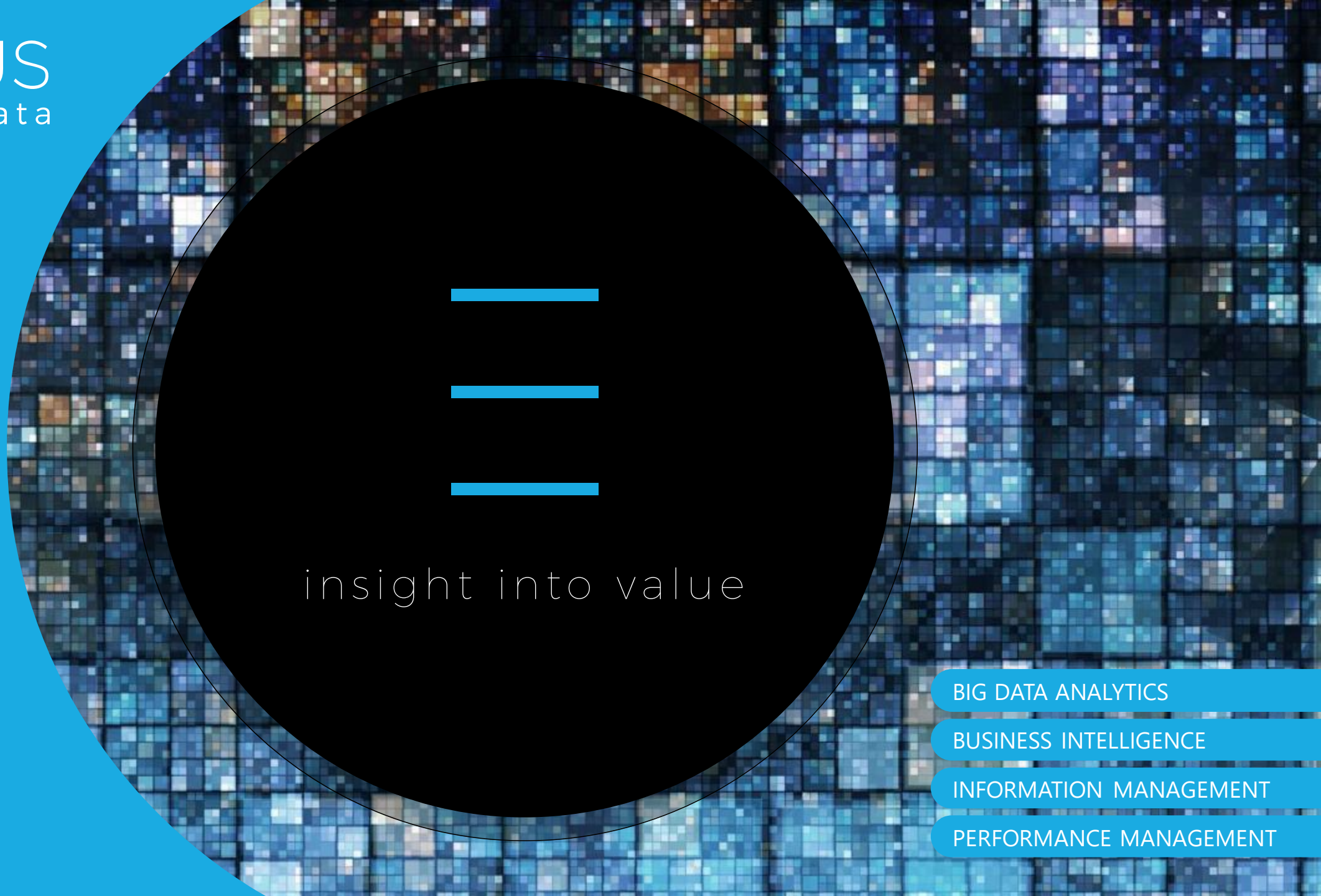
KEYRUS
data



insight into value

BIG DATA ANALYTICS

BUSINESS INTELLIGENCE

INFORMATION MANAGEMENT

PERFORMANCE MANAGEMENT

CLASSIFICATION BINAIRE EN PRÉSENCE DE DONNÉES FORTEMENT DÉSÉQUILIBRÉES

Avec exemple illustratif en R et Caret

Khalil El Mahrsi, Consultant Senior Data Science

7 novembre 2017

KEYRUS
data

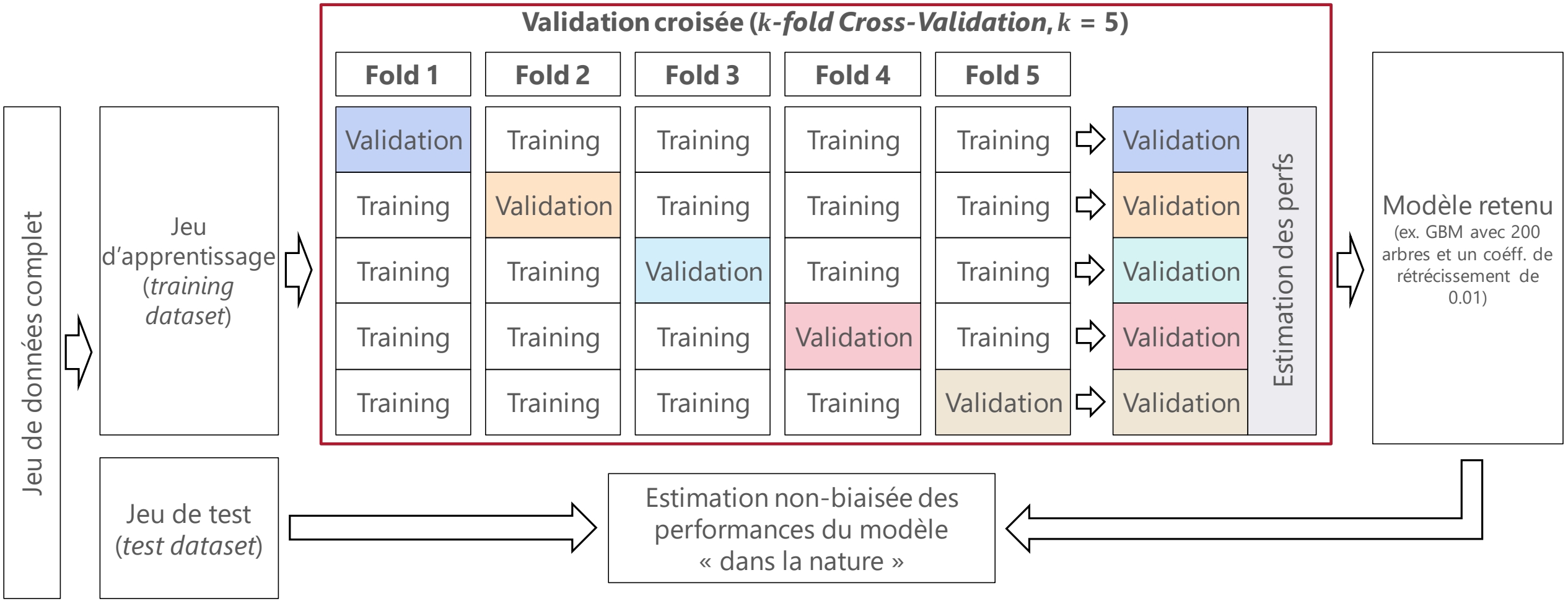
PROBLÈMES D'APPRENTISSAGE SUPERVISÉ

- On dispose d'un ensemble de variables descriptives (ex. montant de la transaction, adresse IP, date et heure, etc.)
- On souhaite « deviner » une variable de sortie (ex. transaction frauduleuse ou non)
- Supervisé = on dispose d'un jeu de données labellisé (c.à.d., où l'on connait la sortie)

- La nature de la variable de sortie définit le type du problème
  - Quantitative (continue) → Régression (ex. prédire le nombre de ventes d'un produit)
  - Qualitative (discrète) → Classification (ex. identifier les chiffres/lettres sur une image)
  - Qualitative binaire (deux modalités) → Classification binaire

