



# Tarification des contrats d'assurance automobile

Guy Tsang, Axel Gardahaut & Léo Dutertre-Ladurée

# INTRODUCTION

## CONTEXTUALISATION ET PRÉSENTATION DES DONNÉES

1

### • 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 Gini :

$$\text{Gini} = 2 \times \text{AUC} - 1$$

# 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

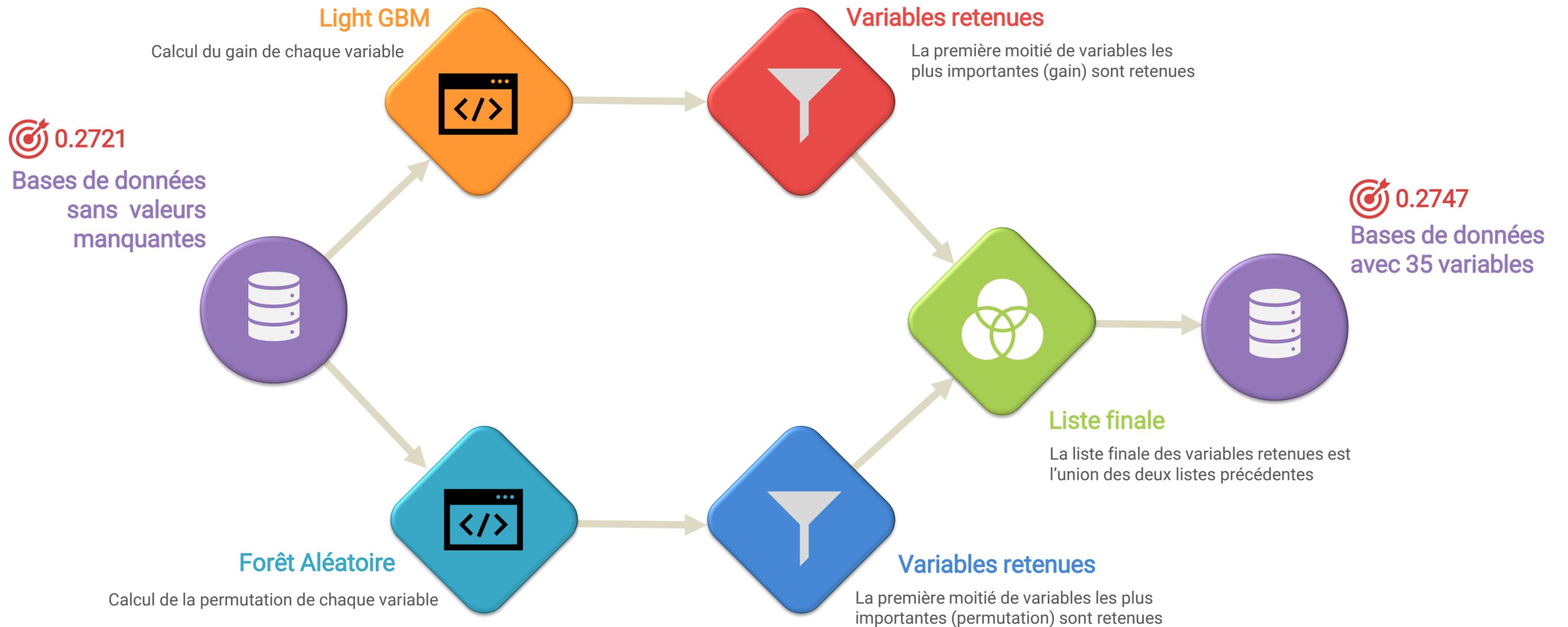
# I/. DÉMARCHE

## B) PRÉ-TRAITEMENT DES DONNÉES



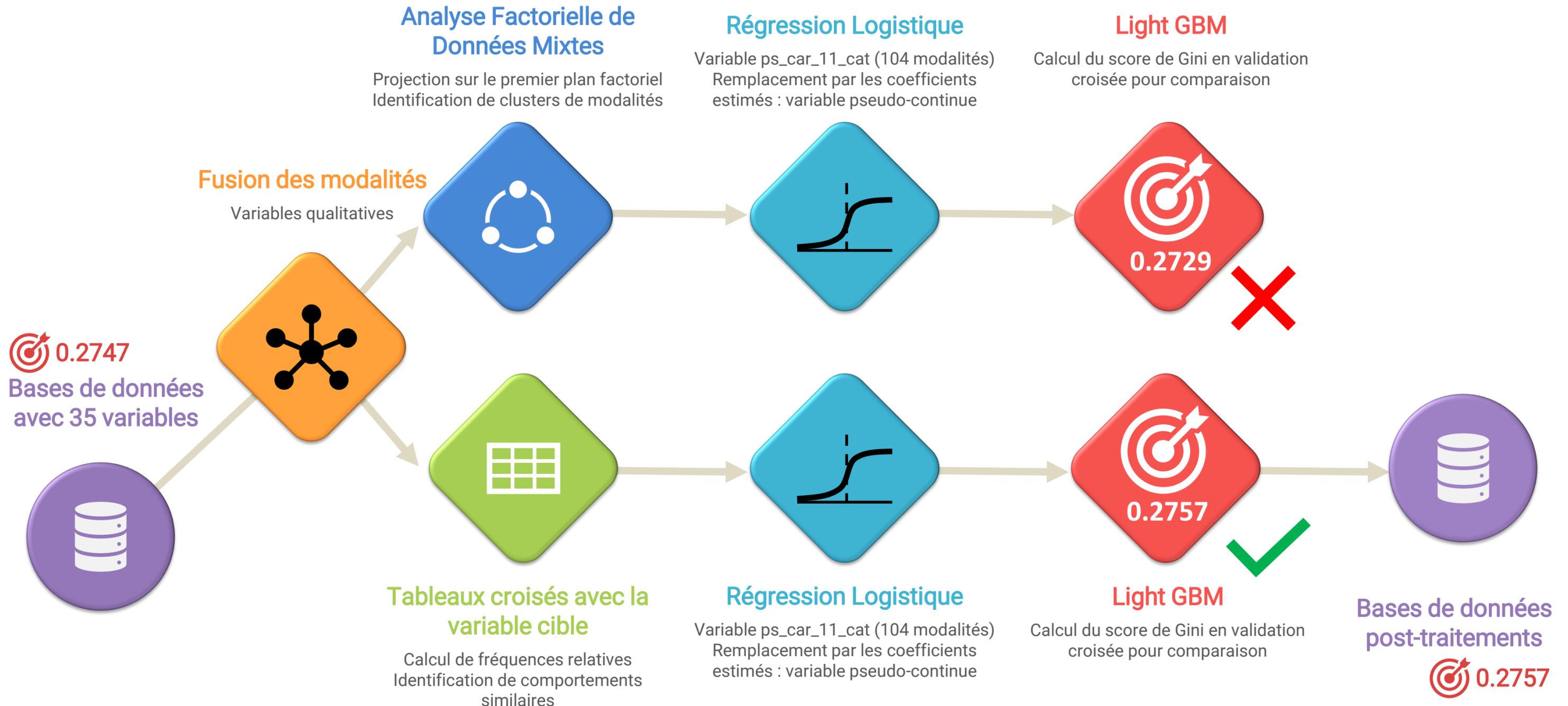
# I/. DÉMARCHE

## C) SÉLECTION DES VARIABLES



# I/. DÉMARCHE

## D) TRAITEMENT DES DONNÉES



# II/. MODÈLES

## A) RÉGRESSION LOGISTIQUE PÉNALISÉE

- Modélisation :

$$\min_{\beta} -\frac{1}{n} \sum_{i=1}^n [y_i \ln p_i + (1 - y_i) \ln (1 - p_i)] + R(\beta)$$

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 :

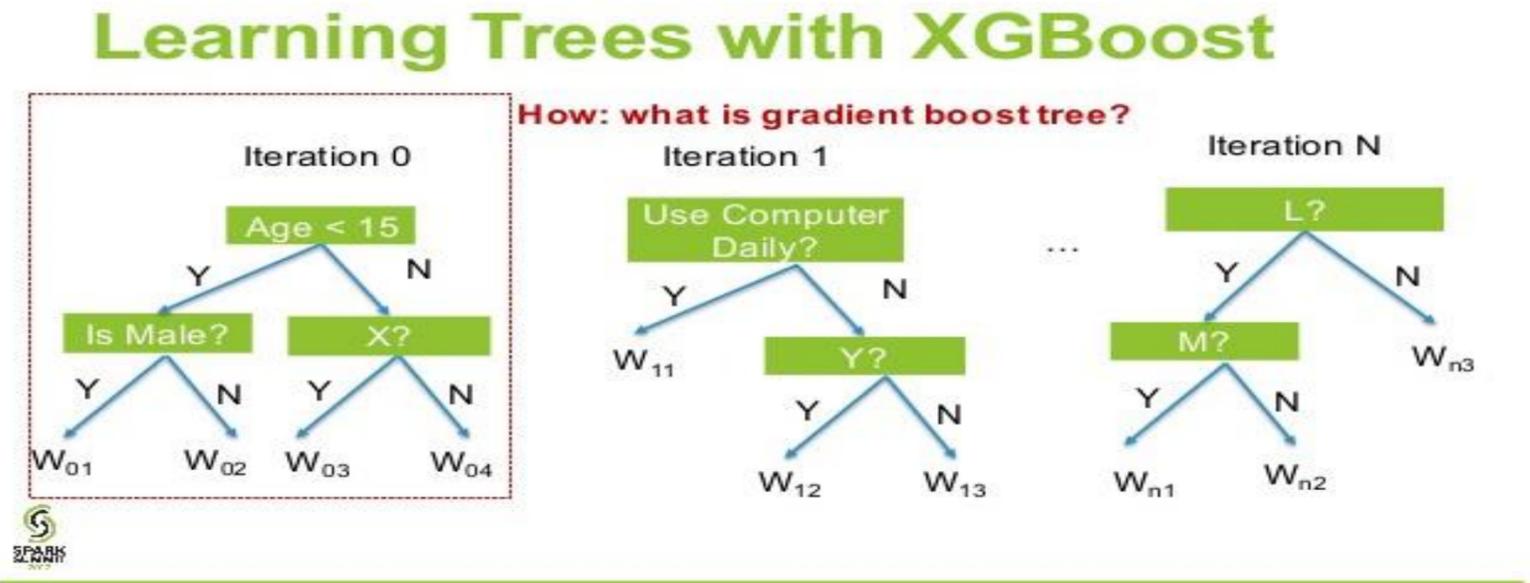
$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

