24e CONGRÈS DES ACTUAIRES

BY Institut.

17 juin 2025 - PARIS

EXPRESSO TECHNIQUE

14h - 14h20

DONNÉES DÉSÉQUILIBRÉES EN ASSURANCE : PHÉNOMÈNE, DÉFIS ET SOLUTIONS

Samuel STOCKIEKER
ACTUNEO

Denys POMMERET
AIX MARSEILLE UNIVERSITÉ

Amina BOURAS
TRILOGYC

🧠 Apprentissage automatique : la qualité des données avant tout

📄 Principe fondamental : « Garbage in, garbage out »

- Même les modèles les plus avancés **ne peuvent apprendre que ce que les données leur permettent d'apprendre**.
- Des données **incomplètes**, **bruitées**, ou **déséquilibrées** mènent à des prédictions **peu fiables ou injustes**.

🧩 Trois piliers de la qualité des données pour la modélisation:

1. **Pertinence** : variables explicatives alignées avec la réalité métier (ex : comportement de souscription, historique sinistre, réclamations…)
2. **Représentativité** : tous les profils importants doivent être présents en nombre suffisant
3. **Équilibre** : attention particulière aux cas rares à fort enjeu

🎯 Pourquoi c'est clé dans notre métier ?

Nos décisions impactent directement des clients : fiabilité, équité, anticipation = impératifs métier.

🎯 Pourquoi s'intéresser au déséquilibre des données en assurance ?

- **Plusieurs cas en assurance**

  - **Détection de fraude :**

    - Cas de fraude très minoritaires (<1 %)
    - Impacts financiers et réputationnels élevés
    - Comportements évolutifs et hétérogènes

  - **Anticipation de la résiliation (churn) :**

    - Résiliations parfois inférieures à 5 %
    - Perte de valeur vie client, coût de reconquête
    - Raisons multiples, souvent peu visibles (insatisfaction, multi-équipement, opportunisme)

  - **Sinistralité: occurrence de sinistres rares (MRH, nouveaux produits, etc.)**

  - **Anticipation des rachats, estimation de taux de transformation**
  - **…**

Ce que ces cas ont en commun :

- Données souvent **très déséquilibrées**
- Signaux faibles, **profils rares mais à fort enjeu**
- Modèles classiques souvent inefficaces sans adaptation

⚠ quel impact pour l'utilisateur ?

Risques concrets liés au déséquilibre :

- Les indicateurs d'évaluation standards (ex. : précision, AUC) masquent souvent la faiblesse du modèle à détecter les vrais résiliants ou fraudeurs
- Modèles qui « jouent la sécurité » : sur-prédiction des cas les plus fréquents **clients normaux**
- Actions de prévention ou de rétention **mal ciblées**, donc inefficaces

Conséquences pour les équipes métier :

- Besoin de comprendre **où les modèles se trompent**
- Temps perdu à comprendre et reparamétrer ou changer de modèle
- Parfois perte de confiance dans le model !
- Risque d'abandon du modèle et retour à des méthodes manuelles

Quelle approche / démarche spécifique face à ce phénomène?

La génération de données synthétiques

Pre-processing : une solution universelle

- On travaille sur le jeu de données initial
- On génère des données synthétiques pour rééquilibrer les « classes minoritaires ».
- On dispose alors d'un jeu rééquilibré que l'on utilise pour l'entrainement de notre modèle

