ÉVALUER UN MODÈLE DE RÉGRESSION

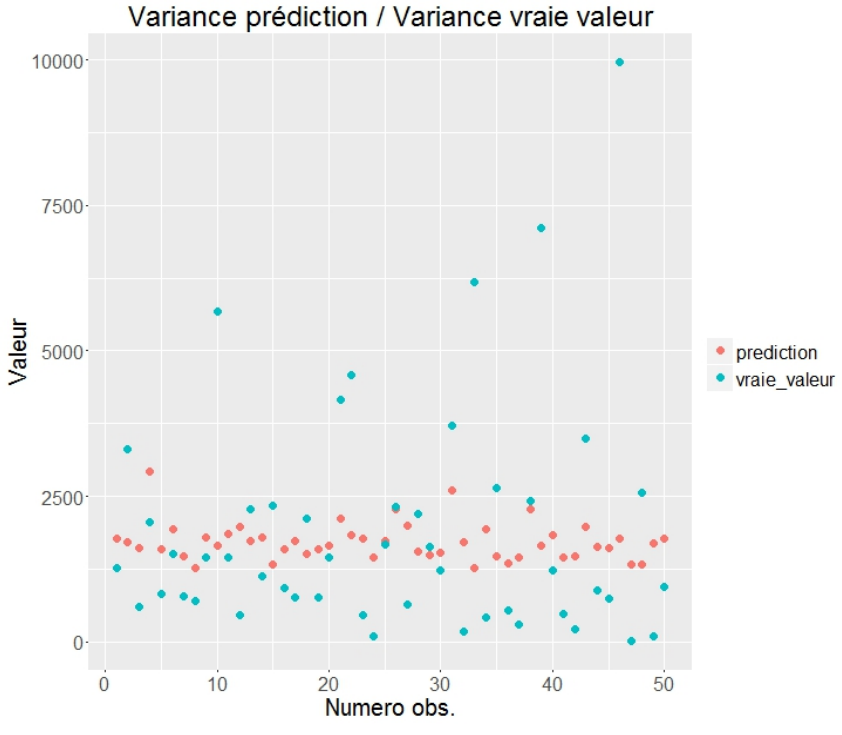
- On utilise souvent des critères quantitatifs pour évaluer un modèle statistique.
  - Critères classiques :
    - $R^2 = 1 - \frac{\sum_{i=1}^{N} \epsilon_i^2}{\sum_{i=1}^{N} (y_i - y)^2} = 1 - \frac{erreur\ quadratique\ du\ modèle}{variance\ empirique\ de\ y} = \frac{variance\ expliquée}{variance\ totale}$, avec $\bar{y} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$
      - S'interprète comme la proportion de variance de Y expliquée par le modèle : on cherche donc à maximiser le R^2
    - RMSE (Root Mean Square Error) : Erreur moyenne quadratique : $\sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$
  - Ces critères permettent de comparer les performances de modèles de régression obtenus avec des méthodes différentes

ÉVALUER UN MODÈLE DE RÉGRESSION

- Remarque sur le R^2 :
  - Selon le problème étudié, le R^2 prend des valeurs très différentes.

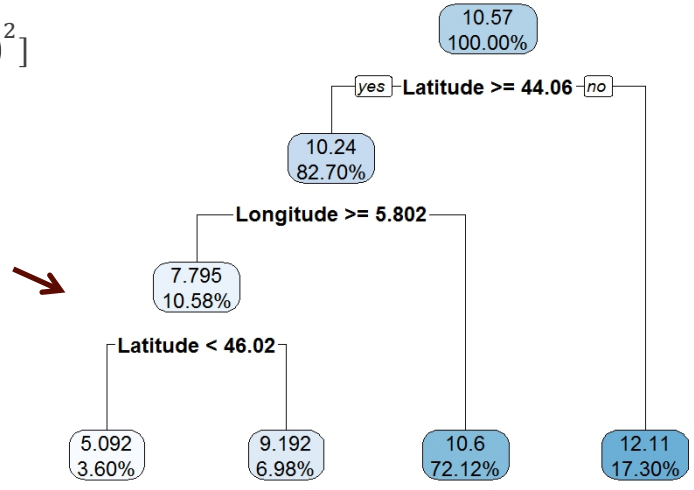
    Exemples :
    - Prédire le coût d'un sinistre en assurance dommage auto à partir des informations sur le véhicule (Marque, type de véhicule, etc…)-> R^2 faible (environ 1,5%)
    - Prédire la température annuelle moyenne d'un lieu en France en fonction de latitude-longitude -> R^2 élevé (environ 90%)