• 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)



Source : Yatai Horizon Consulting

# 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

10

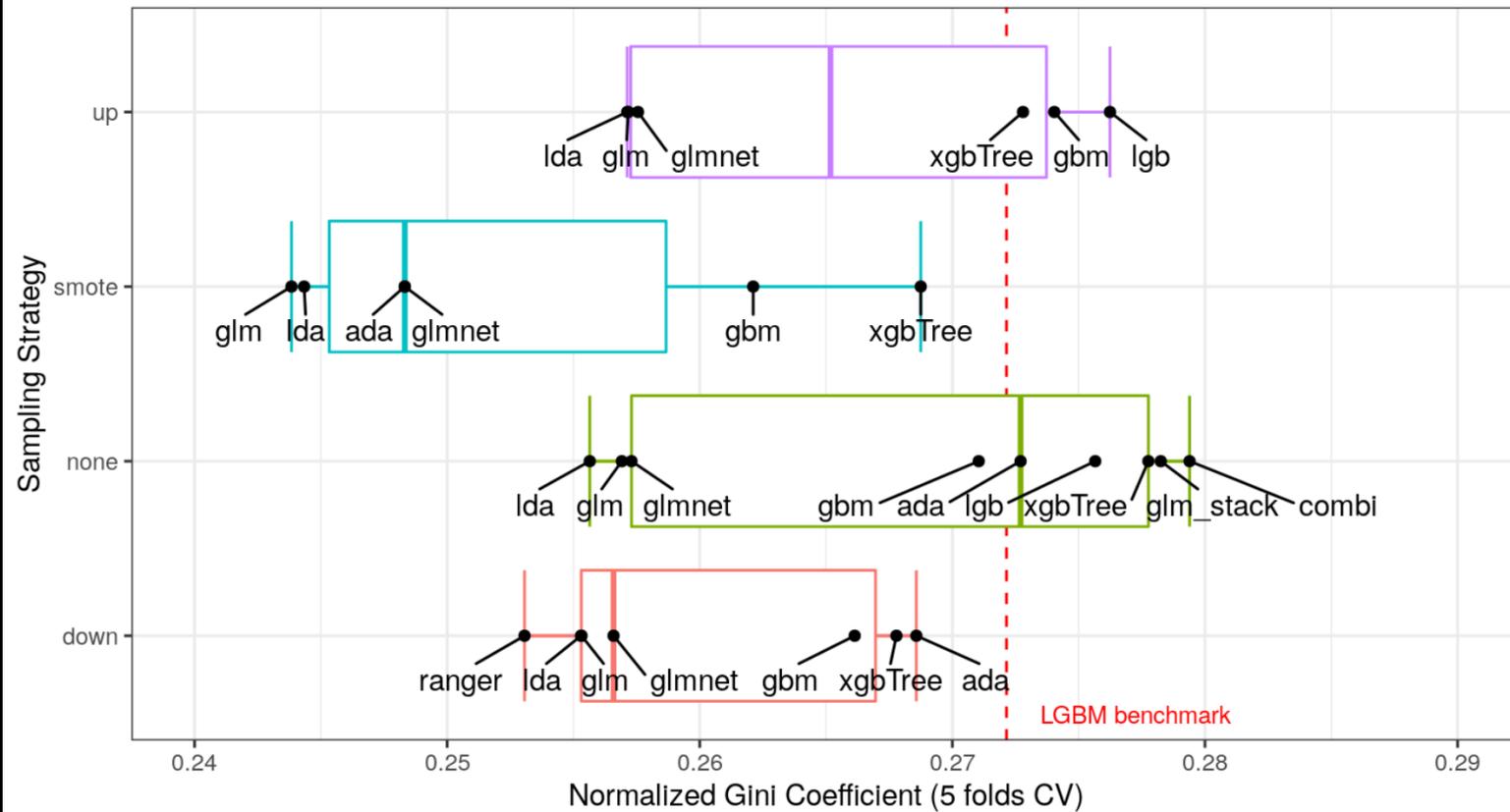
- Principe
  - Agrégation de modèles différents construits à partir d'un méta-classifieur
  - Les modèles n'ont pas besoin d'être de même nature (LDA + Arbre par exemple)
  - Possibilité d'avoir plusieurs couches de stacking
- Avantages
  - ✓ Permet d'améliorer les prédictions en cas d'informations complémentaires
  - ✓ 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

