MACHINE LEARNING

Tarification des contrats d'assurance automobile

Guy Tsang, Axel Gardahaut & Léo Dutertre-Ladurée

INTRODUCTION
CONTEXTUALISATION ET PRÉSENTATION DES DONNÉES

• NATURE DES DONNÉES

Les variables sont anonymisées :
- Variables concernant l'assuré
- Variables concernant la région de l'assuré
- Variables concernant la voiture de l'assuré
- Variables calculées

• BASE D'APPRENTISSAGE

Nombre de variables : 57
Nombre d'individus : 416 648

• VARIABLE CIBLE

Réclamation dans un délai d'un an :
- Proportion de 1 dans la base d'apprentissage : 3.67%
- Proportion de 1 dans la base de test : 3.60%

• MÉTRIQUE CHOISIE

Coefficient normalisé de Giini :

$$Gini = 2 \times AUC - 1$$

PLAN

I/. DÉMARCHE

A) Benchmark
B) Pré-traitement des données
C) Sélection des variables
D) Traitement des données

II/. MODÈLES

A) Régression logistique pénalisée
B) XGboost
C) LightGBM
D) Stacking

III/. RÉSULTATS

A) Comparaison des modèles
B) Sélection du meilleur modèle & Performance

I/. DÉMARCHE
A) BENCHMARK

- Seuil à absolument dépasser avec :
  - Le traitement de la base de données
  - La sélection de variables
  - L’utilisation de modèles alternatifs
  - L’hyperparamétrisation des modèles

- Benchmark par LightGBM
  - Sans hyperparamétrisation (définis selon des valeurs usuelles en pratique)
  - Première étude de l’importance des variables du jeu de données
  - Coefficient Normalisé de Gini en validation croisée 5 blocs : 0.2719

I/. DÉMARCHE
B) PRÉ-TRAITEMENT DES DONNÉES

Bases de données initiales
0.2719

1 Détection des valeurs manquantes
- Identification des variables présentant des valeurs manquantes.
- Décision sur le traitement adapté selon la part représentée par les valeurs manquantes.

2 Pertinence des transformations
- Construire une variable dichotomique indiquant si la valeur est manquante ?
- Considérer les valeurs manquantes comme une modalité de la variable ?
- Remplacer par une médiane ?
- Imputation sophistiquée ?