  - Dans l'évaluation d'un modèle, ATTENTION au risque de sur-apprentissage !



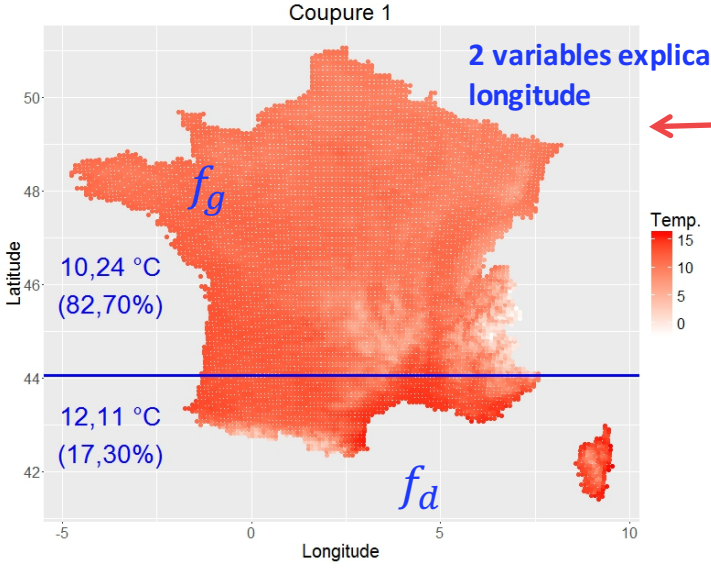
CART (CLASSIFICATION AND REGRESSION TREE) – PRINCIPE DU MODÈLE

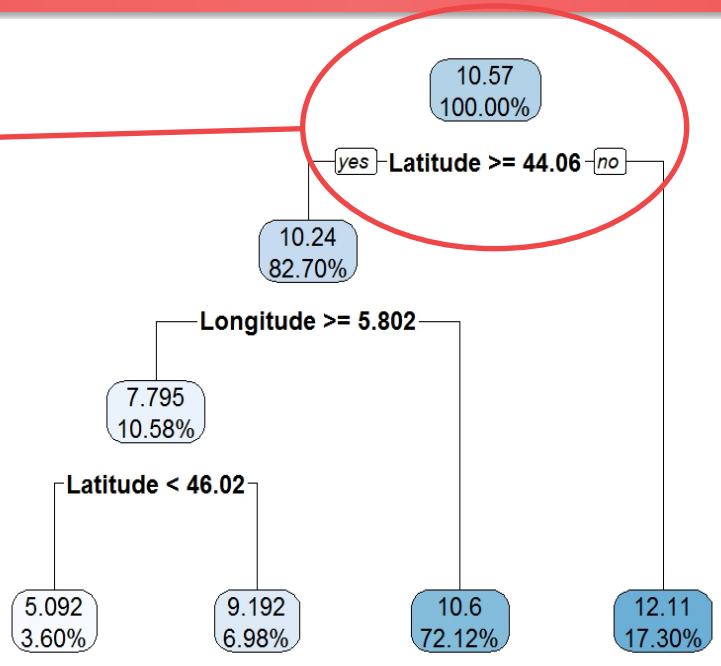
- Principe du modèle CART
  - On cherche la fonction f^{**} telle que $f^{**} = \underset{f}{\operatorname{argmin}}\ E[(Y - f(X))^2]$
  - Comme $E[(Y - f(X))^2]$ est inconnue, en pratique on cherche $f^* = \underset{f}{\operatorname{argmin}} \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2\right]$
  - On construit un estimateur $\tilde{f}$ de f^* qui prend la forme d'un arbre :