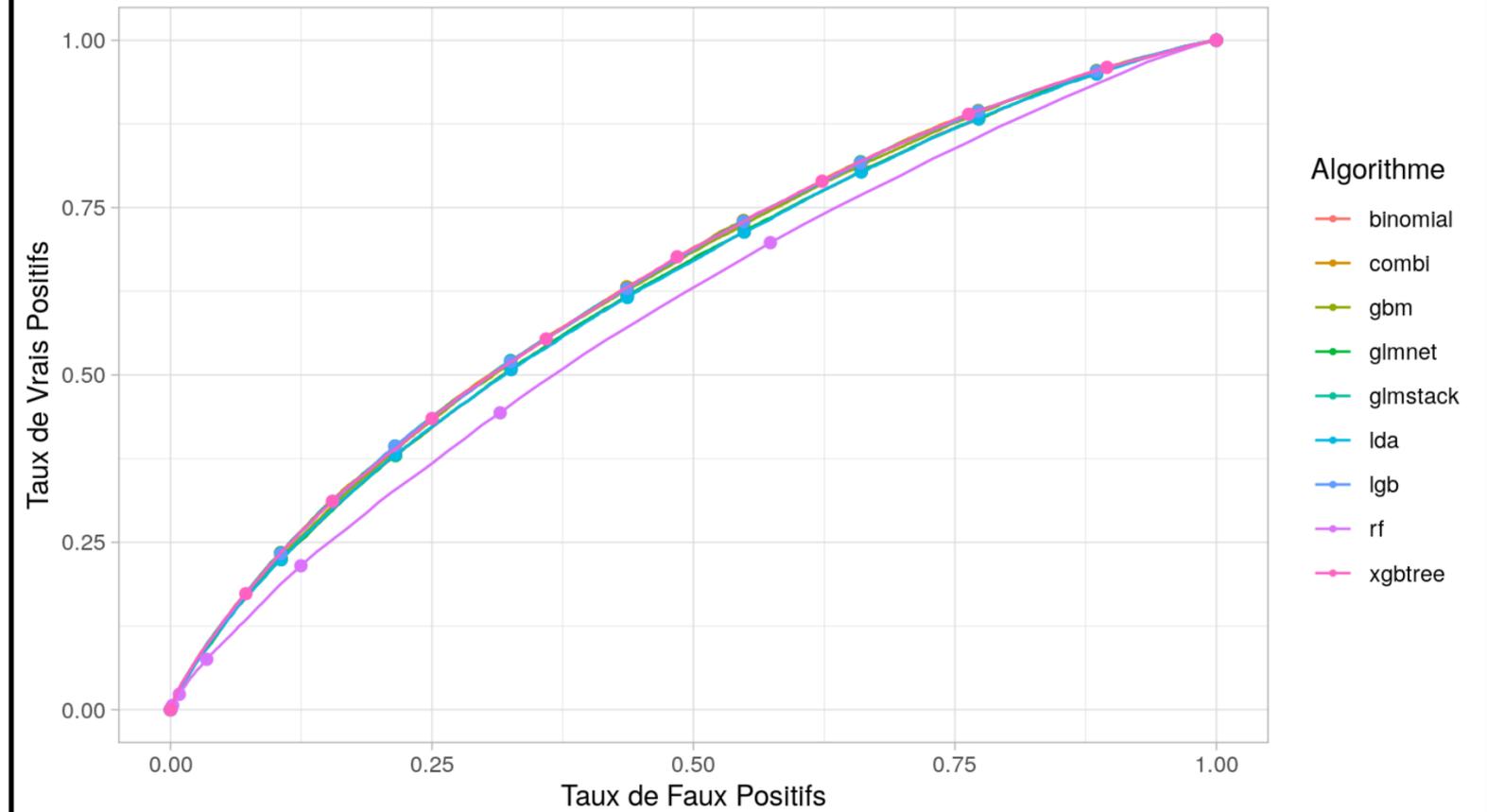
Comparaison des performances (validation croisée 5 blocs)

6 Modèles ont été retirés avant la construction des boxplots car trop mauvais  
 $glm\_stack = \text{stacking de XGB et LGB avec un GLM} \quad // \quad combi = 0.1 * pred\_LGB + 0.9 * pred\_XGB$