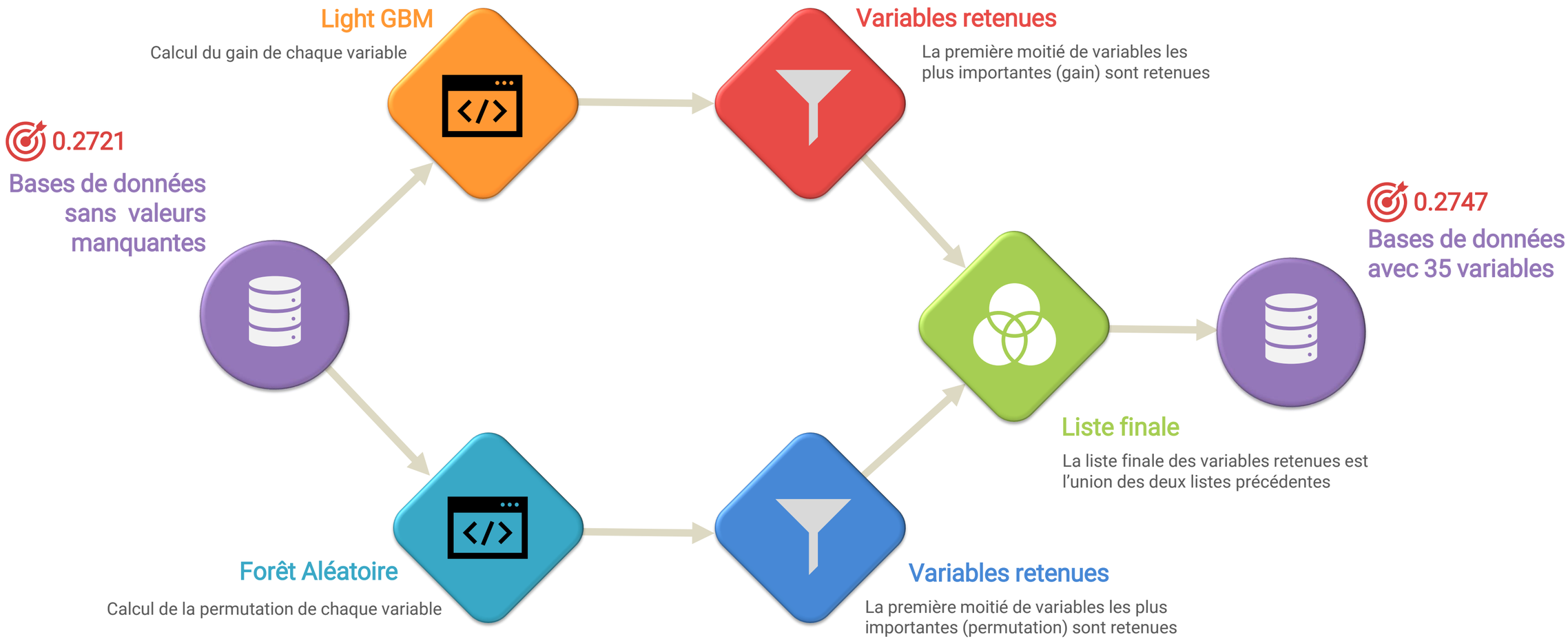
3 Variables à fort ratio de NA
- 2 variables qualitatives
- Remplacement des NA par une nouvelle modalité spéciale (-999)

4 Imputation
- Variables à faible taux de NA
- Imputation par forêt aléatoire
- Apprentissage sur la base train

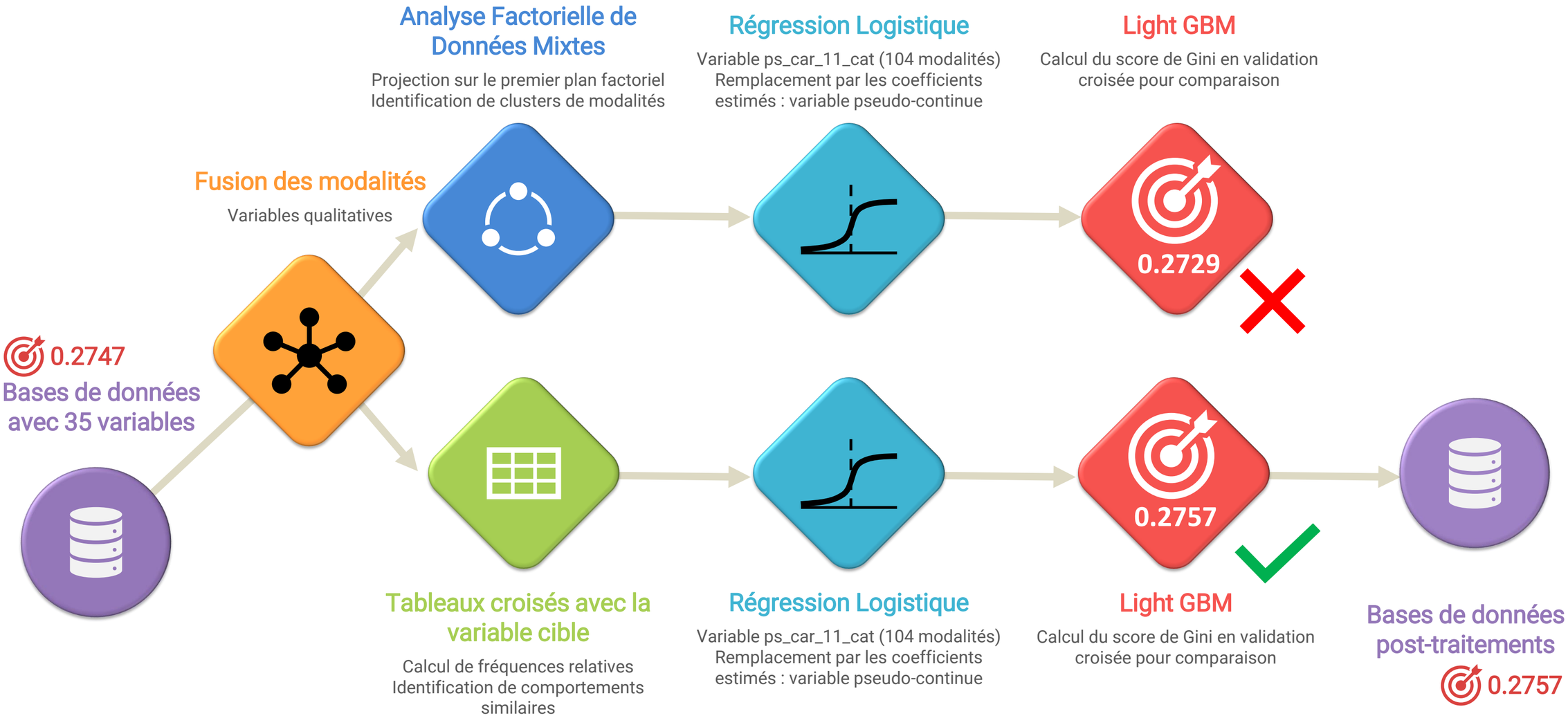
5 Comparaison des distributions avant/après
- Changements sur les fréquences relatives des modalités des variables qualitatives
- Changements sur la distribution des variables quantitatives