- Comparaison GLM
  - Dans le GLM, on suppose $g(E[Y|X]) = {}^t\beta.X$ et le problème est de trouver $\beta^* = \underset{\beta}{\operatorname{argmin}}\, E[(Y - g^{-1}({}^t\beta.X))^2]$
  - Le problème à résoudre dans le GLM est donc un cas particulier du problème à résoudre pour CART : celui où $f(x) = g^{-1}({}^t\beta.x)$
  - Dans le GLM on fait une hypothèse plus forte et l'estimation de β est plus simple (β est de dimension finie). Dans CART on ne fait aucune hypothèse et le problème à résoudre est plus complexe (f est de dimension infinie) : on peut néanmoins chercher une solution approchée par la construction d' arbre.

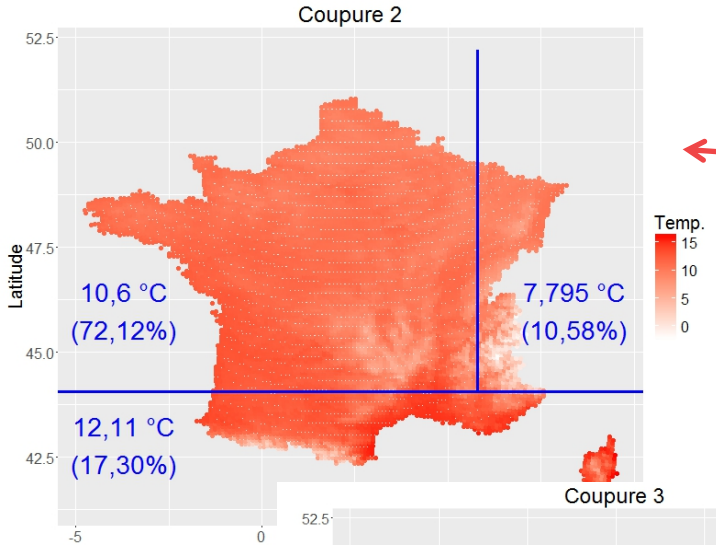
ALGORITHME DE CONSTRUCTION DE L'ARBRE : EXEMPLE DE LA PRÉDICTION DE LA TEMPÉRATURE MOYENNE EN FRANCE