Courbes ROC

Prédictions 5fCV des modèles entraînés sans resampling

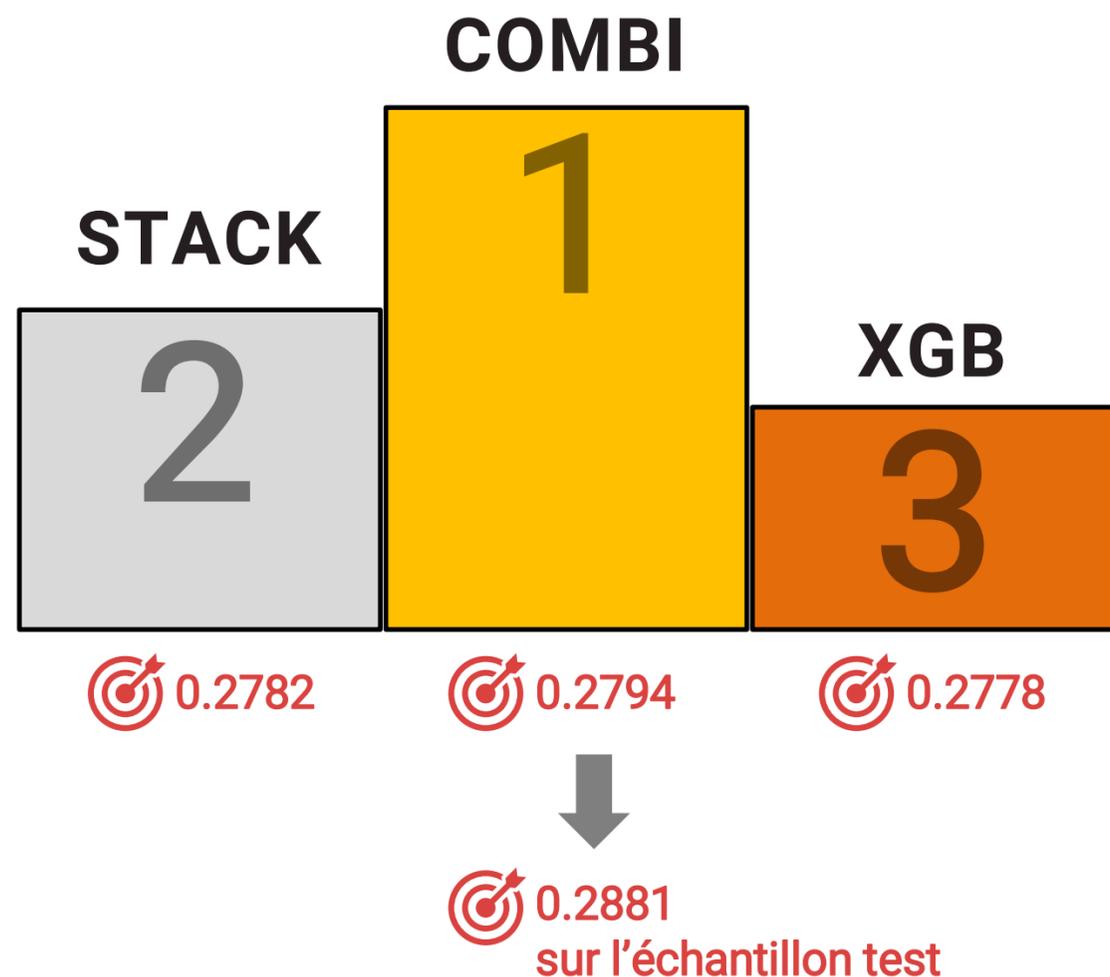


	GLM	LDA	GLMNET	SVM	KPPV	ARBRE	FORÊT	GBM	ADA	LGB	XGB	STACK	COMBI
Up	Modèle construit	Modèle construit	Modèle construit	Modèle non construit	Modèle non construit	Modèle construit	Modèle non construit	Modèle non construit	Modèle non construit	Modèle construit	Modèle construit	Modèle non pertinent	Modèle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