DÉROULEMENT (CLASSIQUE) DE L'APPRENTISSAGE

Plusieurs modèles sont mis en compétition pour en choisir le meilleur

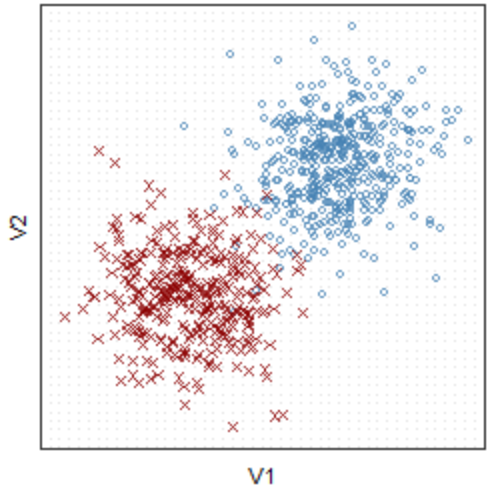


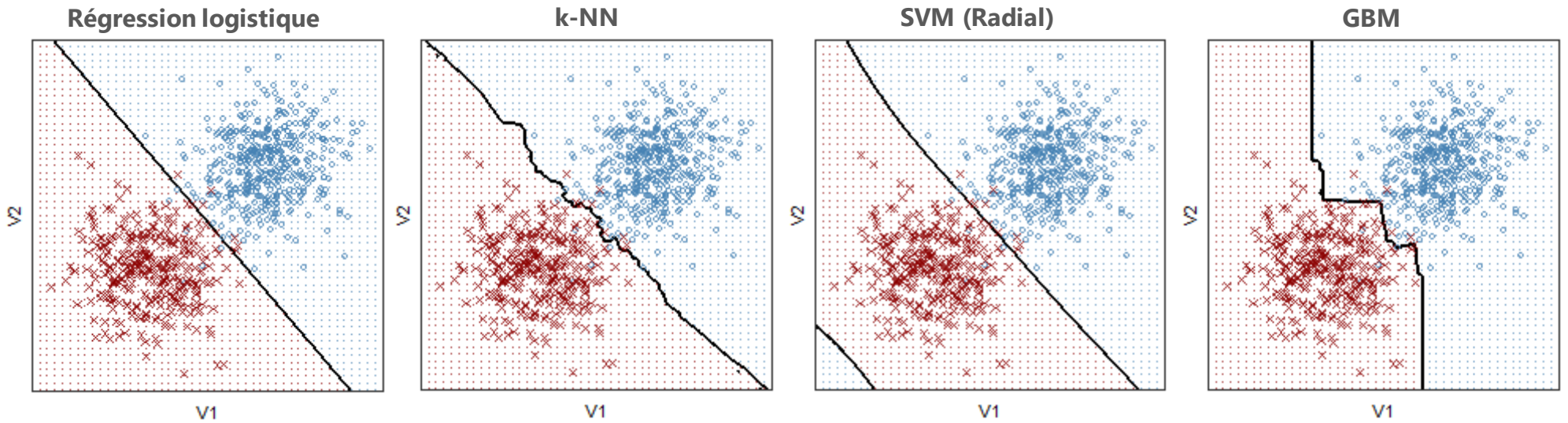
KEYRUS data

CLASSIFICATION BINAIRE : ON NOUS A (ENCORE) MENTI À L'ÉCOLE !!!

Cas d'études assez simples

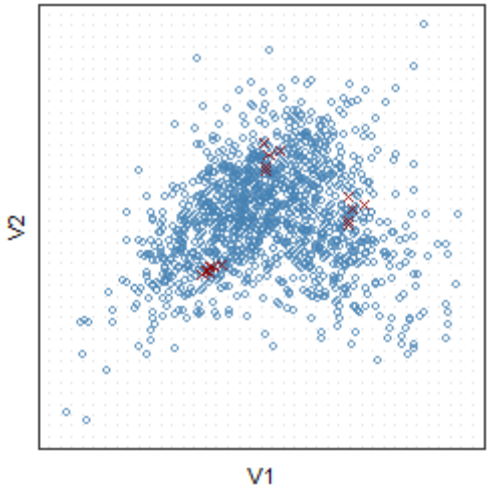
- Données nettoyées et propres
- Variable à prédire bien expliquée par les variables descriptives
- Classes à nombres d'effectifs similaires
- Facilement séparables sans trop d'efforts

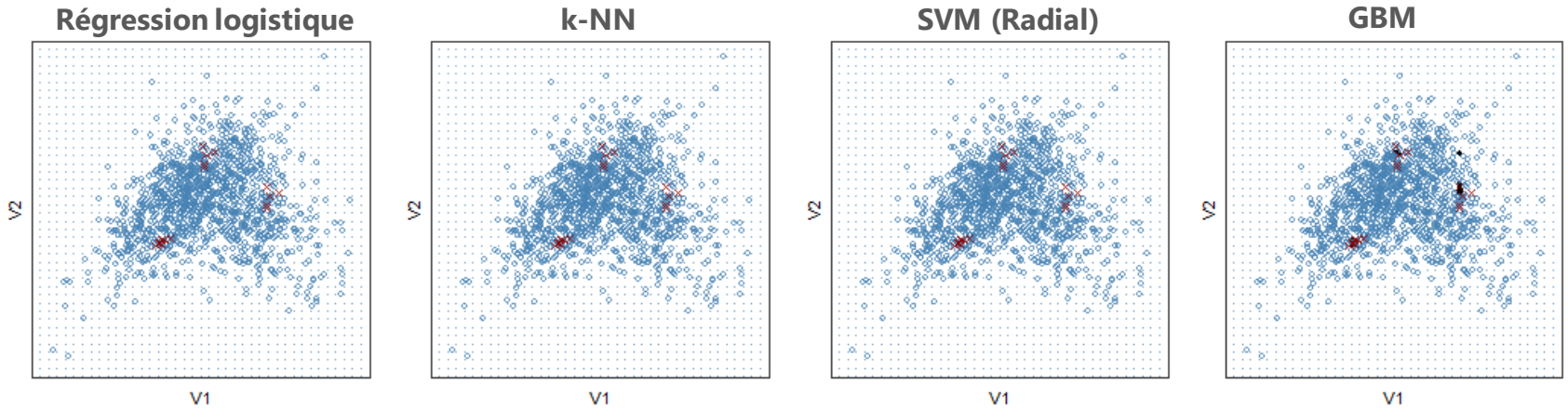




LE DÉSÉQUILIBRE DE CLASSES, ÇA N'ARRIVE PAS QU'AUX AUTRES

- Dans la pratique, c'est souvent plus compliqué…
  - Problèmes de qualité des données (ex. valeurs manquantes, biais, asymétrie, etc.)
  - **Parfois la classe d'intérêt est sous-représentée…**
  - **… et noyée dans des observations de la classe majoritaire**
- On parler de déséquilibre quand une classe constitue moins de 20% du jeu de données
  - Parfois, c'est bien pire (1-2%) → Classes fortement déséquilibrées
- Un cas qui survient dans plusieurs contextes
  - Identification des transactions bancaires frauduleuses
  - Prédiction des défaillances pour la maintenance prédictive
  - Prédiction du taux de clics pour la publicité en ligne





OÙ SE SITUE LE PROBLÈME ?