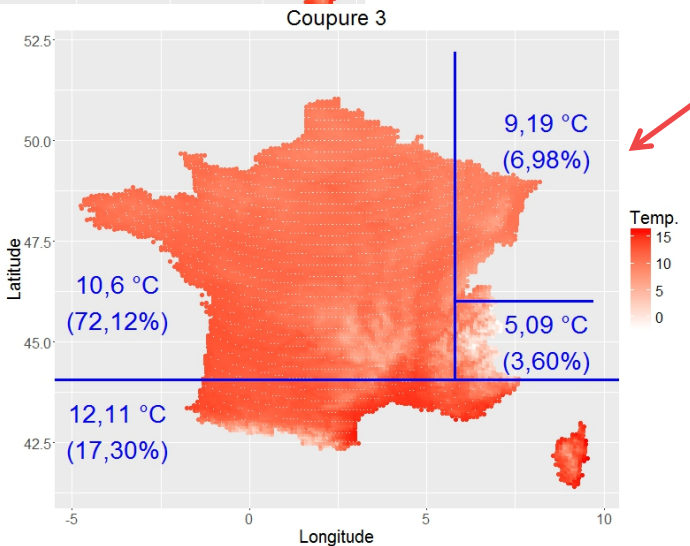


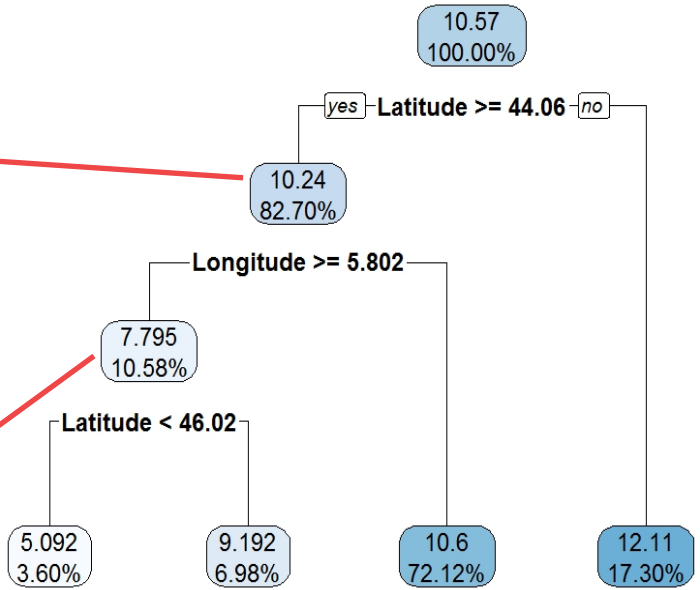


- Variance d'une feuille f :
  - soit $\bar{y}$ la moyenne de y dans f. Alors $Var(f) = \frac{1}{|f|}\sum_{i \in f}(y_i - \bar{y})^2$
- Principe :
  - Procédure récursive. En un nœud :
    - On choisit la coupure (de la forme $X_j \overset{(>)}{<} c$) qui minimise $|f_g| \cdot Var(f_g) + |f_d| \cdot Var(f_d)$: la somme des variances des feuilles filles pondérée par les effectifs des feuilles
    - Remarque : On a toujours $|f_g| \cdot Var(f_g) + |f_d| \cdot Var(f_d) \leq |f| \cdot Var(f)$ de manière que la variance totale de l'arbre décroît pendant l'algorithme

ALGORITHME DE CONSTRUCTION DE L'ARBRE : EXEMPLE DE LA PRÉDICTION DE LA TEMPÉRATURE MOYENNE EN FRANCE







ALGORITHME DE CONSTRUCTION DE L'ARBRE : CRITÈRE D'ARRÊT

- Notations :
  - On note F un arbre et $Var(F)$ sa variance, i.e. la somme des variances de ses feuilles pondérée par les effectifs des feuilles
  - Soit F_0 l'arbre initiale (tous les individus sont dans la feuille initiale) et soient $F_1, F_2, \dots, F_{10}$, …les arbres successifs construits pendant la procédure
  - Remarques :
    - $Var(F_0) = Var(Y)$
    - Le R^2 (train) associé à un arbre F est : ${R_F}^2 = 1 - {Var(F)}/{Var(Y)}$

- Critères d'arrêt :
  - L'algorithme s'arrête lorsque $Var(F_i) - Var(F_{i+1}) < c.Var(Y)$ (soit ${R_{F_{i+1}}}^2 - {R_{F_i}}^2 < c$) , avec c le paramètre de complexité de l'arbre (noté « cp » dans R)
  - C'est-à-dire : lorsque l'augmentation du R^2 dû à l'ajout d'une feuille est trop faible, on s'arrête.
  - Le critère d'arrêt précédent est le critère d'arrêt « principal ». En pratique les critères suivants se combinent avec le critère principal dans le choix de l' arrêt (fonction « rpart.control » dans R):
    - Nombre minimal d'individus d'un nœud pour qu'une coupure soit cherchée.
    - Nombre minimal d'individus d'une feuille
    - Profondeur maximale : longueur maximale d'un chemin qui descend depuis le haut de l'arbre vers les feuilles.

