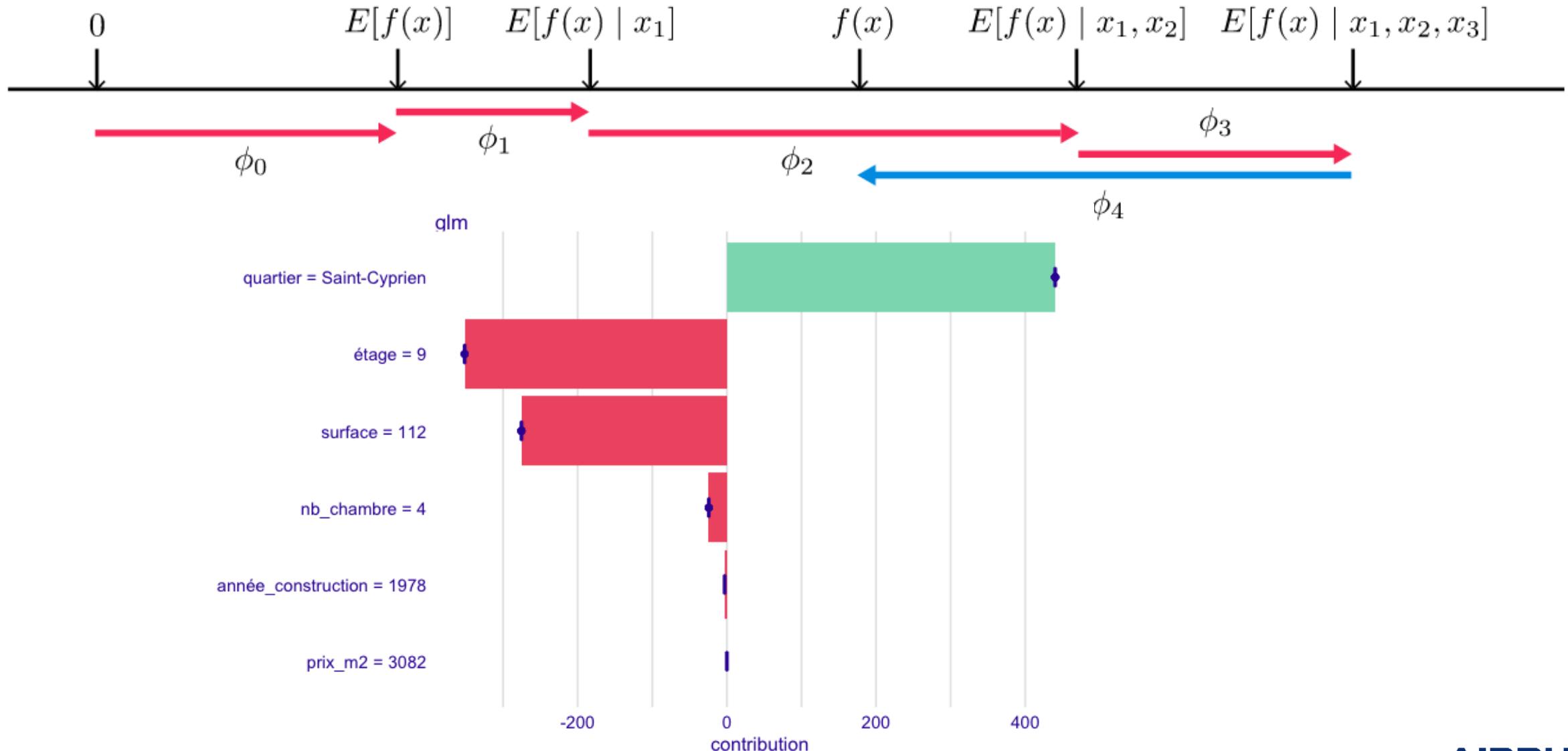
Explication d'une prédition unique: Breakdown / Waterfall plot



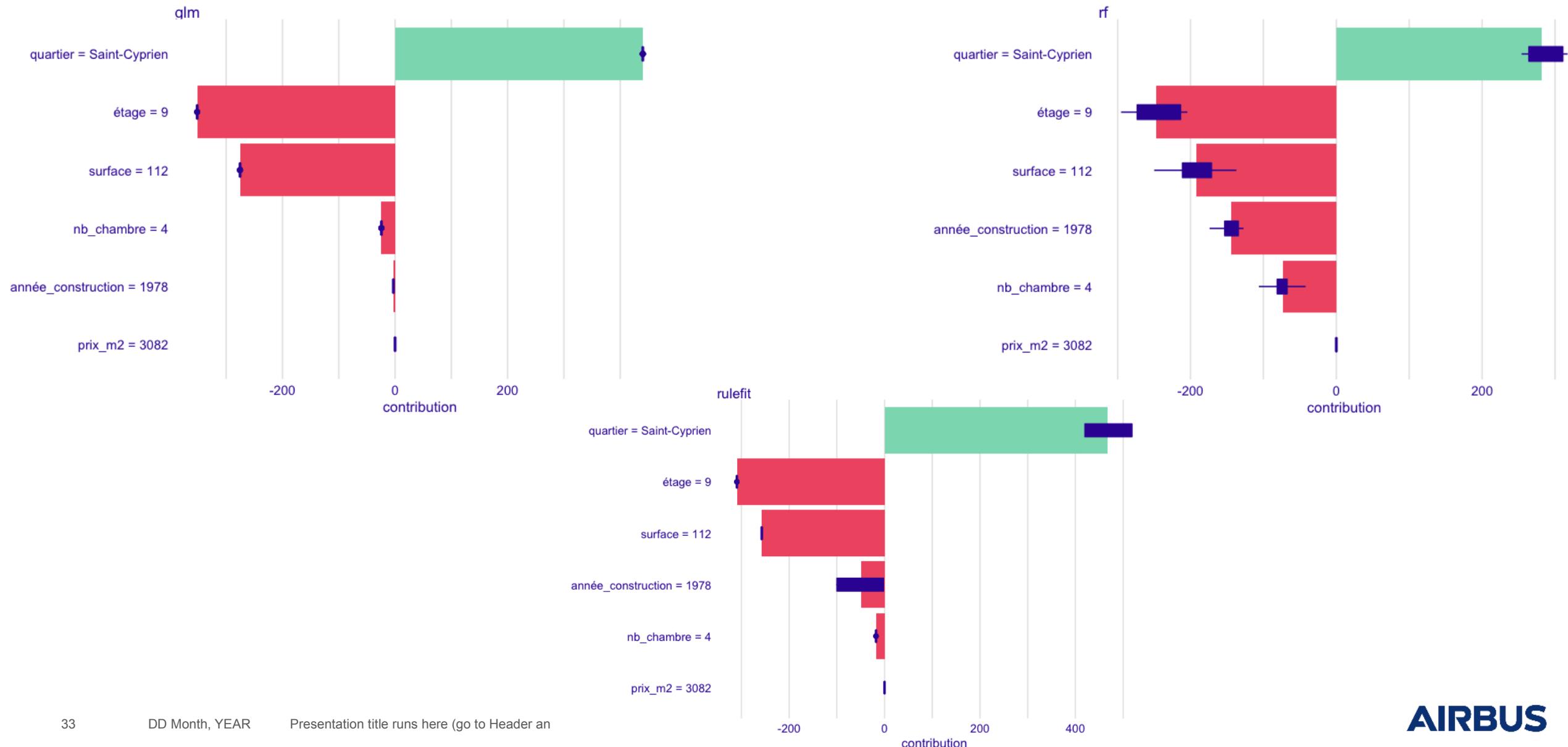
Explication d'une prédition unique: Breakdown / Waterfall plot