- Classiquement, le choix du meilleur modèle est basé sur le taux de bonne classification

$$\text{Taux de bonne classification} = \frac{\text{Nombre d'observations bien classées}}{\text{Nombre total d'observations}}$$

  - En cas de déséquilibre, prédire la classe majoritaire devient plus « rentable »
  - Phénomène de l'« accuracy paradox »

Quelle(s) métrique(s) utiliser pour comparer les modèles ?

- Par défaut, un seuil (cutoff) de probabilité de 0.5 est utilisé pour déterminer la classe

Comment choisir un cutoff plus adéquat ?

- Le déséquilibre de classes empêche d'apprendre à partir de la classe minoritaire

Comment combattre le déséquilibre et faciliter l'apprentissage de la classe minoritaire ?

AGENDA

Introduction

Choisir une métrique adéquate

Combattre le déséquilibre de classes

Choisir un bon cutoff

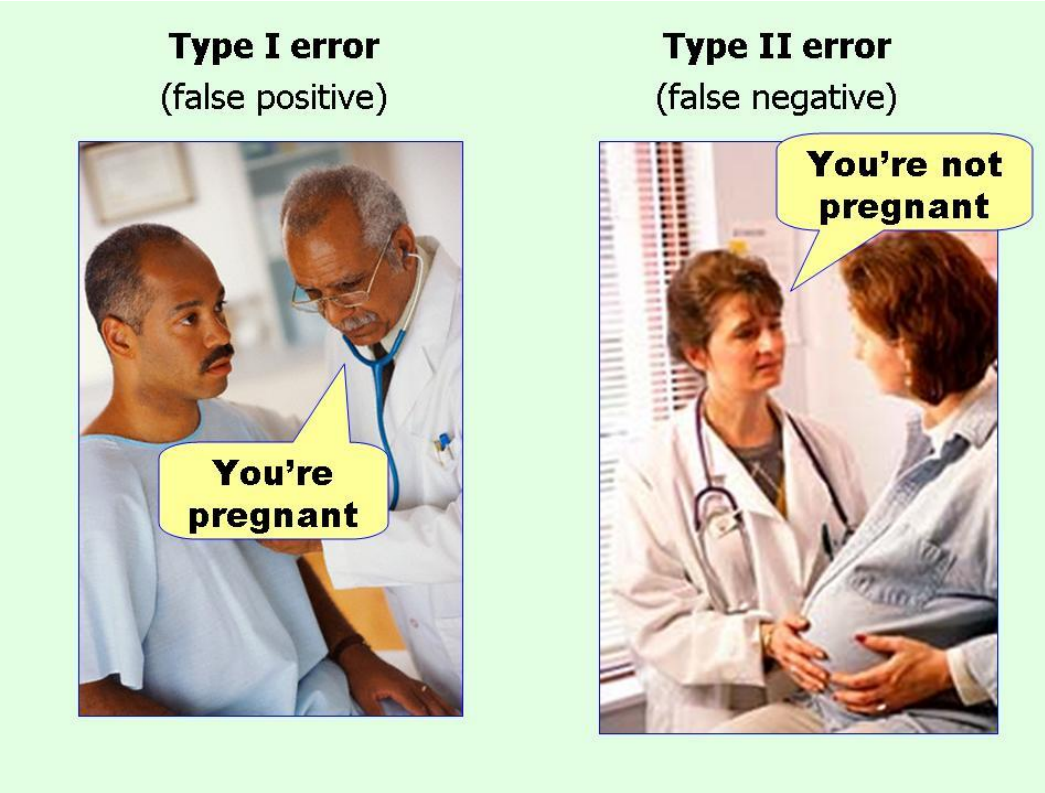
Mise en œuvre sur un exemple

TOUTES LES ERREURS NE SE VALENT PAS

- Le taux de bonne classification (utilisé par défaut) ne fait pas la différence entre erreur commise sur la classe d'intérêt et erreur commise sur la classe majoritaire

- La matrice de confusion permet d'avoir une meilleure vision des performances d'un classifieur binaire…
- … et de calculer plusieurs indicateurs utiles
  - Rappel (sensibilité, TPR) = TP / (TP + FN)
  - Précision = TP / (TP + FP)
  - Spécificité (TNR) = TN / (TN + FP)
  - F1-score = (2 x Précision x Rappel) / (Précision + Rappel)
  - Etc.