ALGORITHME DE CONSTRUCTION DE L'ARBRE : FORMES DES COUPURES

- Remarques :
  - Formes des coupures en fonction du type de la variable explicative
    - Quantitative : $X_j \overset{(>)}{\leq} c$
    - Qualitative : $X_j \in$ {ensemble de modalités}
  - Pour utiliser l'algorithme CART sur une variable Y qualitative, on remplace la variance par l'impureté de Gini* , une fonction qui comme la variance mesure la « pureté » d'une classe. Pour le reste, l'algorithme est identique à celui présenté ici.

* : à ne pas confondre avec le coefficient de Gini

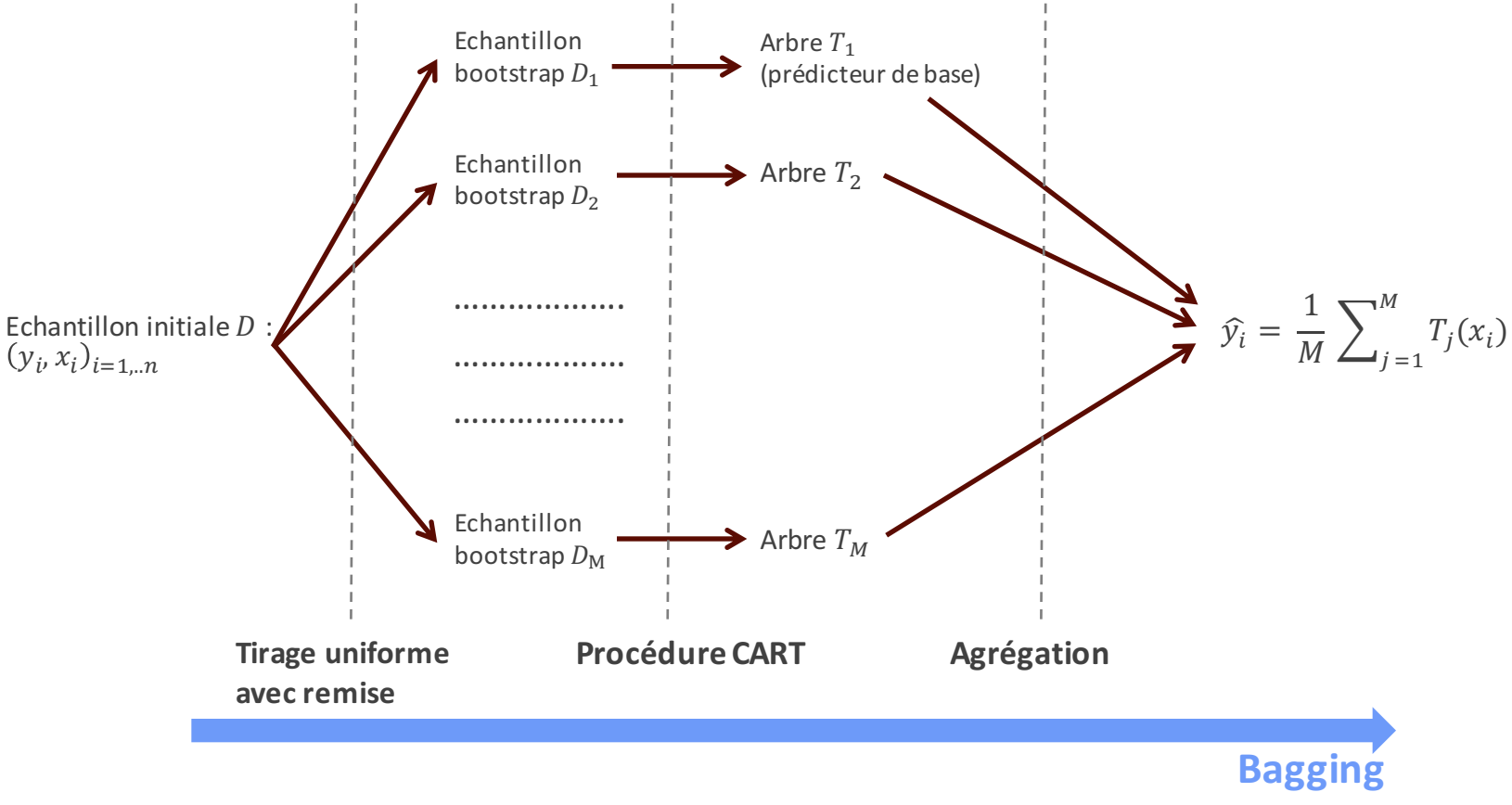
VALIDATION CROISÉ

- Principe de Validation croisé :
  - La validation croisée permet de mesurer la performance d'un modèle sans qu'il y ait de risque de sur-apprentissage

  - Principe :
    - On coupe la base de données B en K groupes de taille $^{N}/_{K}$: $\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_\mathrm{K}$
    - Pour i allant de 1 à K, on calibre un CART sur $B \setminus G_i$, puis on utilise le modèle calibré pour prédire Y pour les individus de G_i
    - Après la boucle, on dispose d'une prédiction pour tous les individus de B. On peut alors calculer le R^2 de cette prédiction (ou d'autres statistiques)

  - Remarque : On peut répéter la procédure de validation croisée plusieurs fois, et moyenner les résultats pour obtenir une meilleure stabilité des résultats

PRINCIPE DE LA FORÊT ALÉATOIRE



RF (RANDOM FOREST)

- Principe de la forêt aléatoire :
  - Faire M arbres CART (une forêt !) puis moyenner les prédictions des M CART pour obtenir la prédiction finale.
  - Pour que les M CART ne soit pas tous identiques et « explorent » diverses combinaisons de X, on introduit de l'aléa dans la construction des M CART :