Bases de données sans valeurs manquantes
0.2721

I/. DÉMARCHE
C) SÉLECTION DES VARIABLES



I/. DÉMARCHE
D) TRAITEMENT DES DONNÉES



II/. MODÈLES
A) RÉGRESSION LOGISTIQUE PÉNALISÉE

- Modélisation :

$$\min_{\beta} -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln p_i + (1-y_i)\ln(1-p_i)\right] + R(\beta)$$
$$\text{sous contraintes : } R(\beta) \leq \tau$$

- Avantages

  - Restriction de l'erreur quadratique moyenne par rapport à régression logistique classique
  - Interprétabilité du modèle et littérature abondante pour les tests notamment
  - Possibilité d'intégrer des interactions entre variables
  - Mécanisme de sélection de variable possible

- Inconvénients

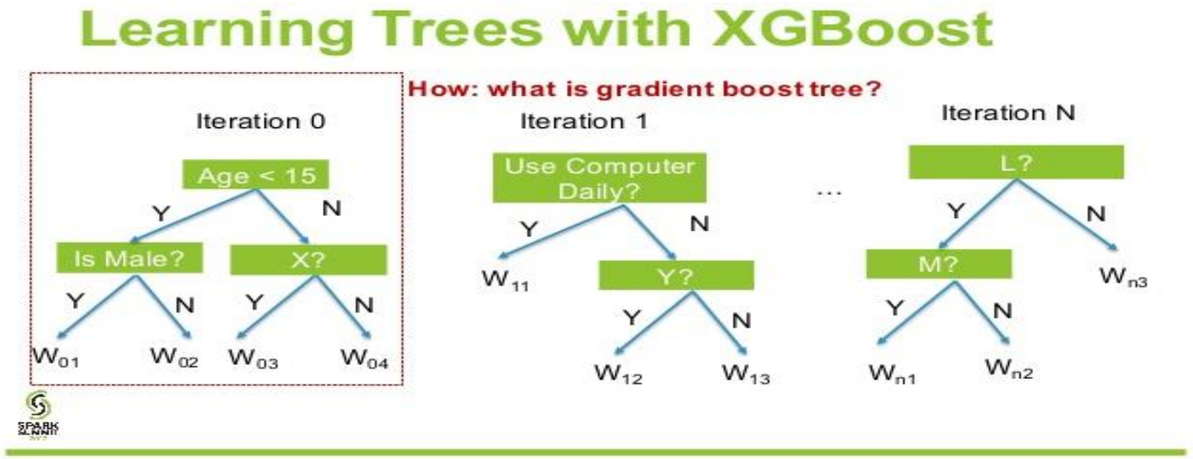
  - Difficulté à gérer automatiquement les phénomènes non-linéaires
  - Problèmes en grande dimension
  - Moins performant dans ce cas de figure