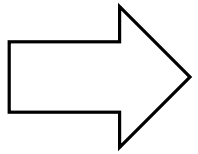
	Classe réelle	
Prédiction	**Oui**	**Non**
Oui	TP	FP
Non	FN	TN



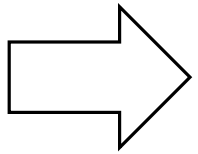
TYPES DE MÉTRIQUES POUR LA SÉLECTION DE MODÈLES

Deux types de métriques

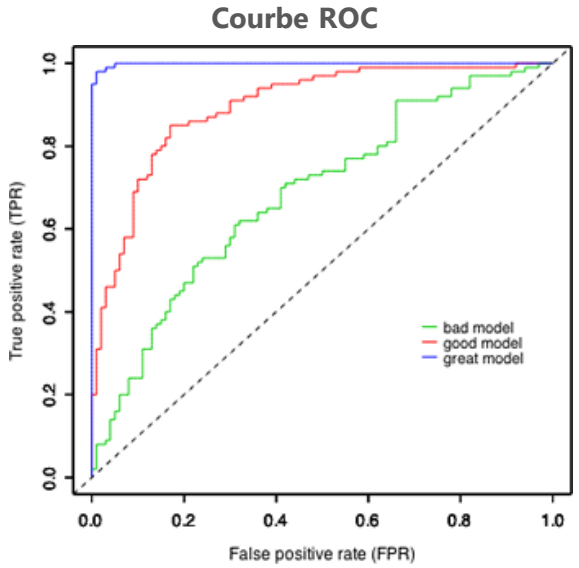
- Métriques dépendantes du cutoff
  - Précision et rappel
  - Sensibilité et spécificité
  - F1-score
- Métriques indépendantes du cutoff
  - ROC AUC (aire sous la courbe ROC)
  - PR AUC (aire sous la courbe Précision-Rappel)

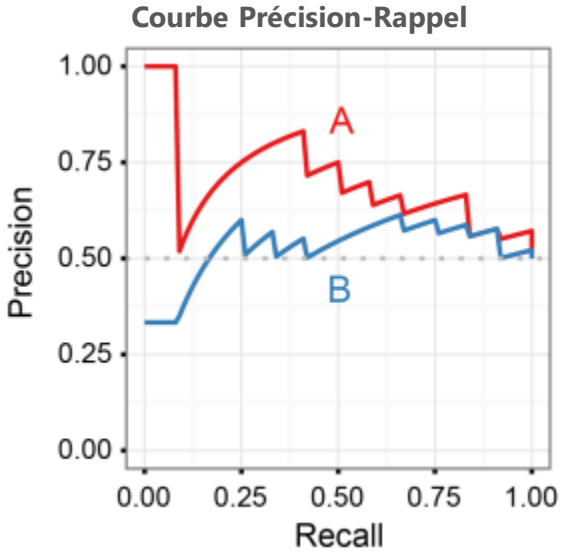


- Certaines doivent être inspectées par paire
- Inadéquates dans le cas déséquilibré
- Vision très restreinte (à un cutoff donné)
- Implémentations utilisant un cutoff de 0.5



- Offrent une vision globale des performances…
- … sur toute la plage de cutoffs





COURBE ROC ET DÉSÉQUILIBRE DE CLASSES

- La courbe ROC est « insensible » au déséquilibre de classes… [Fawcett, 2015]
- … et donne donc une vision excessivement optimiste des performances [Branco et al., 2015]
- La courbe Précision-Rappel s'avère être plus utile en cas de déséquilibre [Saito and Rehmsmeier, 2015]

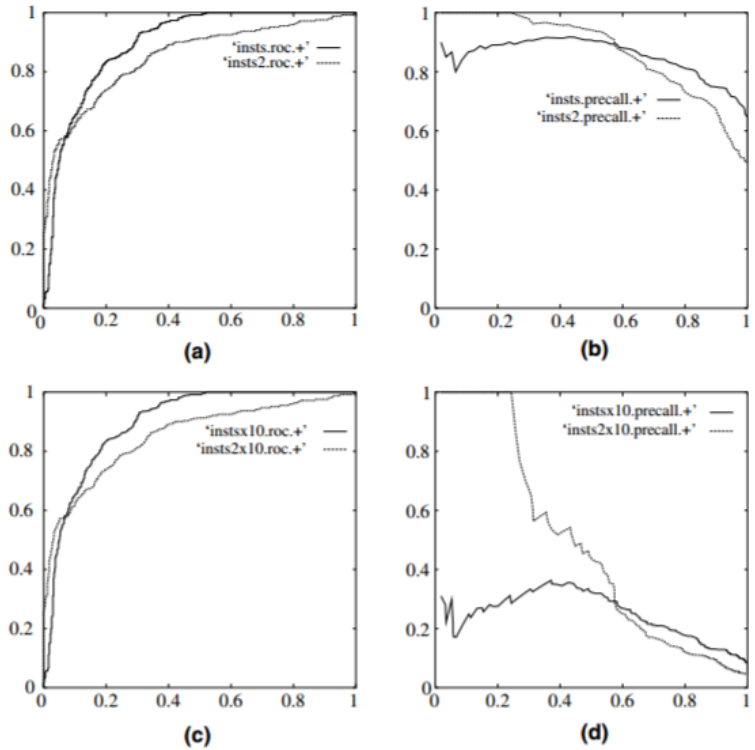


Fig. 5. ROC and precision-recall curves under class skew. (a) ROC curves, 1:1; (b) precision-recall curves, 1:1; (c) ROC curves, 1:10 and (d) precision-recall curves, 1:10.