    - Chaque arbre CART est construit sur un échantillon d'observations **bootstrap** (tirage uniforme avec remise). Combiné avec le principe d'agrégation des différents CART, cela s'appelle le « **bagging** » (bootstrap aggregating)
    - Dans la construction d'un arbre, lors de la recherche de la meilleure coupure on ne teste pas l'ensemble des variables. En chaque nœud on tire un sous-ensemble de variables (paramètre *mtry)* parmi lesquelles on recherche la coupure optimale. En plus de l'exploration, on gagne ici du temps de calcul.
  - Les arbres CART ne sont **pas élagués** (gain de temps) et le critère d'arrêt est en général basé sur le nombre d'individus des feuilles (gain de temps là aussi)
  - On peut donc penser que chaque arbre est beaucoup sur-appris, toutefois la moyennisation permet de réduire le sur-apprentissage

LES PLUS / LES MOINS

- Avantage
  - Prédiction plus performante que celle d'un seul CART (en termes de R^2, Gini,…)
  - Pas besoin de recourir à une stratégie train/test pour éviter le sur-apprentissage grâce à la prédiction **out of bag (OOB):**
    - Idée : Pour prédire le Y d'un individu du train, on utilise seulement les arbres où cet individu n'apparaît pas dans l'échantillon du bagging (il est alors « out of bag »). En pratique, cette méthode donne une très bonne approximation de la performance de validation.
  - Donc pas besoin d'avoir recours au bootstrap/validation croisée pour évaluer performance du modèle

- Inconvénient :
  - Par rapport au CART, on perd la représentation du modèle sous forme d'arbre binaire
  - Contrairement à GLM et CART, il n'y a pas de formule simple qui donne la prédiction du Y pour des caractéristiques X données.