■ 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



# Annexes

Guy Tsang, Axel Gardahaut & Léo Dutertre-Ladurée

# ANNEXES (1/5)

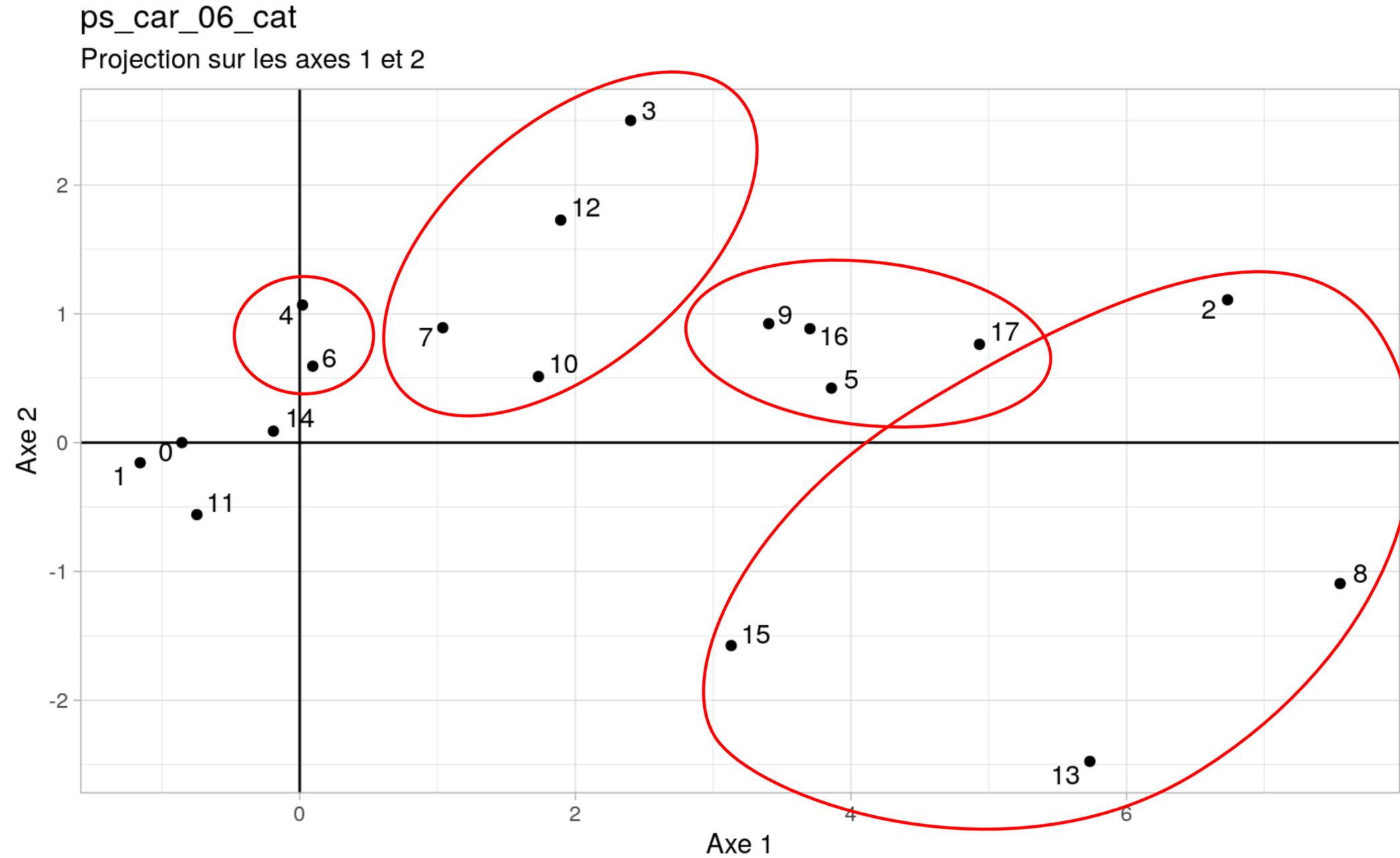
## HYPERPARAMETRES DU LGBM DE BENCHMARK

- Hyperparamètres

- Taux d'apprentissage : 0.01
- Profondeur maximale d'un arbre : 10
- Nombre maximal de feuilles par arbre : 20
- Nombre minimal d'observations dans une feuille : 10
- Part de variables à utiliser par itération : 0.8
- Bagging sur les observations : 0.8