Figures issues de [Fawcett, 2015]

AGENDA

Introduction

Choisir une métrique adéquate

Combattre le déséquilibre de classes

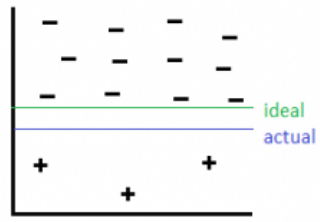
Choisir un bon cutoff

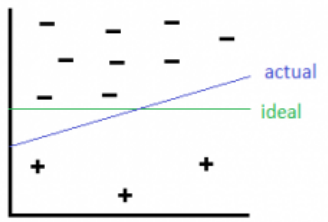
Mise en œuvre sur un exemple

COMBATTRE LE DÉSÉQUILIBRE À LA SOURCE: RÉ-ÉCHANTILLONNAGE

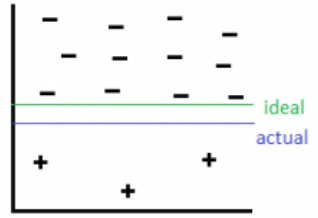
Trois possibilités pour rééquilibrer les données

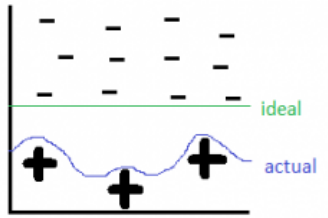
- Sous-échantillonner la classe majoritaire
  - Accélération des temps de calcul
  - Risque de supprimer des observations utiles



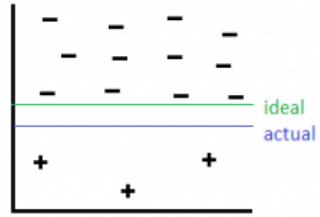


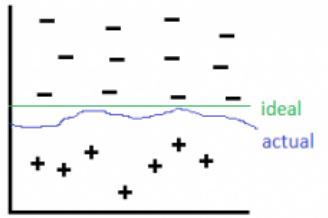
- Sur-échantillonner la classe minoritaire
  - Risque de sur-apprendre un nombre réduit d'exemples de la classe minoritaire
  - Augmentation des temps de calcul





- Ré-échantillonnage hybride (SMOTE et co.)
  - Synthétiser de nouvelles observations de la classe minoritaire…
  - … et sous-échantillonner la classe majoritaire
  - Combine les avantages et les inconvénients du sous-échantillonnage et du sur-échantillonnage





Figures issues de http://www.chioka.in/class-imbalance-problem/

EXEMPLE D'ÉCHANTILLONNAGE HYBRIDE: SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE)

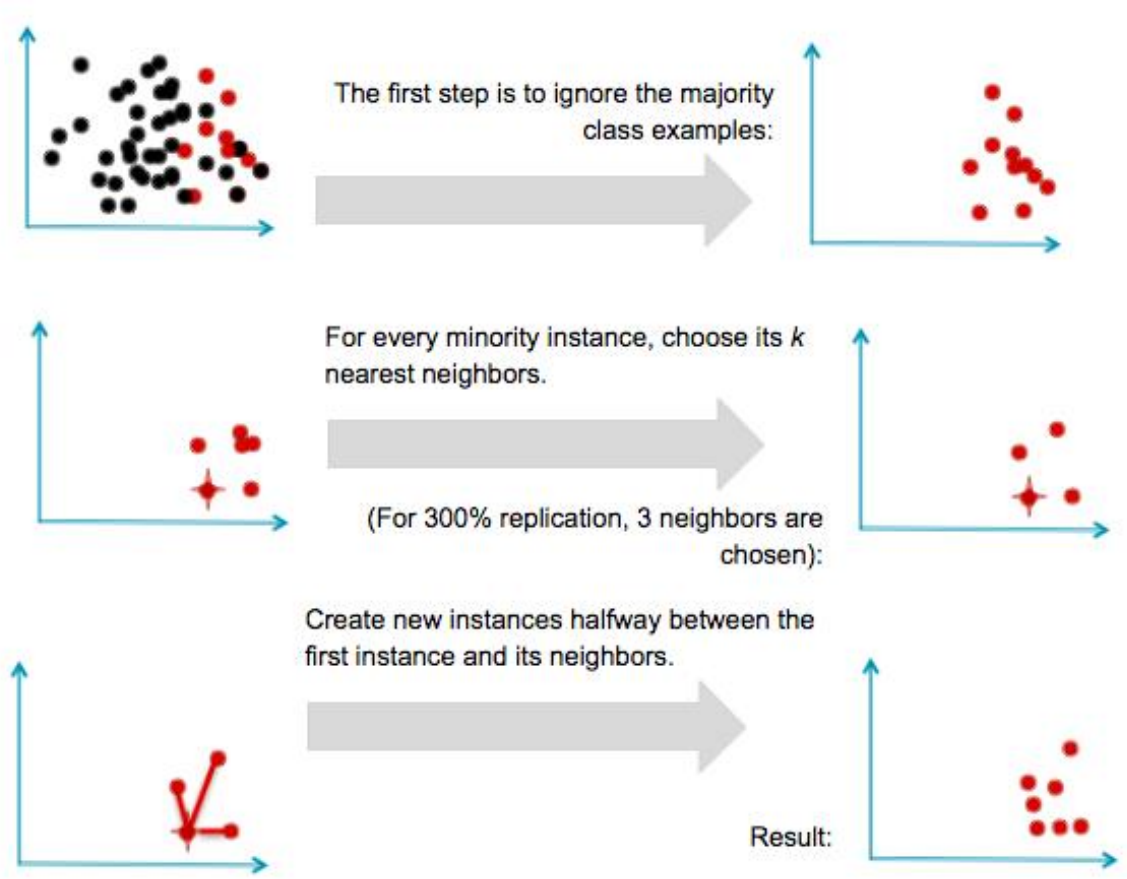
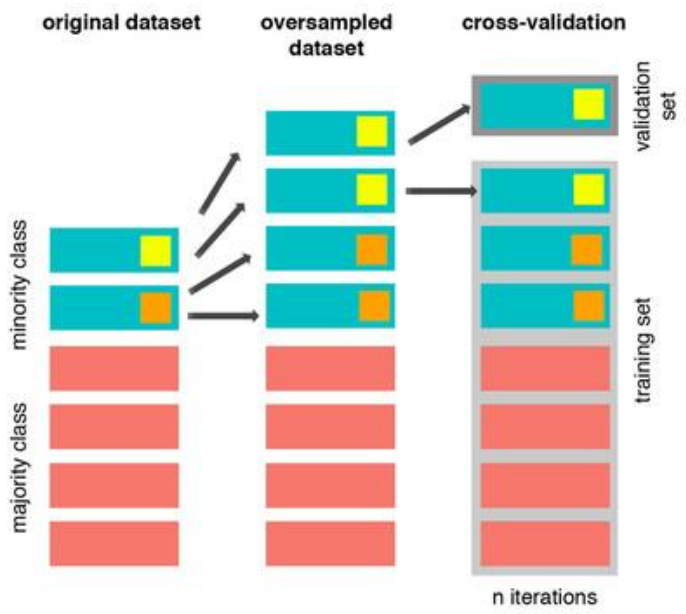