Explication d'une prédition unique : Shapley Plot

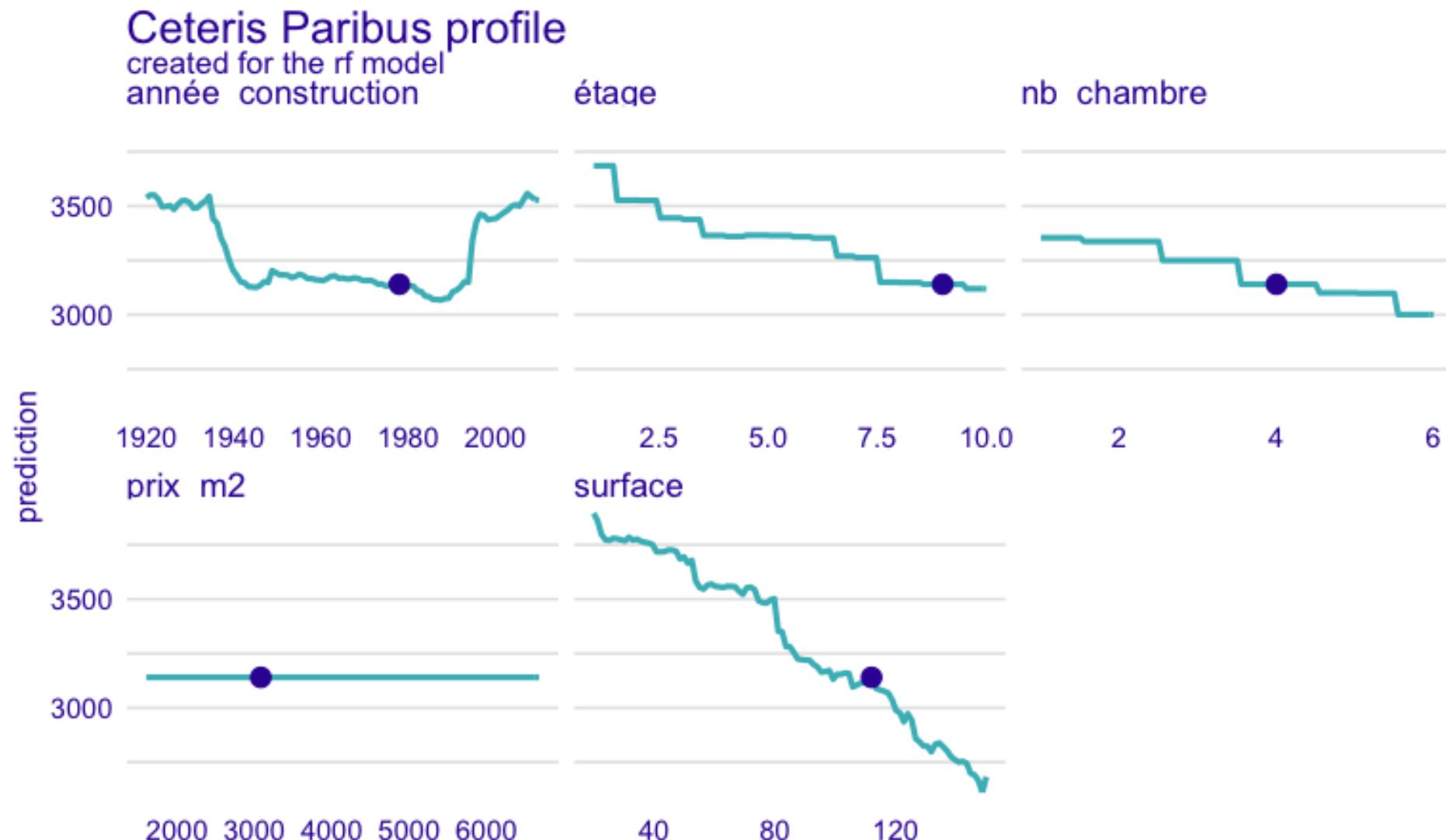


Explication d'une prédition unique : Shapley Plot



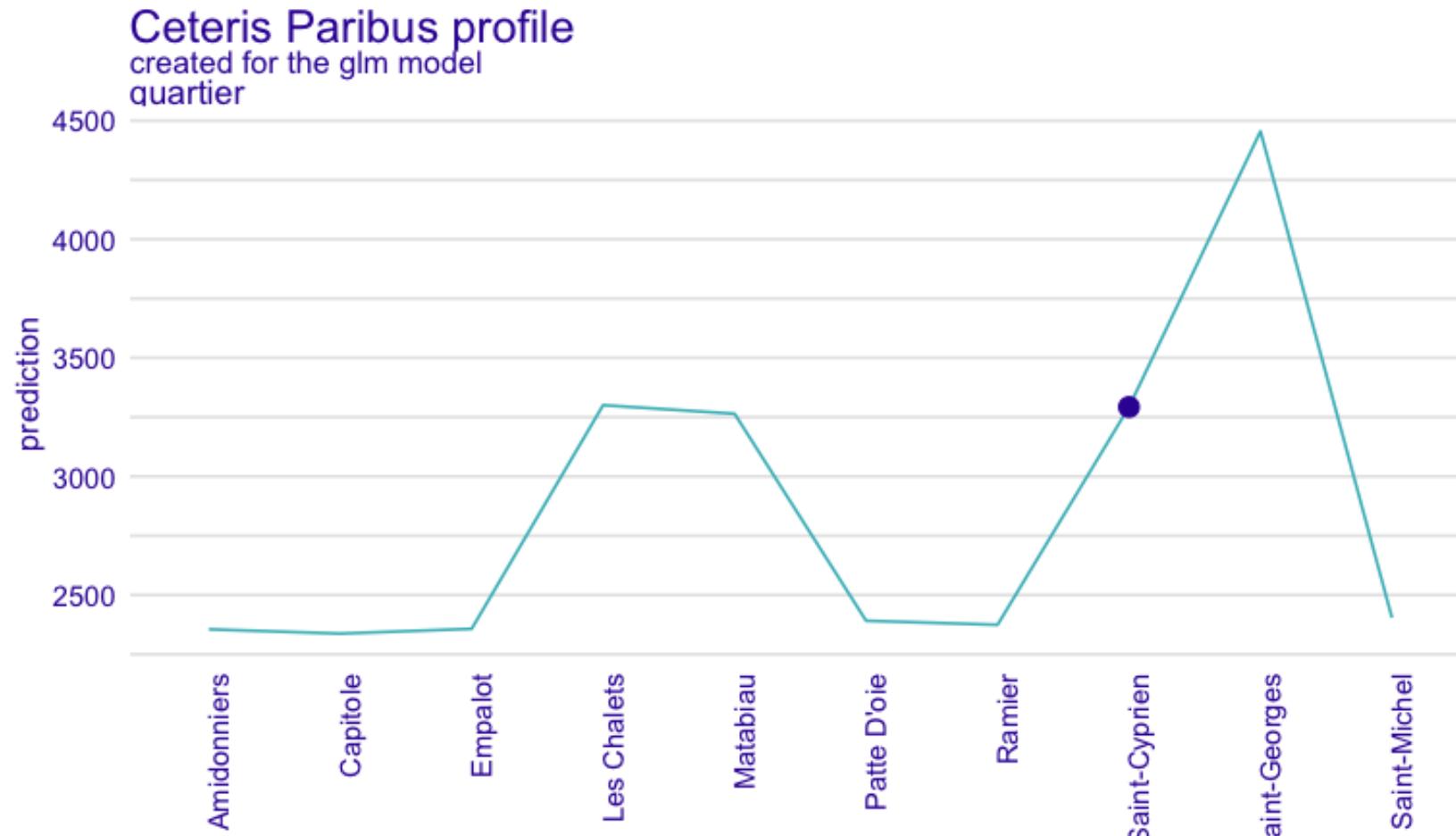