# ANNEXES (2/5)

## STRATEGIE DE FUSION PAR AFDM

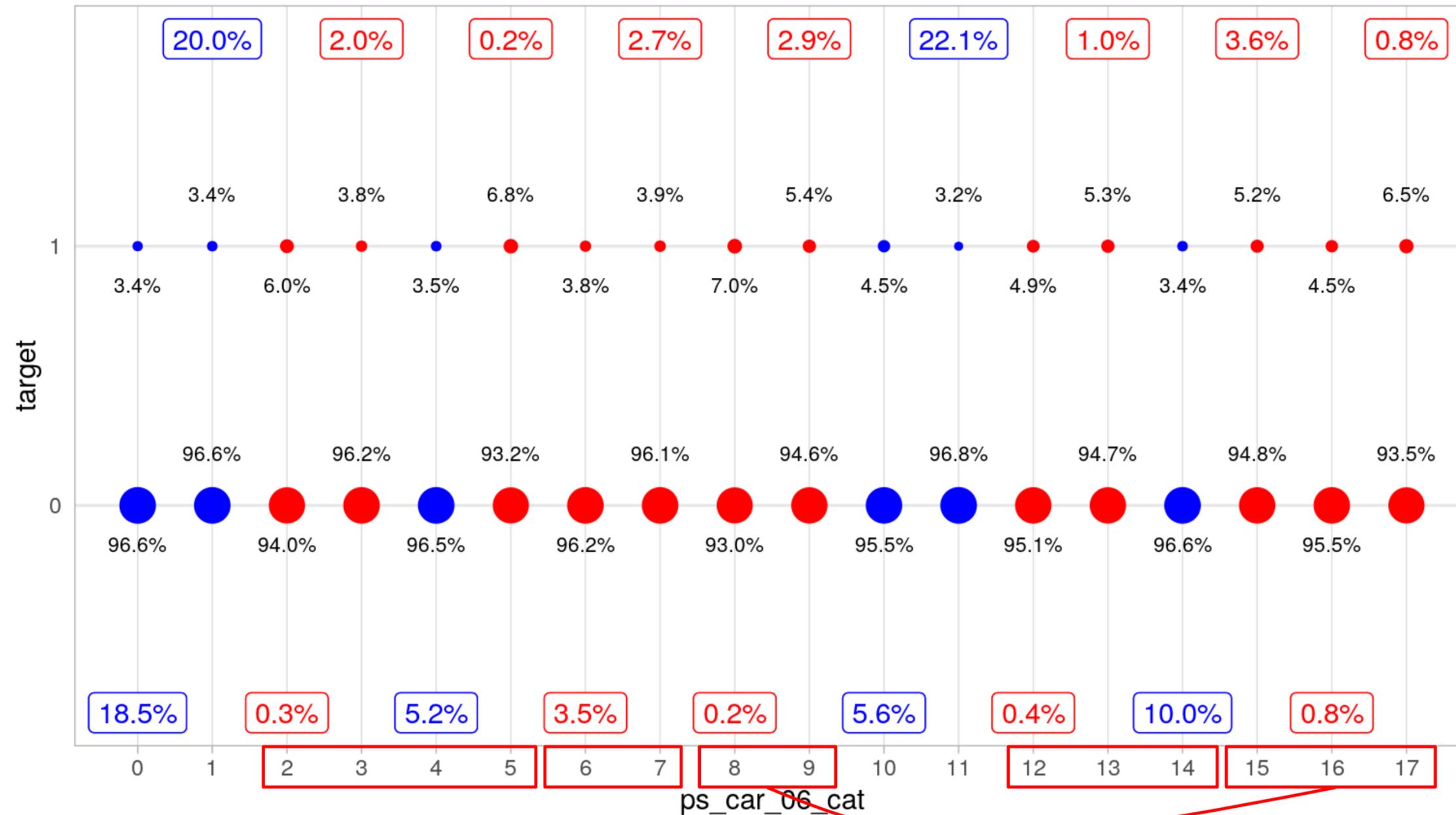


# ANNEXES (3/5)

## STRATEGIE DE FUSION PAR TABLEAUX