PRINCIPAUX PARAMÈTRES DE LA FORÊT ALÉATOIRE

- Comme pour CART, on peut avec RF jouer sur différents paramètres pour construire le modèle :
  - **nodesize :** Nombre minimal d'individus d'une feuille (5 par défaut)
  - **nodedepth** : Longueur maximale d'un chemin qui descend depuis le haut de l'arbre vers les feuilles (profondeur maximale de d'arbre)
  - **mtry** : Nombre de variables tirées en chaque nœud parmi lesquelles on cherche la coupure optimale
  - **ntree** : nombre d'arbres de la forêt (1000 par défaut)

ALGORITHME POUR CALIBRER UNE FORÊT ALÉATOIRE

1) Lancer une première forêt aléatoire pour avoir une première évaluation de la performance prédictive du modèle.
    1) Ici, si on dispose d'un grosse base (par exemple 1 000 000 individus, 50 variables), penser à lancer la forêt sur un sous-échantillon de 20 000 – 30 000 individus → gros gain de temps et faible perte en terme de qualité de modèle

2) Optimiser un ou plusieurs paramètres de manière à maximiser la performance du modèle
    1) Dans l'exemple on a optimiser le paramètre « nodedepth »
    2) En général, lorsque la performance d'un modèle est faible (faible R^2 ou autre critère d'évaluation), diminuer la taille des arbre construits améliore la performance du modèle. (souvent le cas en assurance auto, assurance RC, assurance habitation, etc …)

REMARQUES SUR LA FORÊT ALÉATOIRE

- Comme CART, la forêt aléatoire traite à la fois les variables X quantitatives et les variables X qualitatives.
- L'idée de moyenner différents arbres permet de résoudre le problème de l'instabilité des modèles CART, et aussi de gagner en performance
- On perd cependant en interprétabilité puisque pour comprendre un modèle, il faudrait analyser M arbres différents.
  - On peut néanmoins exploiter la prédiction donnée par la forêt aléatoire pour représenter graphiquement l'impact de chaque variable
- Exporter la forêt aléatoire
  - Pour prédire le Y d'un nouvel individu, il faut moyenner la prédiction des M arbres pour cette individu. Pour utiliser la forêt aléatoire dans un système (par exemple sur Excel, SAS, etc..), il faut donc exporter les M arbres construits dans le système cible.

GRADIENT BOOSTING

SYNTHÈSE : PLUS ET MOINS DES MÉTHODES

	GLM	CART	Forêt Aléatoire & Gradient Boosting Machine
Précision du modèle	- : Variable d'un problème à un autre.	+ : adapté à tout problème mais instable	++
Stabilité (sensibilité du résultat aux données)	+ : dépend fortement de la taille de l'échantillon d'apprentissage	- : dépend fortement de la taille de l'échantillon d'apprentissage	++ : Ok, on peut facilement augmenter la stabilité en construisant des arbres moins profonds
Interprétabilité	++	++	+ : Modèle plus complexe mais on a des sorties graphiques interprétables
Facilité à calibrer	- : Nécessité de sélectionner les variables, utiliser une méthode train-test	- : Nécessité de sélectionner les variables, utiliser une méthode train-test + élagage de l'arbre	++ : La performances est bonne même sans sélection de variable. On peut facilement optimiser certains paramètres
Exportation du modèle sur d'autres système	++ : simple	++ : simple	- : Besoin de faire du code dans le langage cible

	X
Obs 1	B
Obs 2	B
Obs 3	A
Obs 4	C
…	

	X.A	X.B	X.C
Obs 1	0	1	0
Obs 2	0	1	0
Obs 3	1	0	0
Obs 4	0	0	1
…			

Marque	Alimentation	PTAC	Vitesse maxi	Contrôle dyn. Stab.	Composite	prédiction coût moyen
Renault	Essence	(1690,1780]	(150,203]	Non	Berline Familiale	1561
Volkswagen	Diésel	(1610,1690]	(150,203]	Oui	Berline Familiale	1962
Renault	Essence	(1850,1940]	(203,329]	Oui	Cabriolet/Coupe	2507