What-if plot : Ceteris Paribus plot, une seule prediction

Influence d'une variable numérique

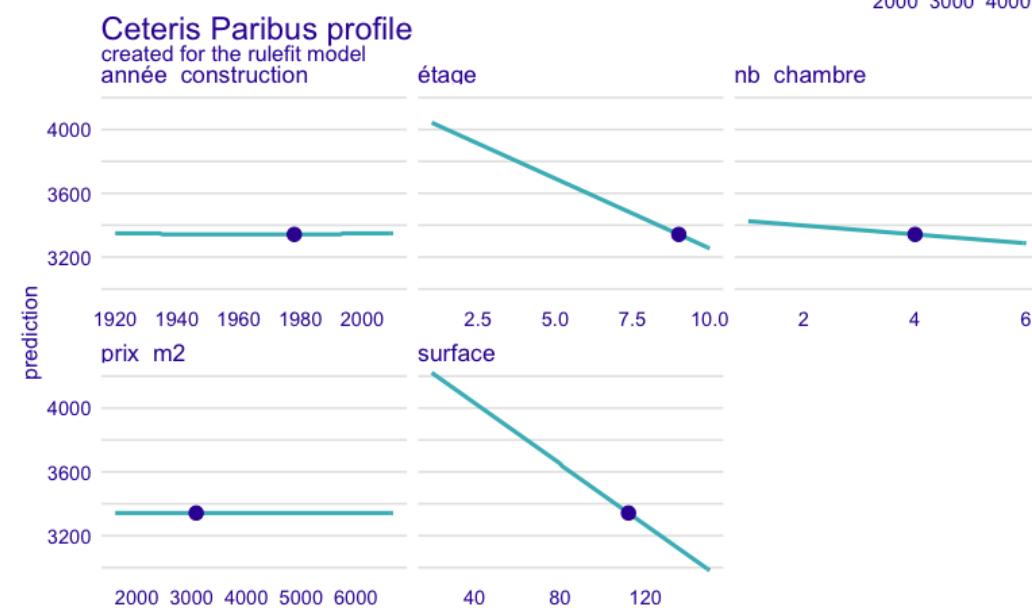
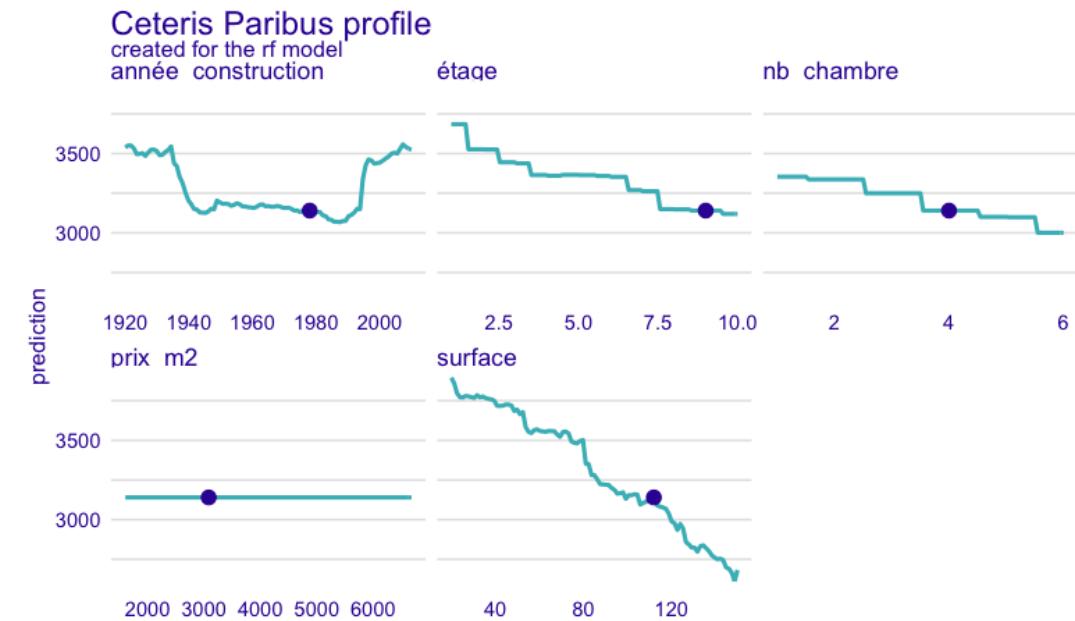
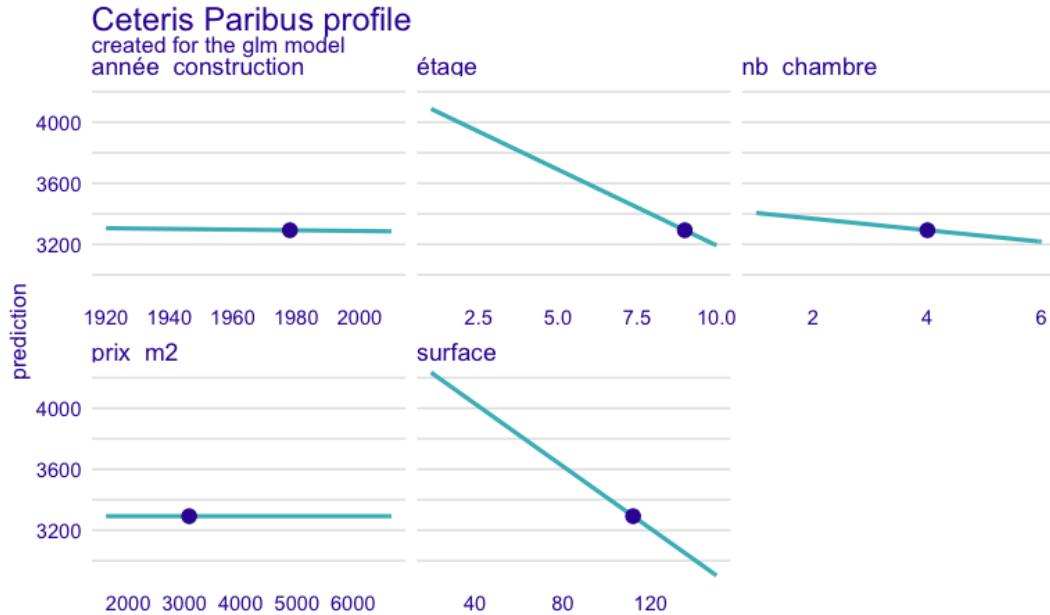