La génération de données synthétiques

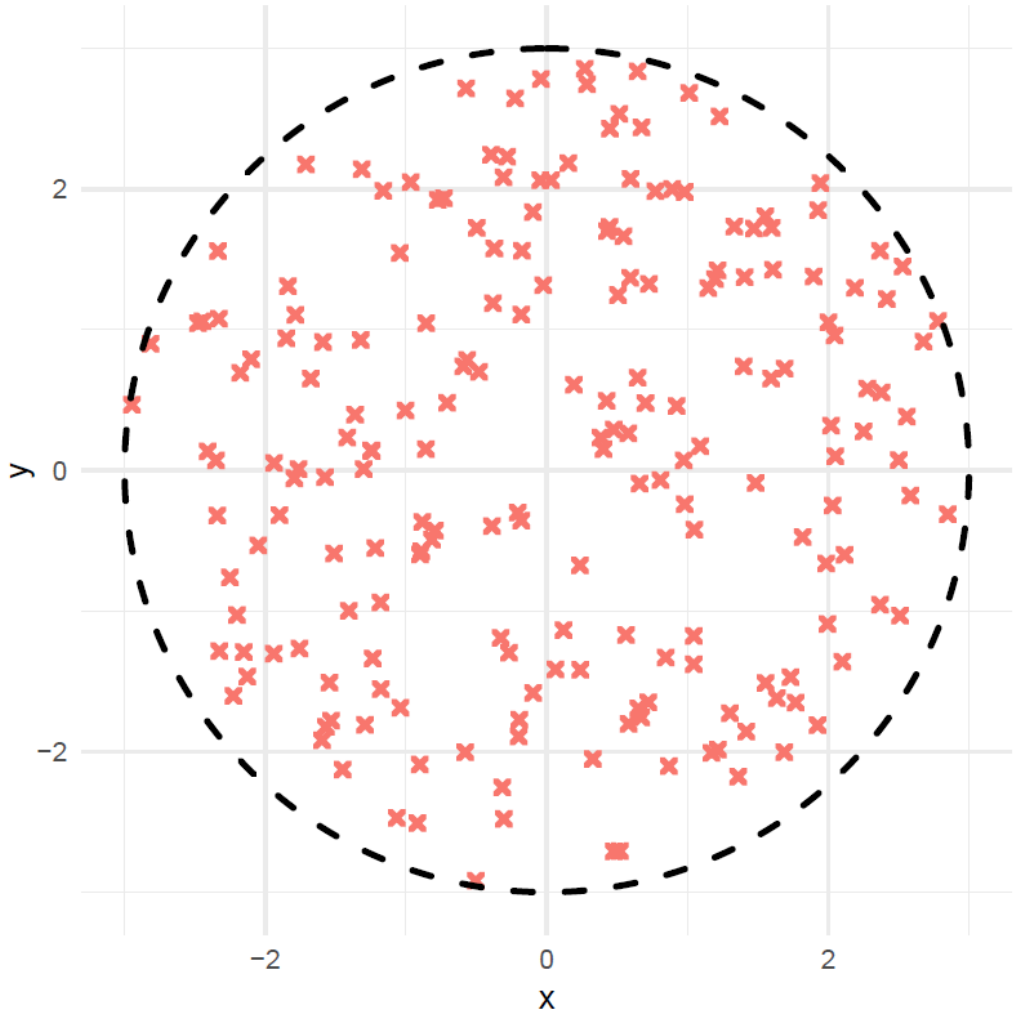
Algorithme SMOTE (Synthetic Minority Oversampling Technique)

➔ Interpolation linéaire dans la classe minoritaire : $M = \{X_1, \ldots, X_n\}$