II/. MODÈLES

B) XGBOOST (2016) : BOOSTING + GRADIENT DESCENT

- Modélisation :

$$\mathcal{L} = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right)$$



Source : Yatai Horizon Consulting

- Avantages
  - Très efficace
  - Gère automatiquement les phénomènes non-linéaires
  - Très customisable car un grand nombre de paramètres
  - Mécanisme de sélection de variable possible
- Inconvénients
  - Interprétabilité du modèle
  - Risques d'overfitting
  - Scalable mais problèmes en très grande dimension
  - Justification théorique des performances complexe (basée sur des heuristiques)

II/. MODÈLES

C) LIGHT GBM (2017) : FAST TREE GRADIENT BOOSTING

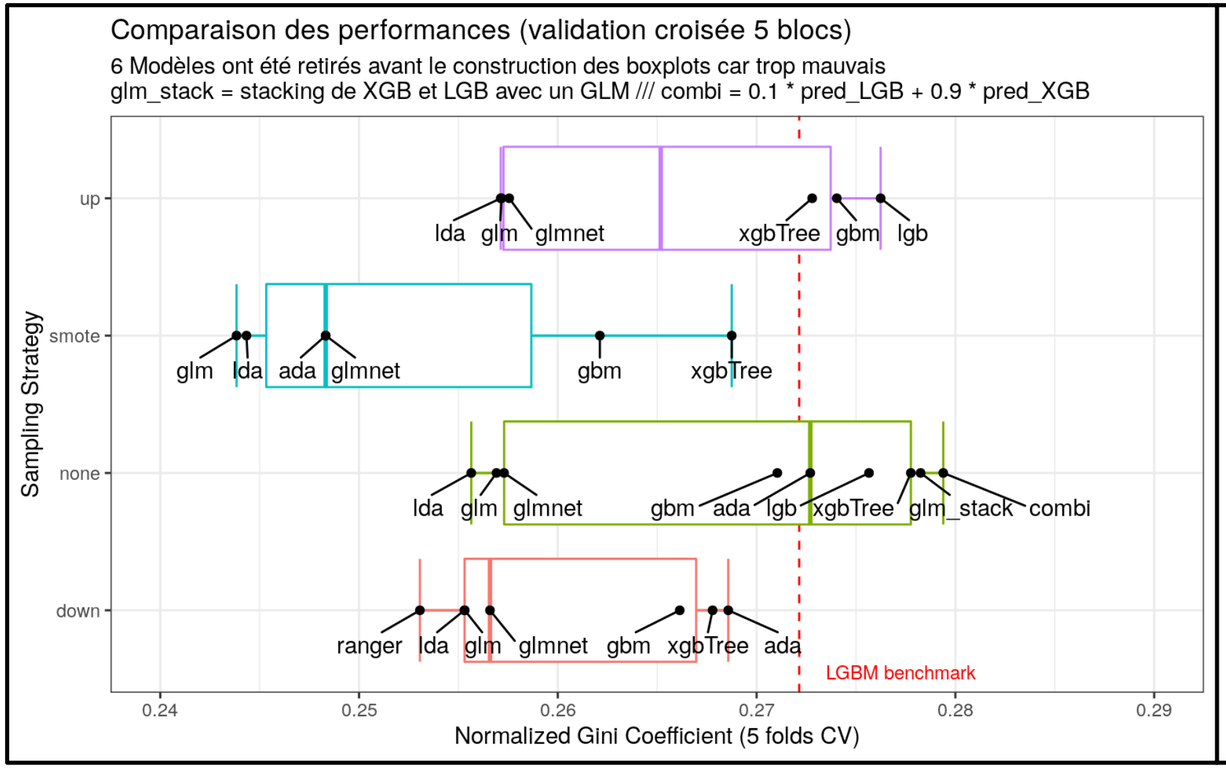
- Même principe que XGboost
- Modélisation : 2 approches différentes
  - GOSS: Sélection des individus les plus informatifs
  - EFB : Réduction des features par regroupement
- Avantages
  - Avantages du XGboost
  - Très efficace et rapide (50x plus rapide que XGboost)
  - Mécanisme de sélection de variable et échantillonnage intégré
  - Justification théorique de certains résultats
- Inconvénients
  - Interprétabilité du modèle encore moins évidente
  - Risques d'overfitting accru