What-if plot : Ceteris Paribus plot, une seule prediction

Influence d'une variable catégorielle

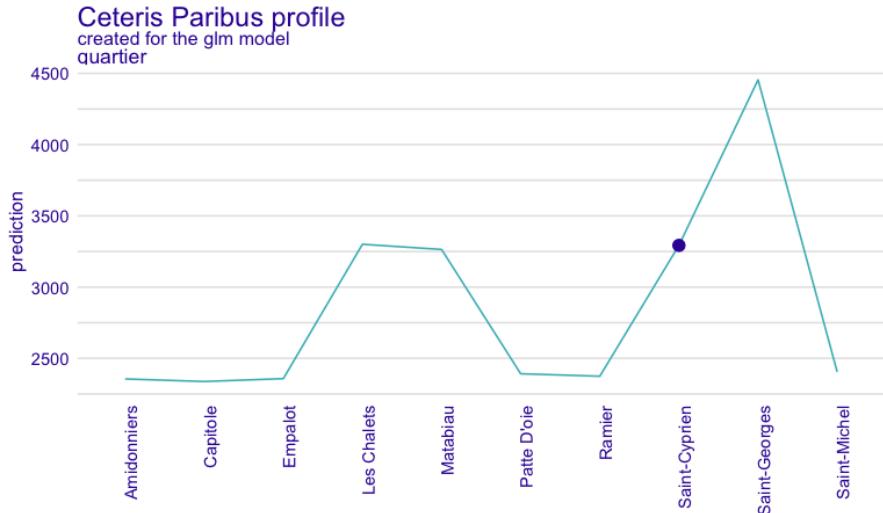


What-if plot : Ceteris Paribus plot, une seule prediction

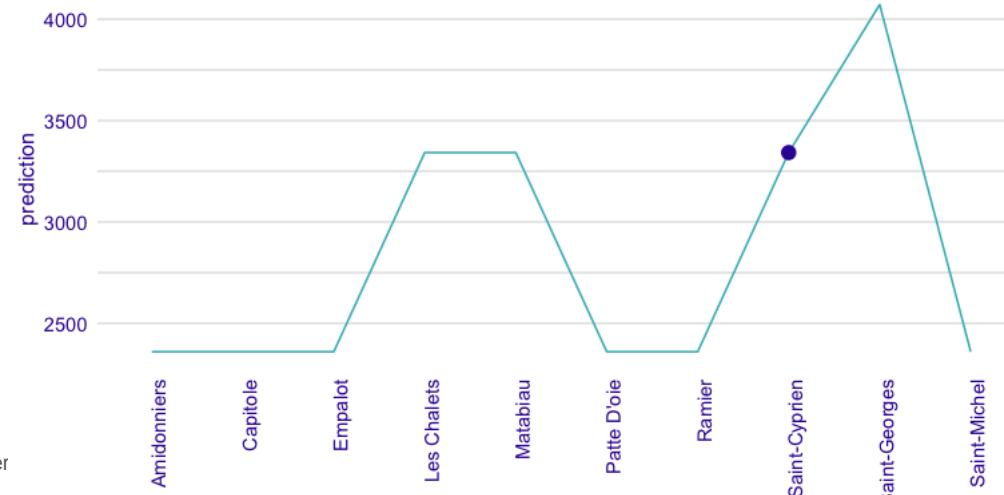


What-if plot : Ceteris Paribus plot, une seule prediction

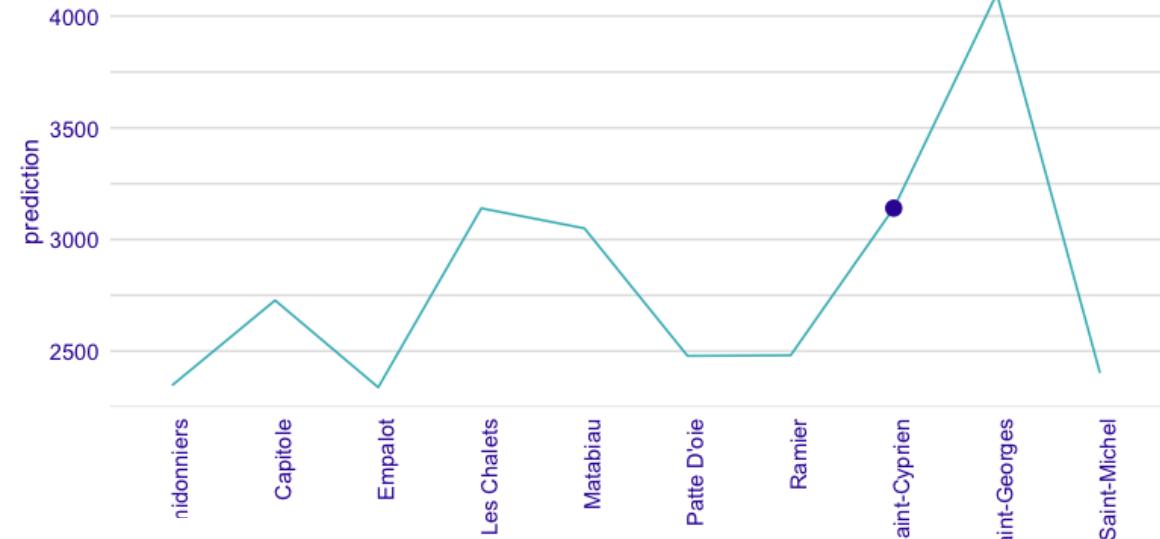