For i = 1, . . . ,m:

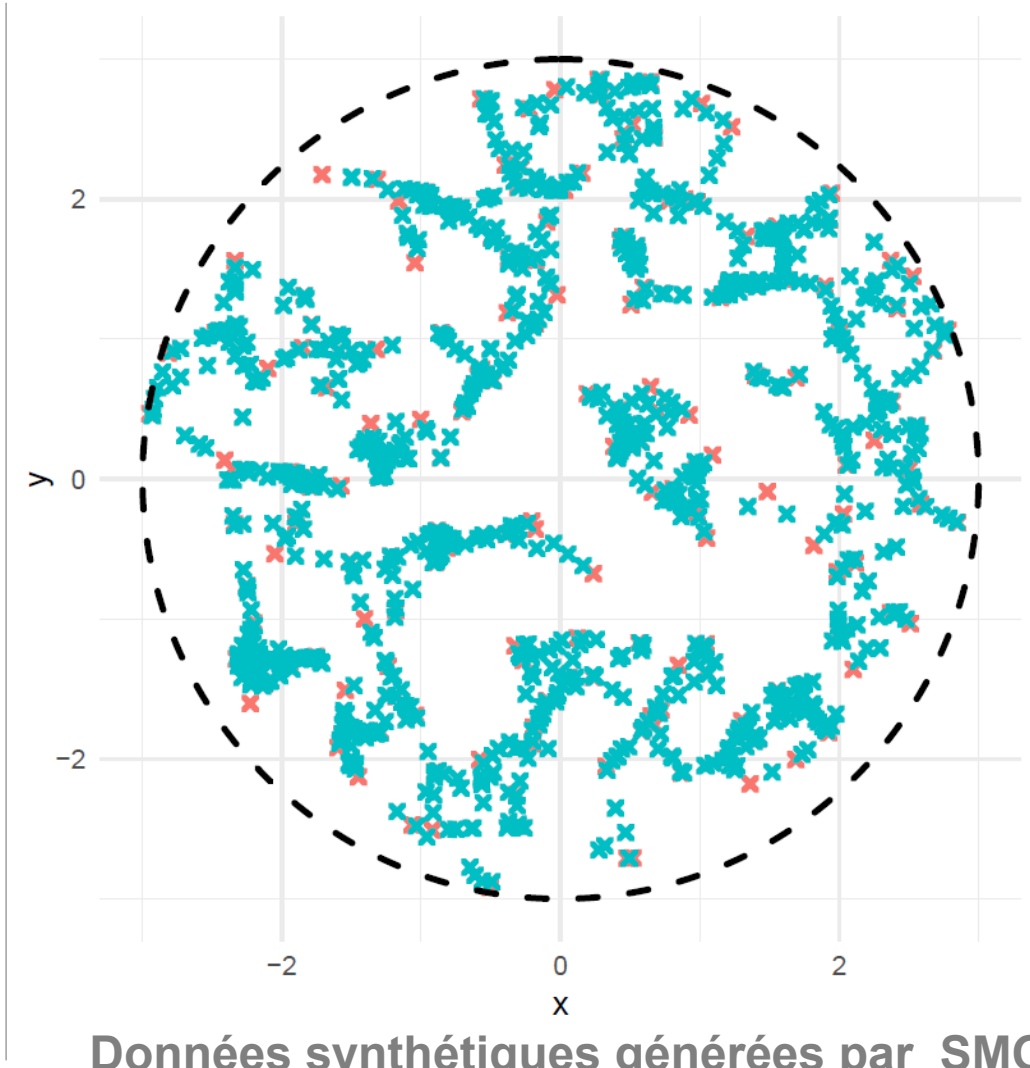
1. Tirage uniforme dans M : X_s
2. Tirage de X_n parmi les k plus proches voisins de X_s
3. Génération de $X^* = X_s + \lambda(X_n - X_s)$, avec λ uniforme