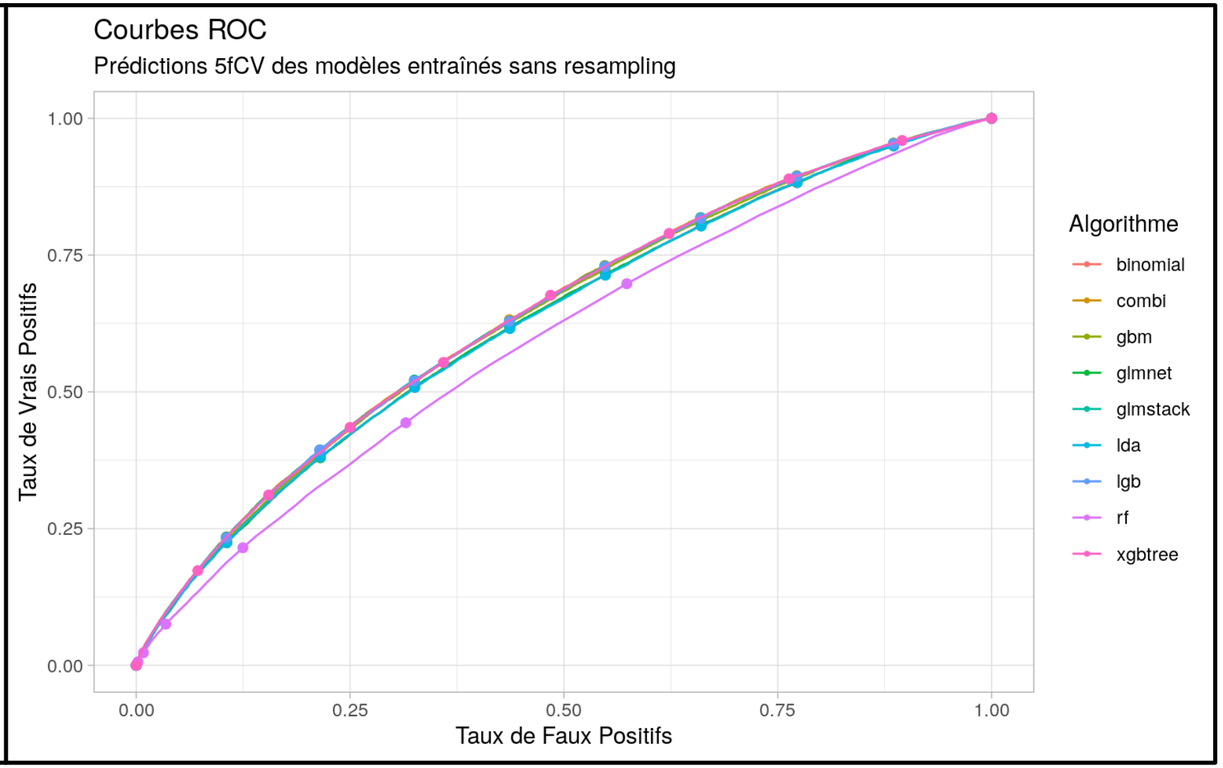
II/. MODÈLES
D) STACKING ET AUTRES MODÈLES