Influence d'une variable catégorielle



Ceteris Paribus profile
created for the rulefit model
quartier



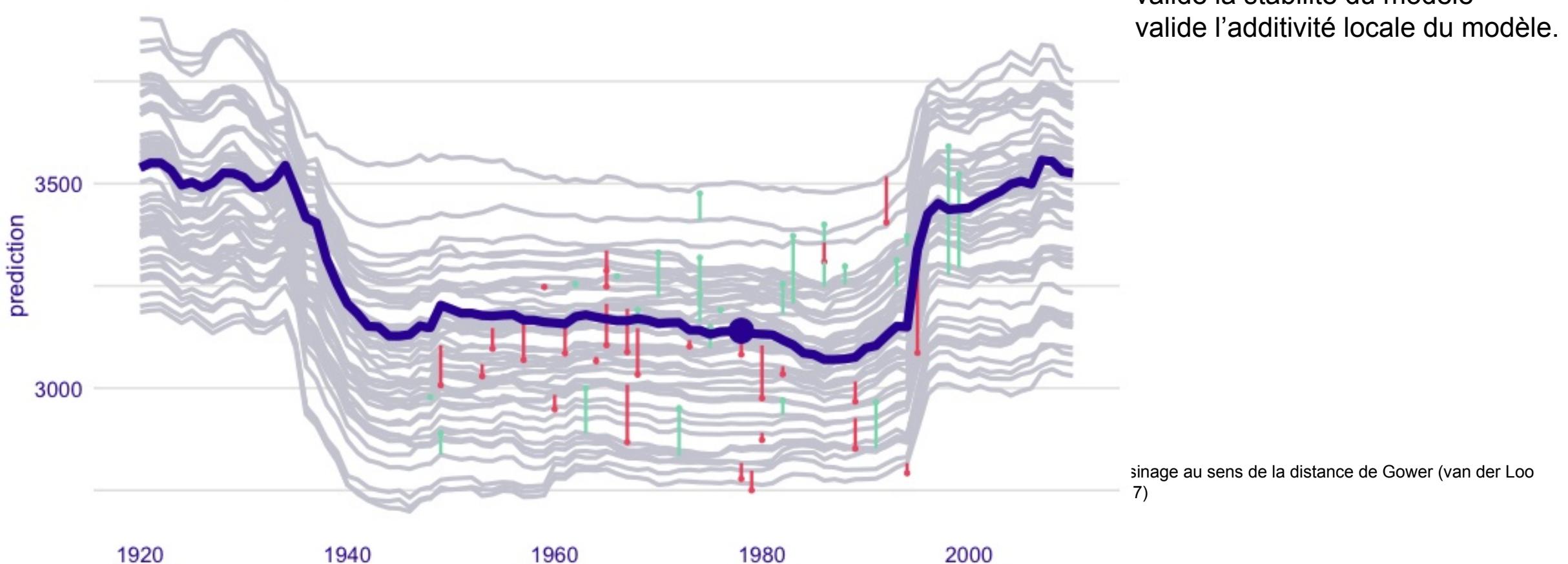
Ceteris Paribus profile
created for the rf model
quartier



Valider la structure du modèle localement : What-if plot avec des points voisins et leur résidus...