La génération de données synthétiques



Données initiales dans une région minoritaire

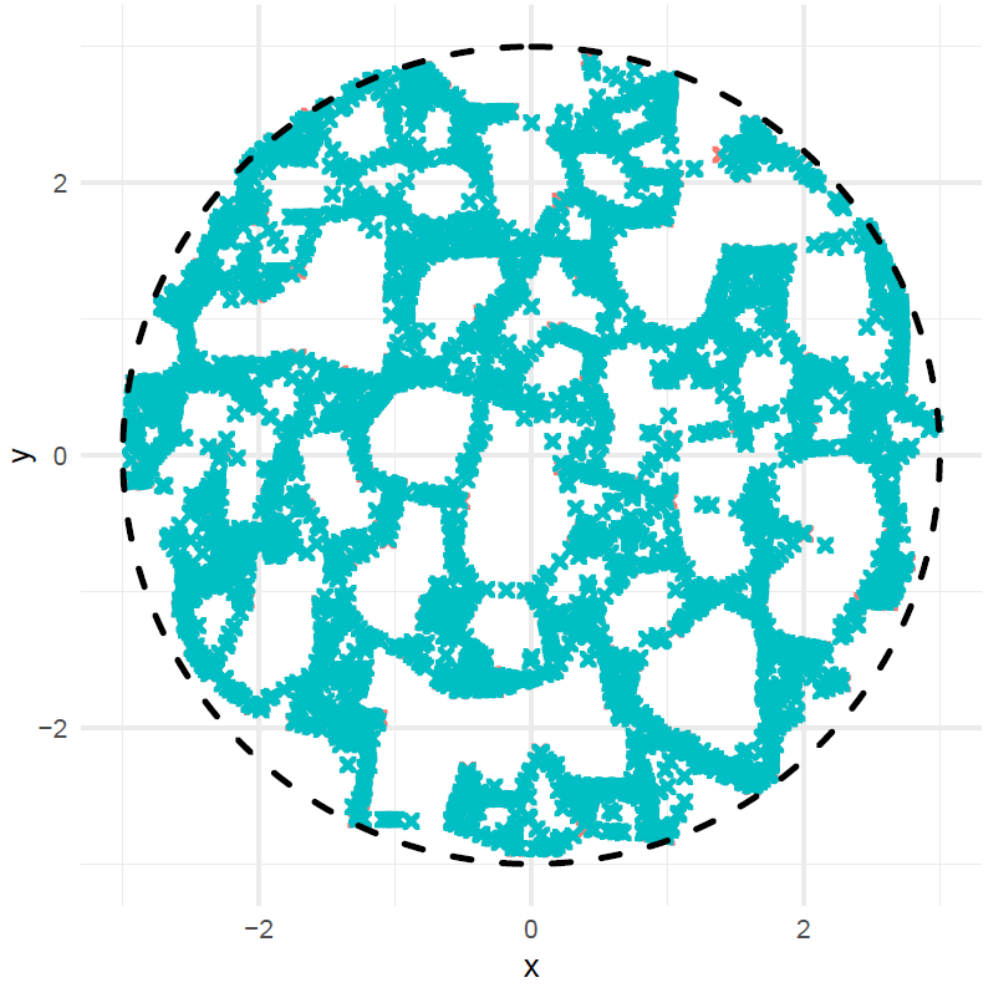
La génération de données synthétiques



Données synthétiques générées par SMOTE

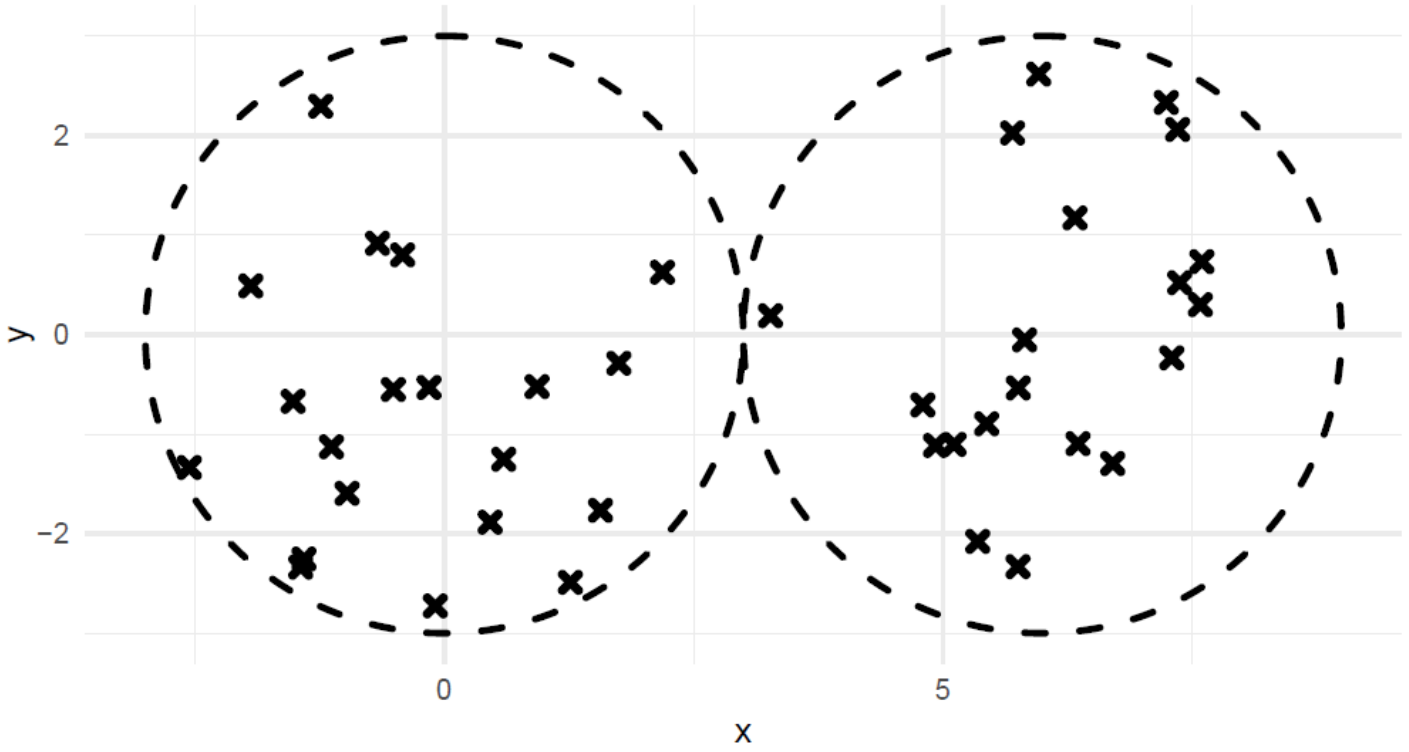
Un phénomène de type « emmental » ou« gruyère »