- Principe
  - Agrégation de modèles différents construits à partir d'un méta-classifieur
  - Les modèles n'ont pas besoin d'être de même nature (LDA + Arbre par exemple)
  - Possibilité d'avoir plusieurs couches de stacking
- Avantages
  - Permet d'améliorer les prédictions en cas d'informations complémentaires
  - Rapidement implémentable à partir des prédictions faites des modèles à agréger
- Inconvénients
  - Interprétabilité du modèle encore moins évidente
  - Risques d'overfitting accru

III/. RÉSULTATS

A) COMPARAISON DES MODÈLES



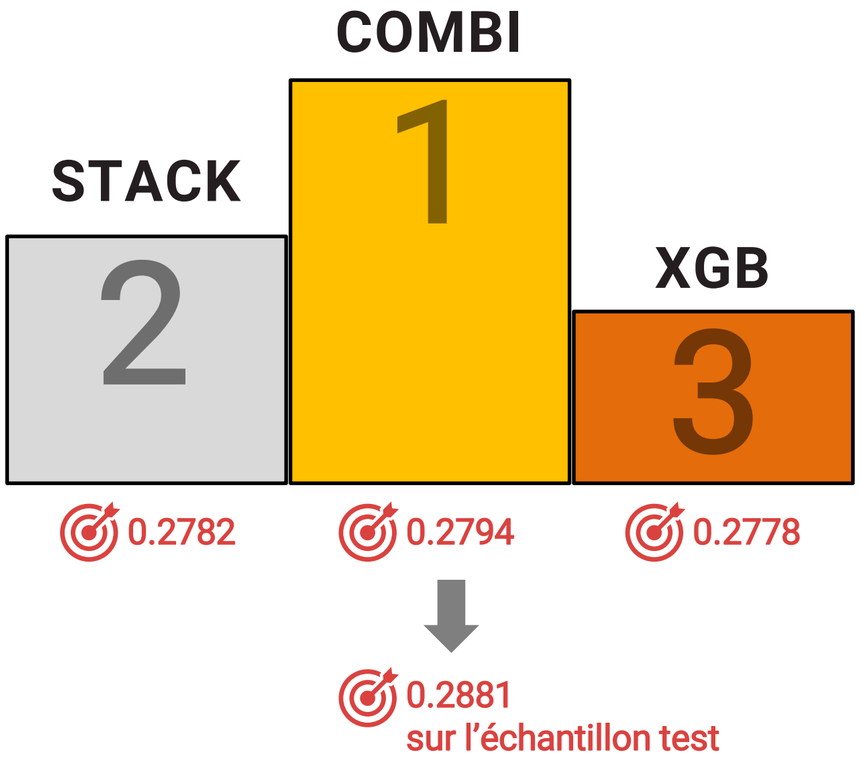


	GLM	LDA	GLMNET	SVM	KPPV	ARBRE	FORÊT	GBM	ADA	LGB	XGB	STACK	COMBI
Up													
Down													
SMOTE													
Aucun													

Modèle construit Modèle non construit Modèle non pertinent

III/. RÉSULTATS

B) SÉLECTION DU MEILLEUR MODÈLE & PERFORMANCE



- Modèle combinaison de scores
  - Modèle issu d'une combinaison linéaire des prédictions CV du LightGBM (10%) et ceux du XGB (90%)
  - Pour information : AUC sur la base test égale à 0.6440
- Classification à partir du modèle retenu
  - Cut obtenu en optimisant le score F1 sur la base train
  - Matrice de confusion et performances de classification

		Observé	
		0	1
Prédiction	0	152 542	4 827
	1	19 596	1 599

	F1	Rappel
Validation	0.1176	0.2482
Test	0.1158	0.2488