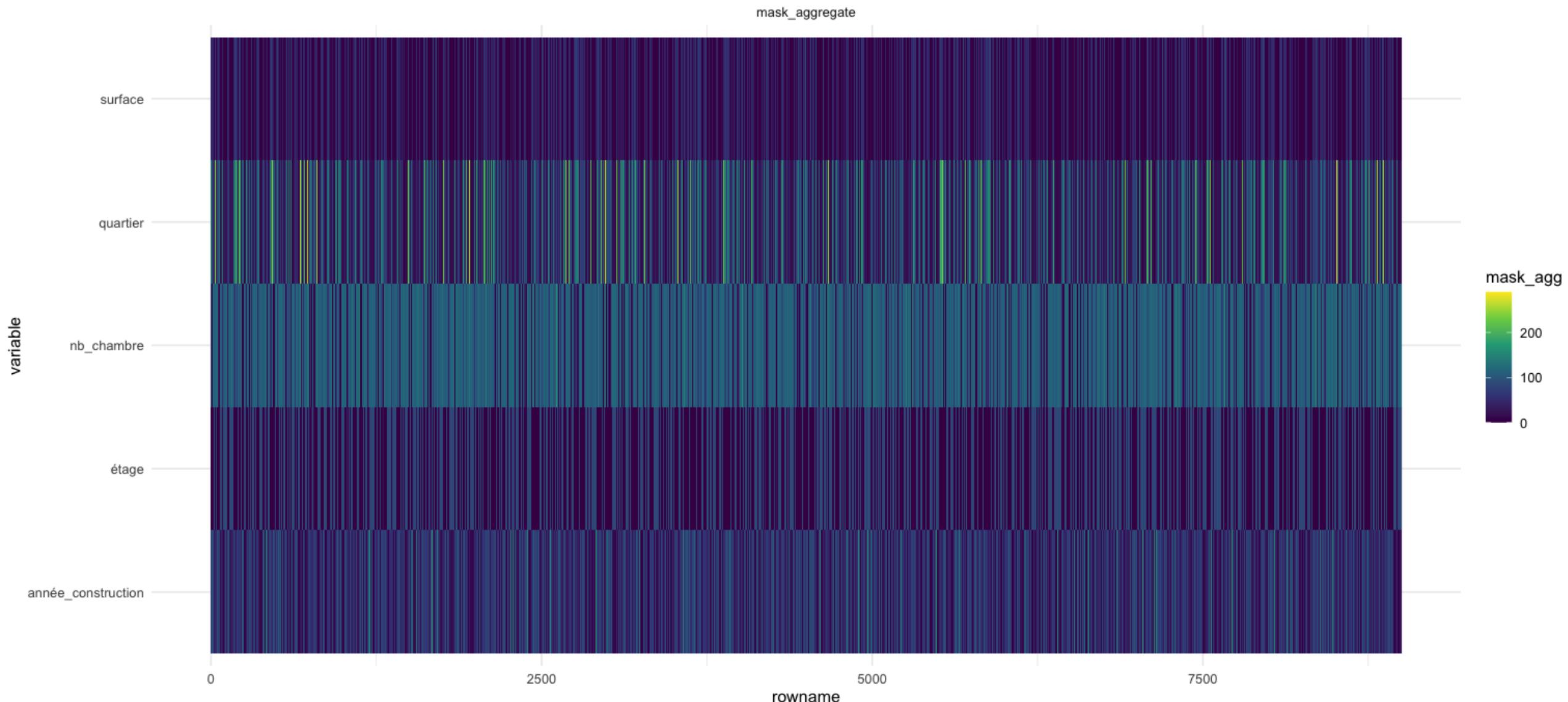
Local stability plot

created for the rf model
année construction

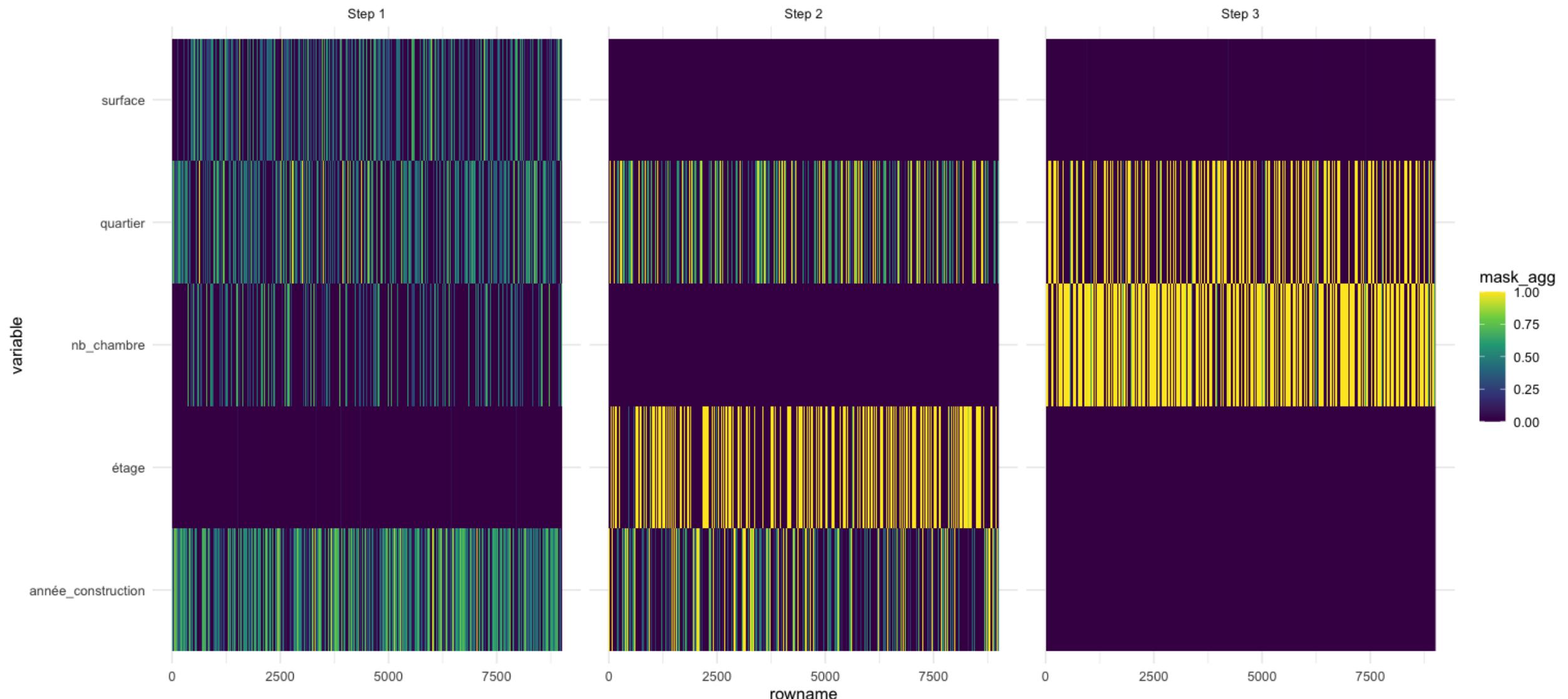


valide la justesse locale du modèle
valide la stabilité du modèle
valide l'additivité locale du modèle.