La génération de données synthétiques



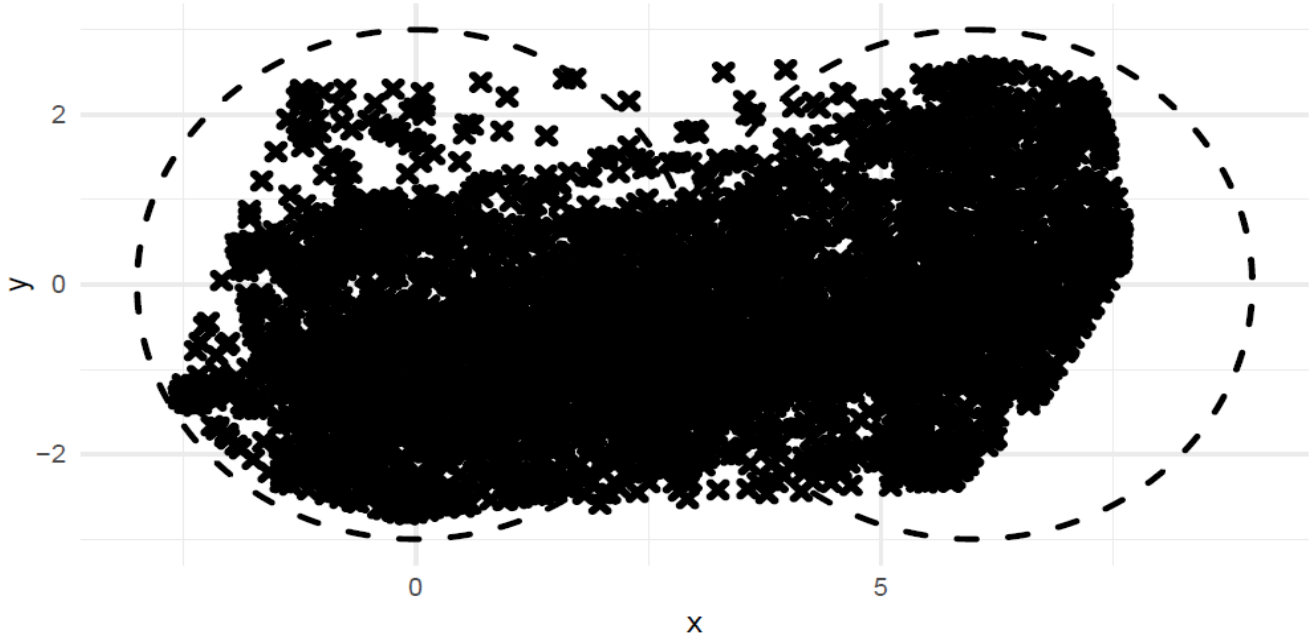
Trop ou pas assez de voisins ? Dans tous les cas, le bord du support est mal exploré (on ne va pas au-delà des individus « extrêmes »)

La génération de données synthétiques



Et si le domaine n'est pas convexe ?

La génération de données synthétiques



Augmenter le nombre de voisins peut être inefficace : peut créer des individus aberrants

Limites de SMOTE

- **Choix du nombre de voisins**

- trop de voisins : on perd la distribution de la classe minoritaire

- Pas assez de voisins : on ne crée par assez de diversité

- **Problèmes aux bords du support**

- On est limité par les observations les plus « extrêmes »

- **Support non convexe**

- On peut créer des individus aberrants

Des solutions existent :

➔ **Borderline SMOTE**

➔ **Résultats théoriques si**

nbre de voisins / nbre d'observations → 0

➔ **…**

Illustration

- Jeu de données *Telematics :*
  - Variable d'intérêt: sinistralité automobile (occurrence de sinistre , charge de sinistre, fréquence)
  - Caractéristiques: informations conducteurs, véhicule, comportement de conduite
- Protocole expérimental :
  - Jeu d'entrainement équilibré (référence)
  - Plusieurs jeux d'entrainement (train set) déséquilibrés
  - Jeu test représentatif
  - K-fold (5)
  - Prédiction autoML (GBM, RF, GLM, RNA, etc.)

Illustration: Classification

- Prédiction de l'occurrence de sinistre (classification supervisé binaire)