Figure issue de https://www.svds.com/learning-imbalanced-classes/

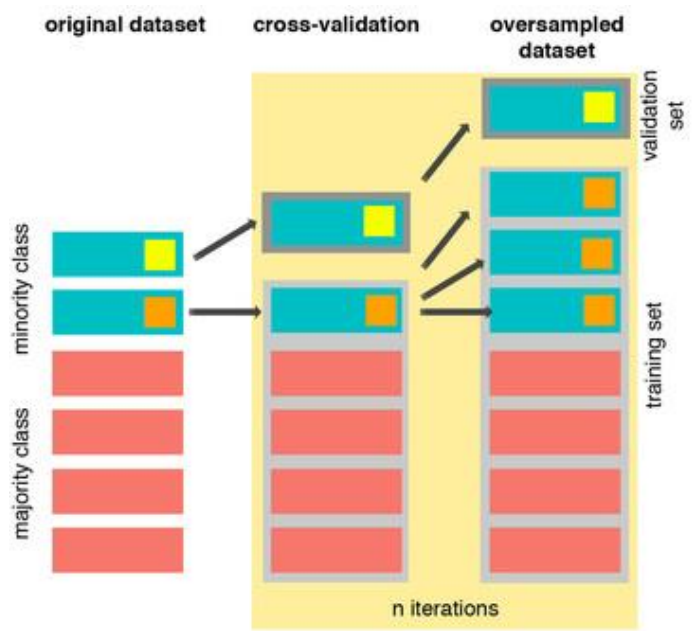
PIÈGES À ÉVITER LORS DU RÉ-ÉCHANTILLONNAGE

- Il ne faut jamais ré-échantillonner le jeu de test. Jamais. Jamais ! JAMAIS !!!
- Lors de la sélection de modèles, le ré-échantillonnage doit être effectué à l'intérieur de la validation croisée
  - À chaque fois, le fold de validation doit rester déséquilibré…
  - … et seuls les k – 1 folds sur lesquels le modèle est appris sont ré-équilibrés

Mauvaise façon



Bonne façon



Figures issues de http://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation

COMBATTRE LE DÉSÉQUILIBRE DANS LE MODÈLE: PONDÉRER LES CLASSES

- Certains algorithmes (ex. SVM, GBM, réseaux de neurones, etc.) offrent la possibilité de pénaliser différemment les erreurs commises sur chaque classe
  - Imposer une pénalité plus grande quand on se trompe sur la classe minoritaire
  - Et vice versa

- Cost-Sensitive Machine Learning

- Stratégie simple: pondérer en fonction du nombre d'exemples dans chaque classe

$$\text{Coût}(C_i) = 0.5 \times \frac{1}{\text{Nombre d'observations appartenant à la classe } C_i}$$

- Les pénalisations peuvent aussi refléter des coûts métier (« réels »)
  - Ex. « Une défaillance non détectée me revient 10 fois plus cher que lever une fausse alerte »
    - $\text{Coût}(\text{défaillance}) = 10$
    - $\text{Coût}(\text{normal}) = 1$