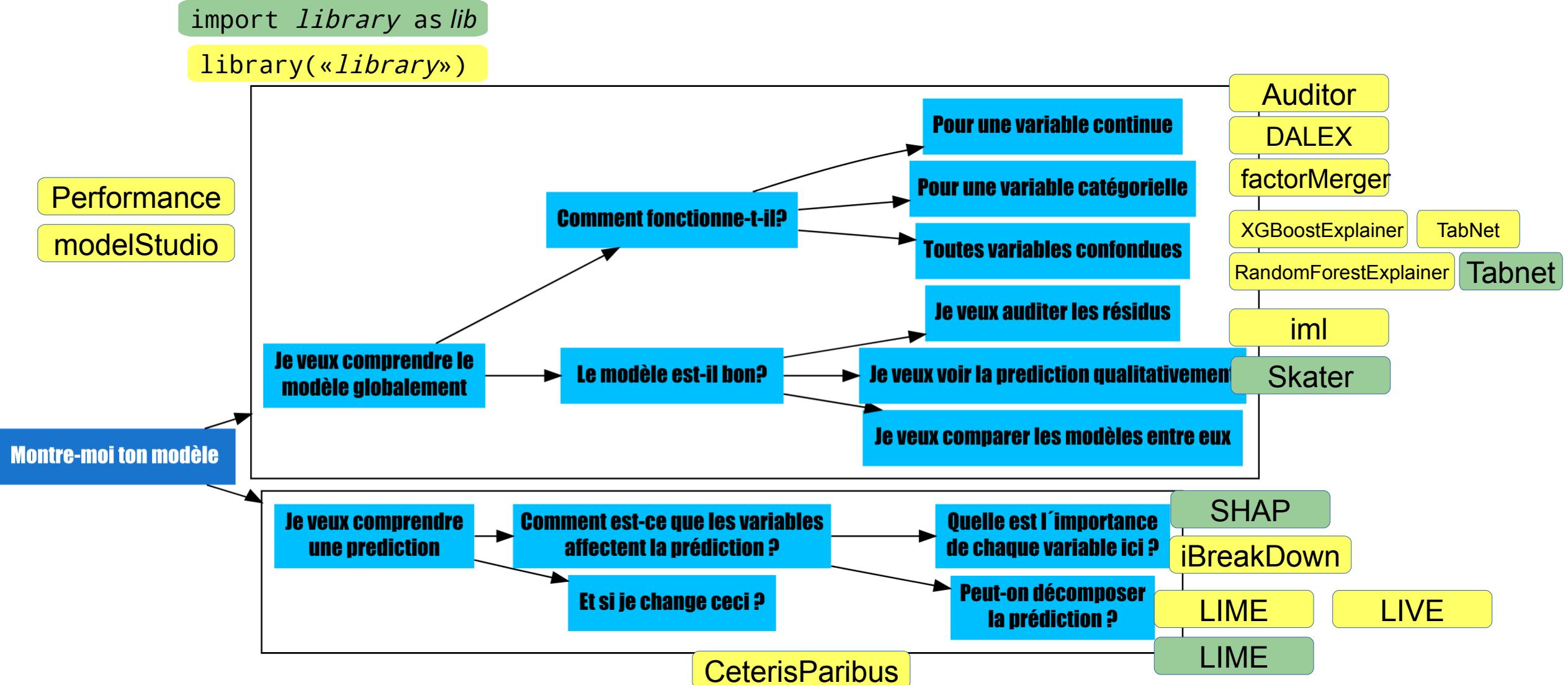
Visualisation spécifique de modèle : TabNet



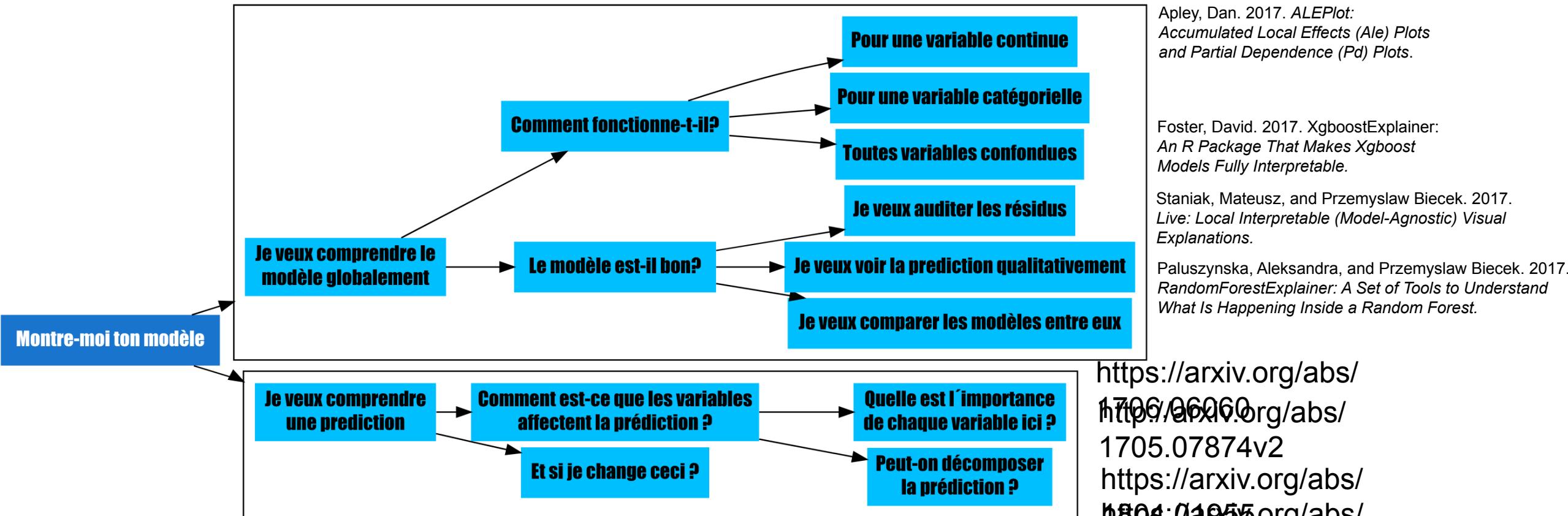
Visualisation spécifique de modèle : TabNet



Et pour les solutions: montre-moi ton package !