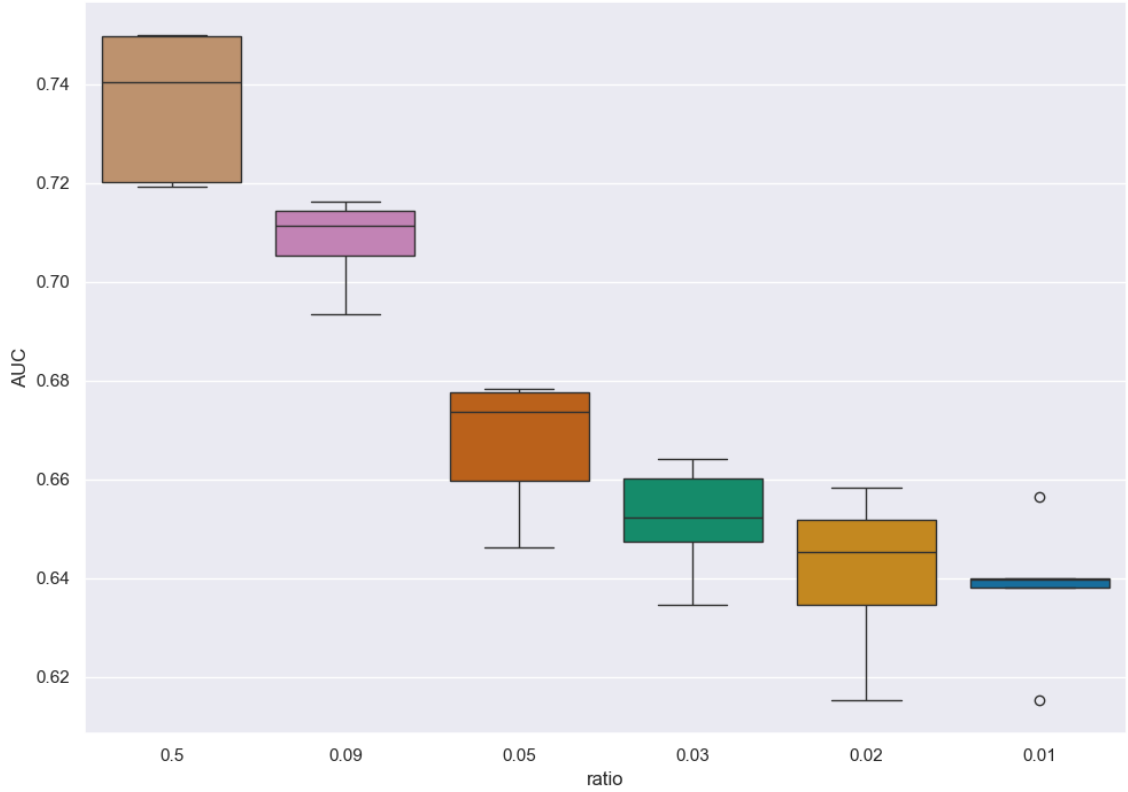


Illustration: Classification

- Prédiction de l'occurrence de sinistre (classification supervisé binaire)

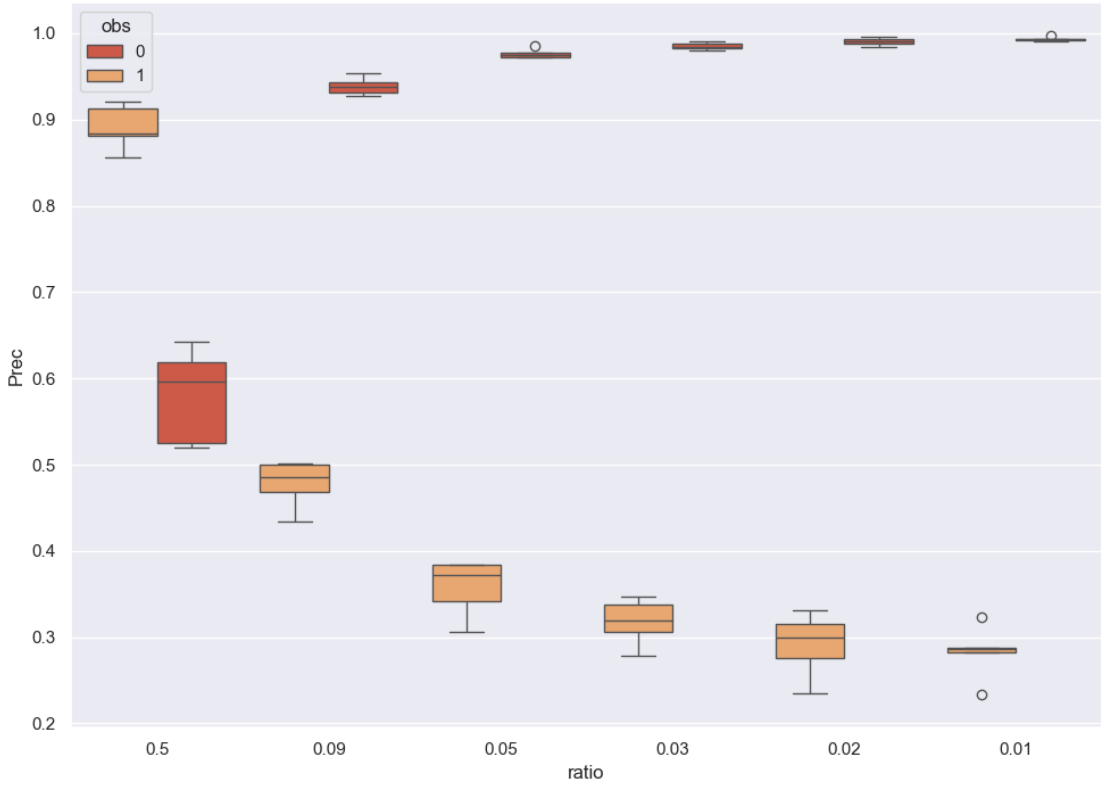


Illustration: Classification

- Prédiction de l'occurrence de sinistre (classification supervisé binaire)