CROISES

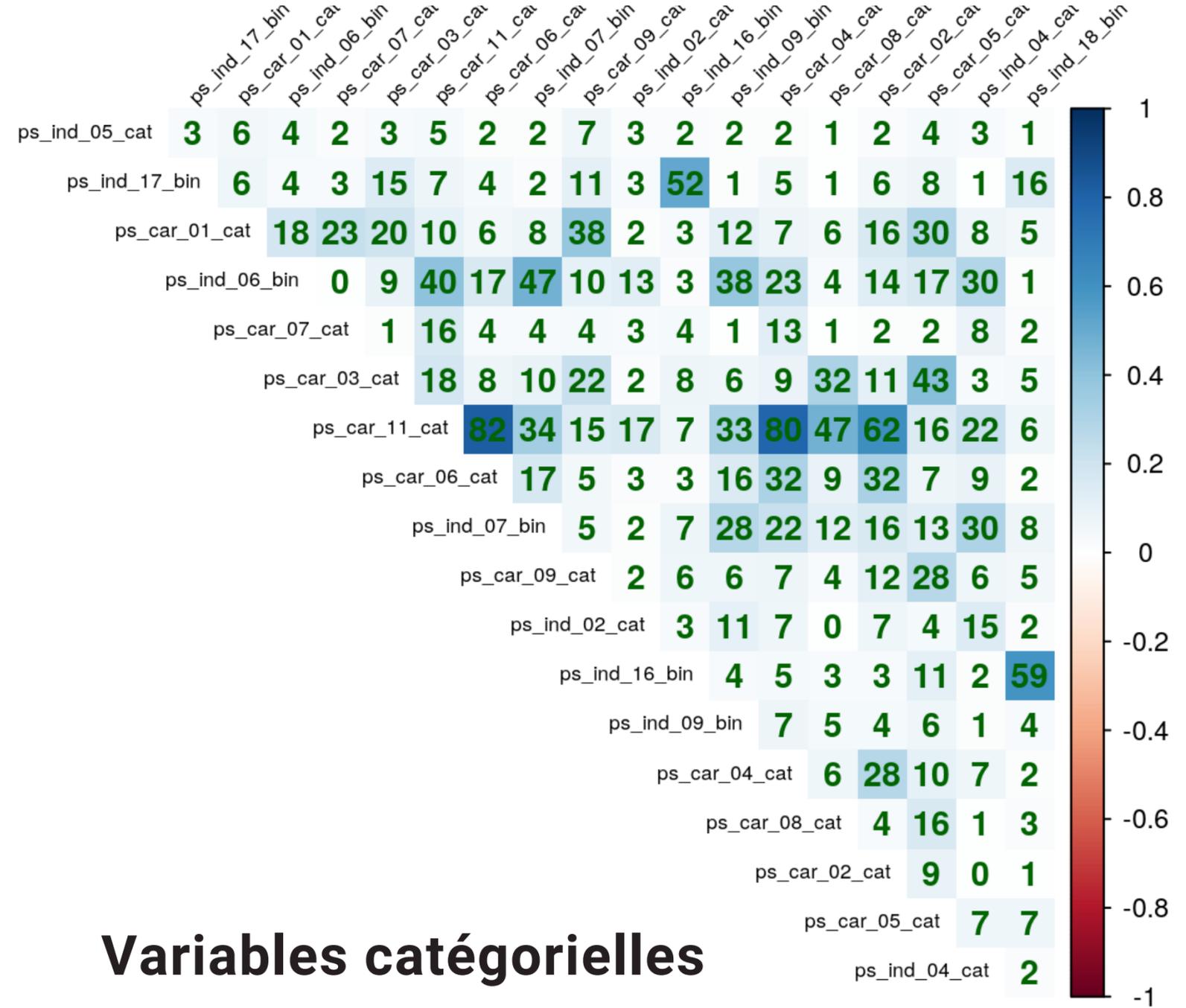
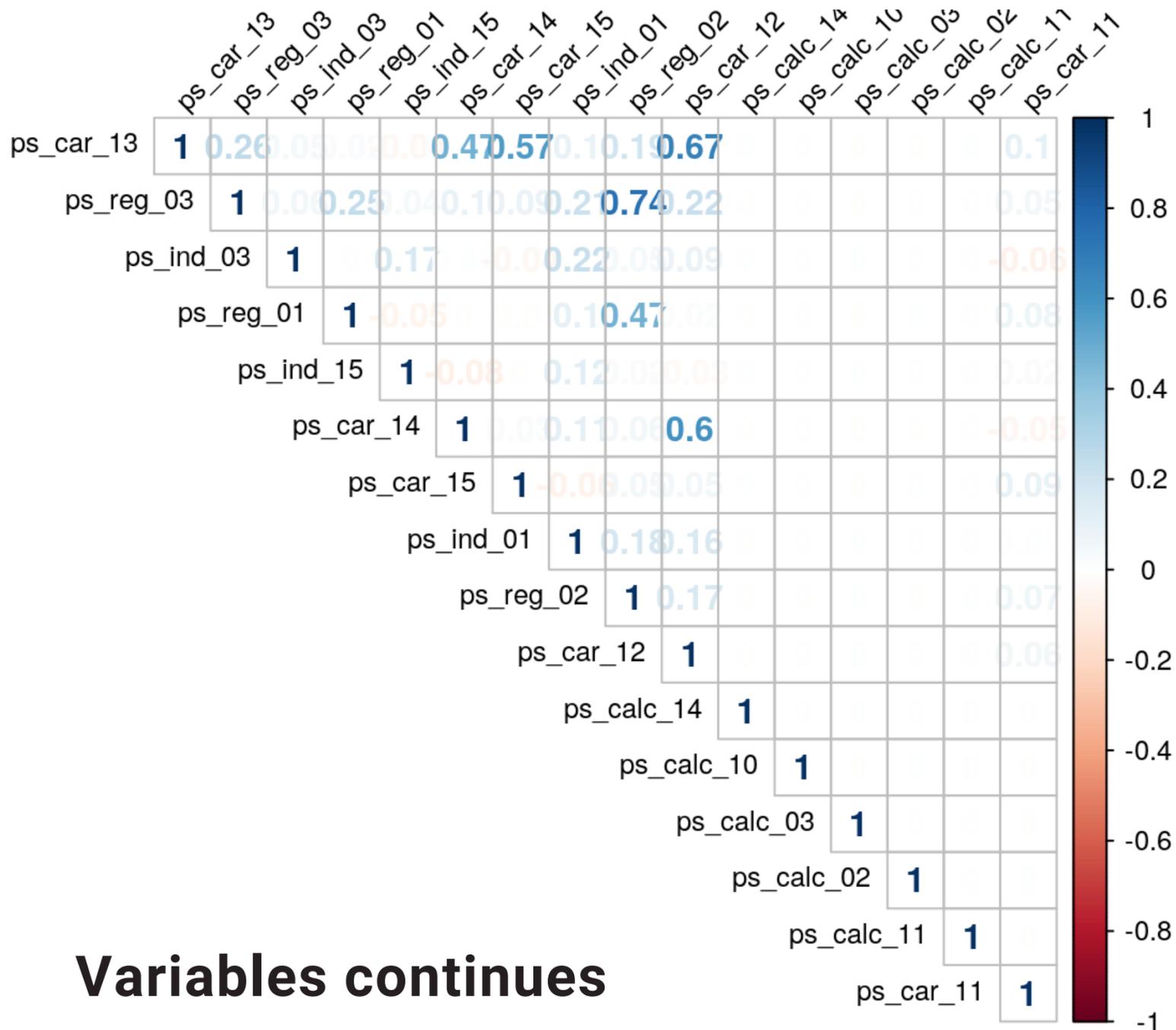
Fréquences marginales par modalité (ps\_car\_06\_cat)