AGENDA

Introduction

Choisir une métrique adéquate

Combattre le déséquilibre de classes

Choisir un bon cutoff

Mise en œuvre sur un exemple

CHOISIR UN CUTOFF APPROPRIÉ

- Par défaut, l'algorithme de classification va affecter une classe « en dur »
  - La plupart des algorithmes estiment la probabilité de la classe d'intérêt
    - Probabilité supérieure à 0.5 → observation affectée à la classe
    - Sinon → observation est affectée à la classe restante
  - Possibilité de récupérer la probabilité et de choisir soi même un cutoff plus approprié

- Le choix du meilleur cutoff est une question ouverte (30+ méthodes dans le package `OptimalCutpoints`)
  - Maximiser l'indice de Youden (Spécificité + Sensibilité – 1)
  - Maximiser l'indice de Kappa
  - Maximiser le F1-score
  - Etc.

- Un seul conseil: laissez le besoin du client vous guider
  - Client A: « Je m'en fous des fausses alertes ! L'essentiel c'est de bien identifier le maximum d'incidents !!! »
  - Client B: « J'aimerai bien prédire le maximum de défaillances possibles, mais une fausse alerte est coûteuse pour moi ! »

- Le choix du cutoff doit être effectué sur un jeu à part. Pas le jeu d'apprentissage. Pas le jeu de test

AGENDA

Introduction

Choisir une métrique adéquate

Combattre le déséquilibre de classes

Choisir un bon cutoff

Mise en œuvre sur un exemple

RÉCAPITULATIF

- « Minimum syndical » à mettre en place
  - Choisir une bonne métrique (ROC AUC, ou de préférence PR AUC)
  - Récupérer les probabilités des classes et optimiser le cutoff
  - Être attentif au besoin exprimé par le client → Guidera le choix de la méthode

- Essayer ensuite d'aller plus loin
  - Mettre en concurrence les différentes stratégies pour combattre le déséquilibre
    - Ré-échantillonner les données (up-sampling, down-sampling, échantillonnage hybride)
    - Pondérer les classes (pour les méthodes qui l'acceptent)
  - Choisir la meilleure (si ça améliore les choses)

- Pièges à éviter
  - Ré-équilibrage du jeu de test → JAMAIS
  - Ré-équilibrage du jeu d'apprentissage → Il faut le faire à l'intérieur de la CV lors de l'évaluation
  - Si vous utilisez `R` et `Caret` → Attention à quelle classe est utilisée lors de l'évaluation

ET SI ÇA NE MARCHE TOUJOURS PAS ?

- Se procurer plus de données (pas toujours possible)

- Essayer avec des algorithmes conçus spécifiquement pour le cas déséquilibré
  - Ex. [Goh et Rodin, 2013]

- Essayer un changement de paradigme (ex. Détection d'anomalies ?)
  - Problème d'estimation de la densité
  - Apprentissage semi-supervisé avec utilisation des auto-encodeurs
  - Etc.

- Peut être que les données ne sont pas adéquates à la tâche tout simplement

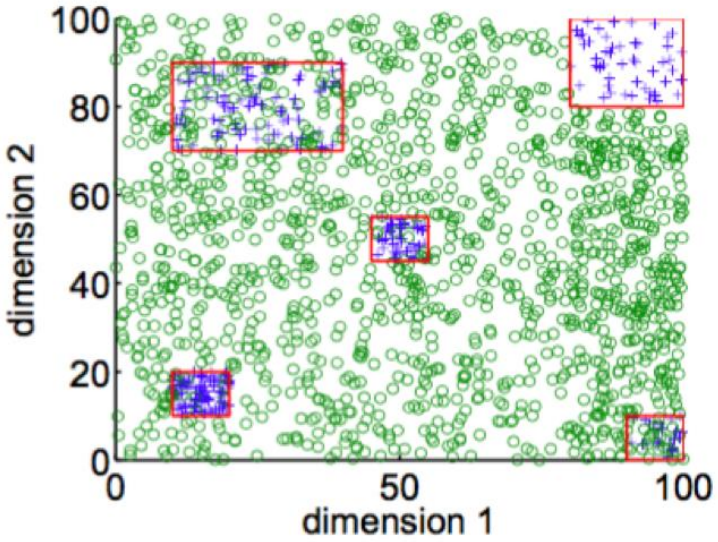


Figure issue de https://www.svds.com/learning-imbalanced-classes/

QUELQUES RÉFÉRENCES POUR ALLER PLUS LOIN

- Learning From Imbalanced Classes
- Simple guide to confusion matrix terminology
- Handling Class Imbalance with R and Caret - An Introduction
- Handling Class Imbalance with R and Caret - Caveats when using the AUC
- The Relationship Between Precision-Recall and ROC Curves [Davis et Goadrich, 2006]
- The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets [Saito et Rehmsmeier, 2015]
- Dealing With Imbalanced Data: Undersampling, Ovesampling and Proper Cross-Validation

MERCI DE VOTRE ATTENTION

Pour nous contacter

contact@keyrus.com

www.keyrus.com

KEYRUS
data