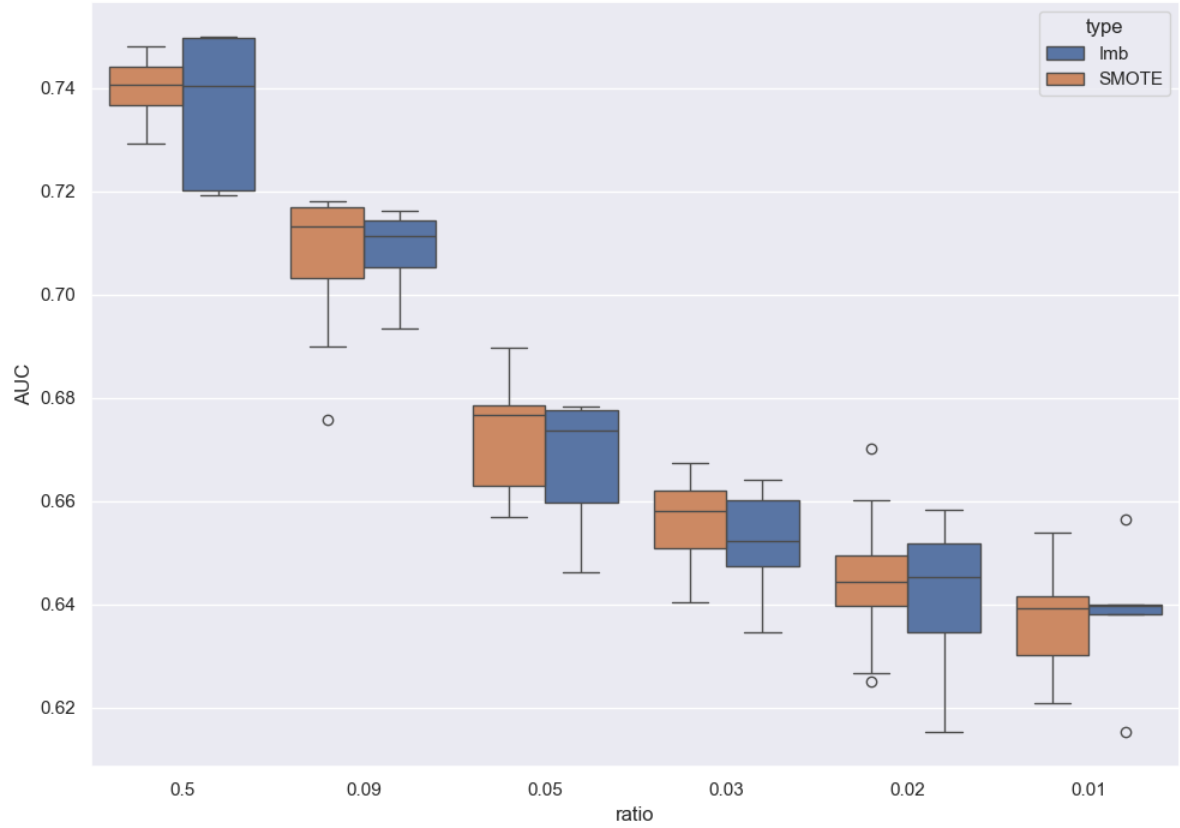
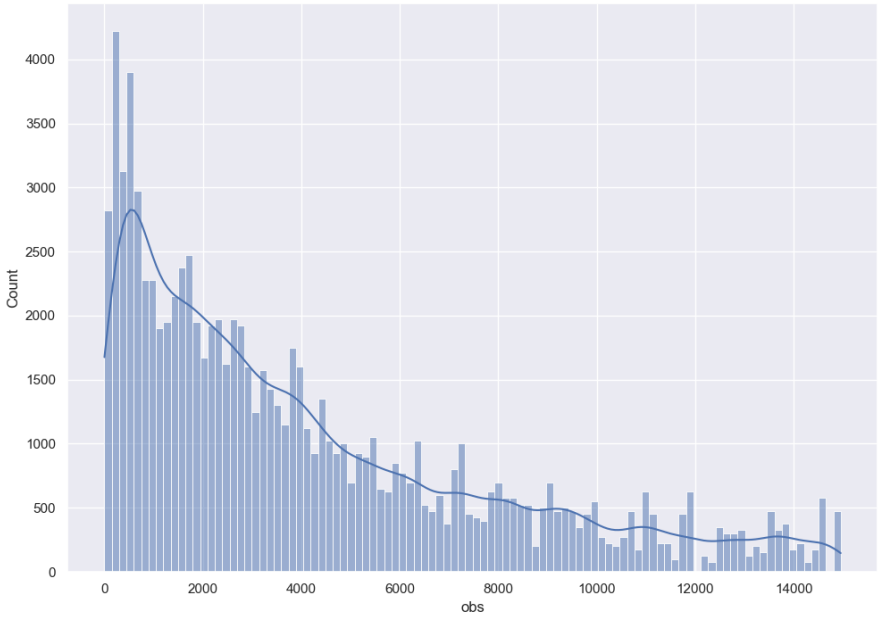


Illustration: et avec la régression ?