Les modalités en rouge ont une fréquence inférieure à 5%



# ANNEXES (4/5)

## CORRELATION ET DEPENDANCE ENTRE VARIABLES



# ANNEXES (5/5)

## GRILLES D'HYPERPARAMETRISATION DES MODELES

- **Arbre de décision**
  - ✓ Profondeur maximale : 1 à 5 par 1
- **Elastic Net**
  - ✓ Alpha (pénalisation) : 0 à 1 par 0.1
  - ✓ Lambda (poids pénalisation) : 0 à 3 par 0.2
- **K plus proches voisins**
  - ✓ Nombre de voisins : 3
- **SVM**
  - ✓ Kernel : non linéaire (RBF)
  - ✓ Sigma (noyau) : 1
  - ✓ C (précision de la marge) : 1
- **Forêt aléatoire**
  - ✓ Variables par itération : 3 à 28 (80%) par 1
  - ✓ Fonction d'entropie : gini
  - ✓ Nombre minimal d'obs par feuille : 1
- **LGB**
  - ✓ Profondeur maximale : 5 à 15 par 5
  - ✓ Nombre max de feuilles : 10 à 25 par 5
  - ✓ Fraction de variables : 0.7 à 0.9 par 0.1
  - ✓ Fraction d'observations : 0.6 à 0.8 par 0.1
- **GBM**
  - ✓ Nombre d'arbres : 300 à 600 par 100
  - ✓ Profondeur maximale : 1 à 6 par 1
  - ✓ Taux d'apprentissage : 0.1
  - ✓ Nombre min d'obs par feuille : 1
- **AdaBoosting**
  - ✓ Nombre d'itérations : 500
  - ✓ Profondeur maximale : 5
  - ✓ Taux d'apprentissage : 0.1
- **XGB**
  - ✓ Nombre d'itérations : 100 à 500 par 100
  - ✓ Taux d'apprentissage : 0.05 ou 0.1
  - ✓ Profondeur maximale : 4 à 6 par 1
  - ✓ Gamma (seuil de gain minimal pour découpe) : 0
  - ✓ Part de variables à utiliser par arbre : 0.7
  - ✓ Poids minimal des feuilles enfant pour découpe : 0
  - ✓ Part des observations par itération : 0.7