Et pour les solutions: montre-moi tes references !



Apley, Dan. 2017. *ALEPlot: Accumulated Local Effects (Ale) Plots and Partial Dependence (Pd) Plots*.

Foster, David. 2017. *XgboostExplainer: An R Package That Makes Xgboost Models Fully Interpretable*.

Staniak, Mateusz, and Przemyslaw Biecek. 2017. *Live: Local Interpretable (Model-Agnostic) Visual Explanations*.

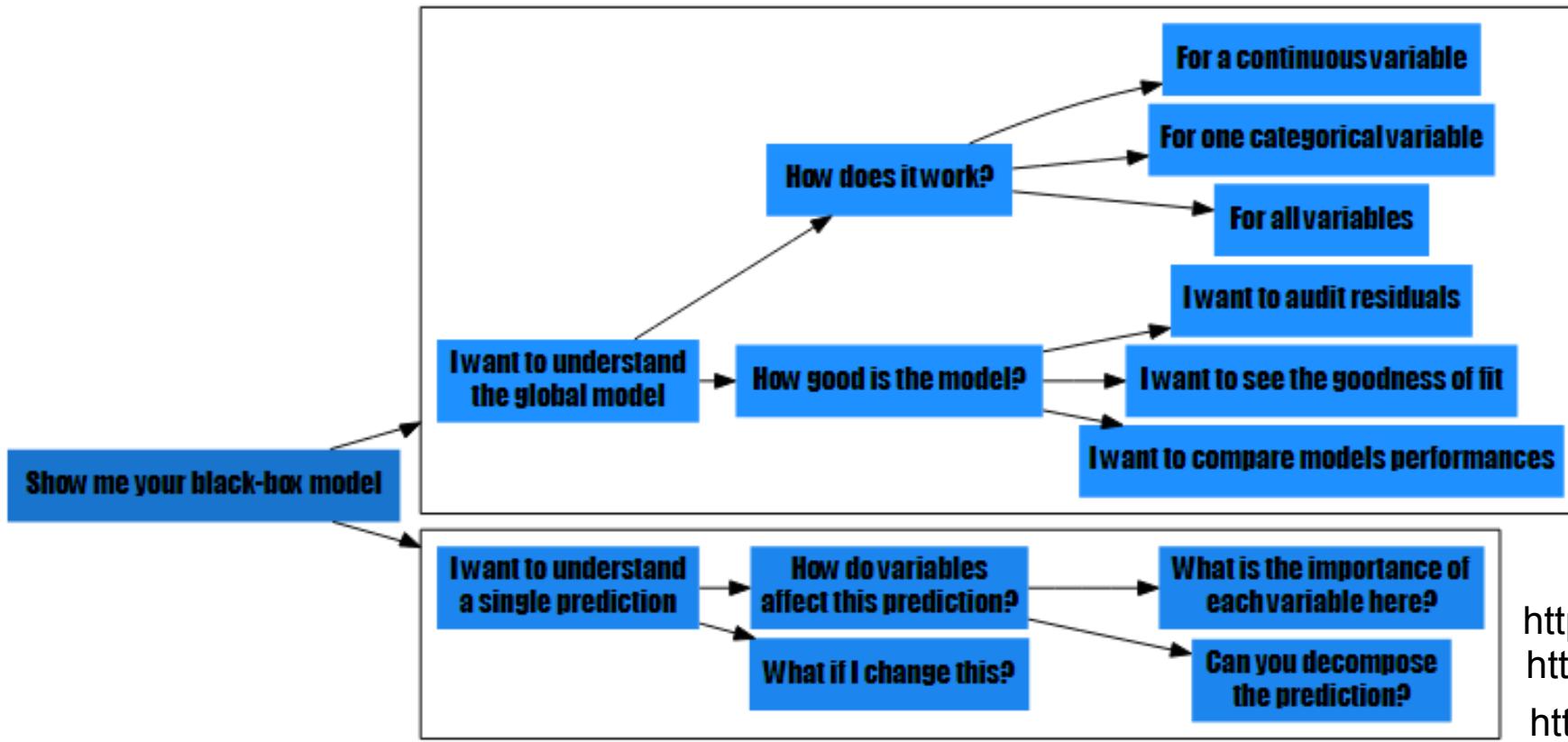
Paluszynska, Aleksandra, and Przemyslaw Biecek. 2017. *RandomForestExplainer: A Set of Tools to Understand What Is Happening Inside a Random Forest*.

<https://arxiv.org/abs/1706.06060>
<http://arxiv.org/abs/1705.07874v2>
<https://arxiv.org/abs/1804.01055>
<https://arxiv.org/abs/1606.05386>

Interpretable Machine Learning book : <https://christophm.github.io/interpretable-ml-book/>

DALEX: Descriptive mAchine Learning Explanations : https://pbiecek.github.io/DALEX_docs/

The solution space: show me the references



Apley, Dan. 2017. *ALEPlot: Accumulated Local Effects (Ale) Plots and Partial Dependence (Pd) Plots*.

Apley, Daniel W., and Jingyu Zhu. 2020. "Visualizing the effects of predictor variables in black box supervised learning models."

Foster, David. 2017. *XgboostExplainer: An R Package That Makes Xgboost Models Fully Interpretable*.

Staniak, Mateusz, and Przemyslaw Biecek. 2017. *Live: Local Interpretable (Model-Agnostic) Visual Explanations*.

Paluszynska, Aleksandra, and Przemyslaw Biecek. 2017. *RandomForestExplainer: A Set of Tools to Understand What Is Happening Inside a Random Forest*.

<https://arxiv.org/abs/1706.06060>

<http://arxiv.org/abs/1705.07874v2>

<https://arxiv.org/abs/1804.01955>

<https://arxiv.org/abs/1606.05386>

Interpretable Machine Learning book : <https://christophm.github.io/interpretable-ml-book/>
DrWhy / DALEX: Descriptive mACHINE Learning Explanations : <http://ema.drwhy.ai/>

Non Tabular Data : Images

Lime Mega-pixel contribution



Post-hoc Model Explanations for VGG-16

**Junco
Bird**