- Prédiction de la charge de sinistre



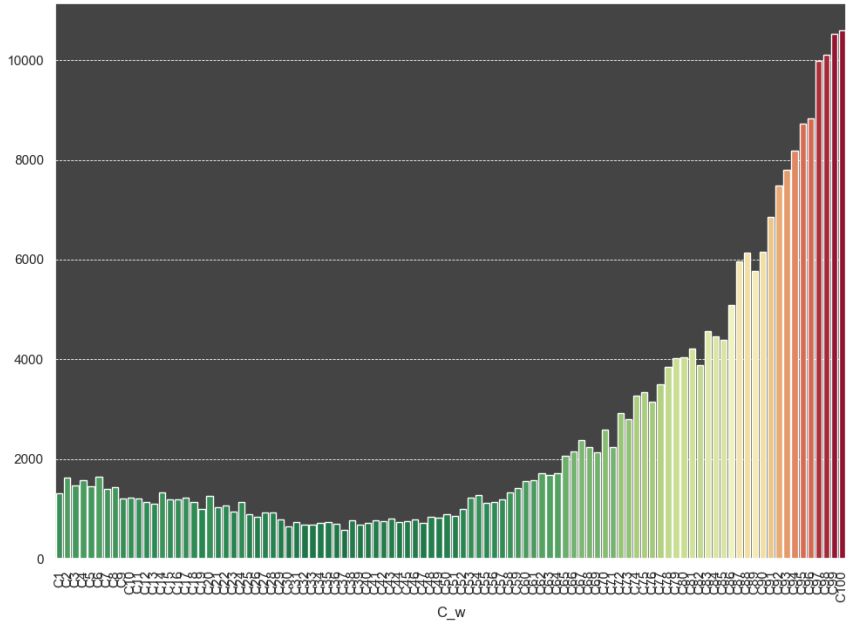


Illustration: et avec la régression ?

- Prédiction de la charge de sinistre

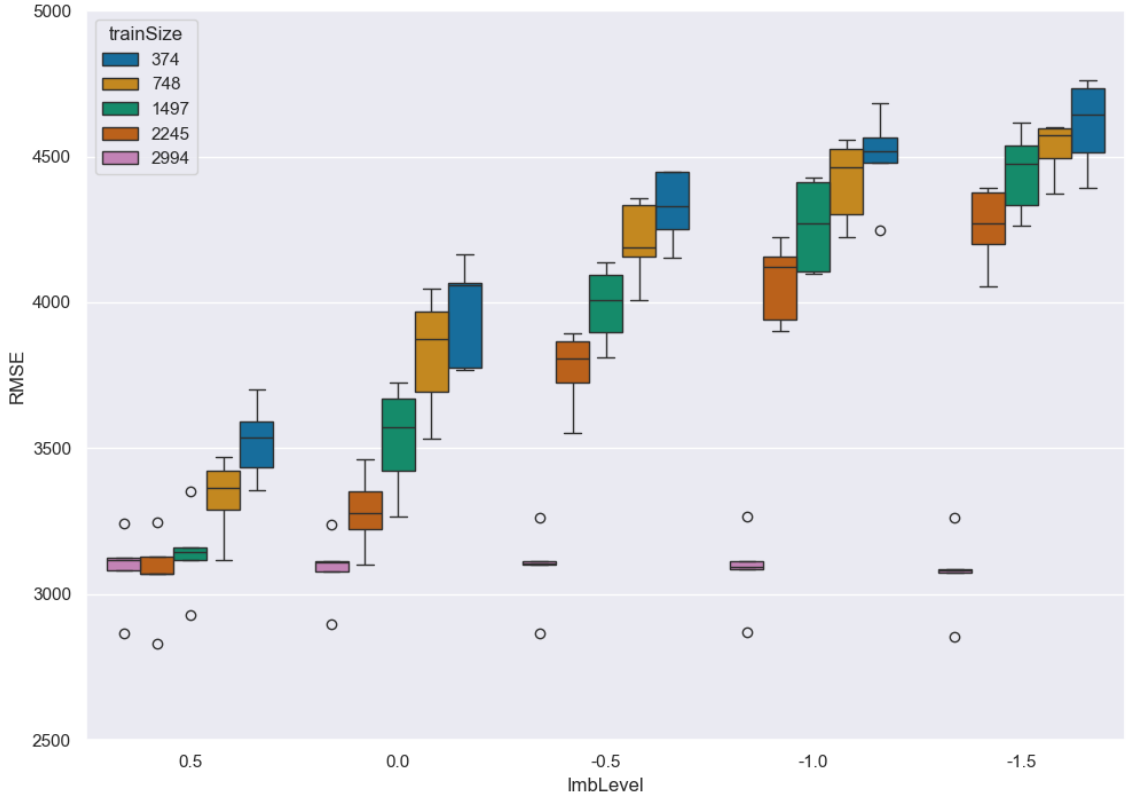
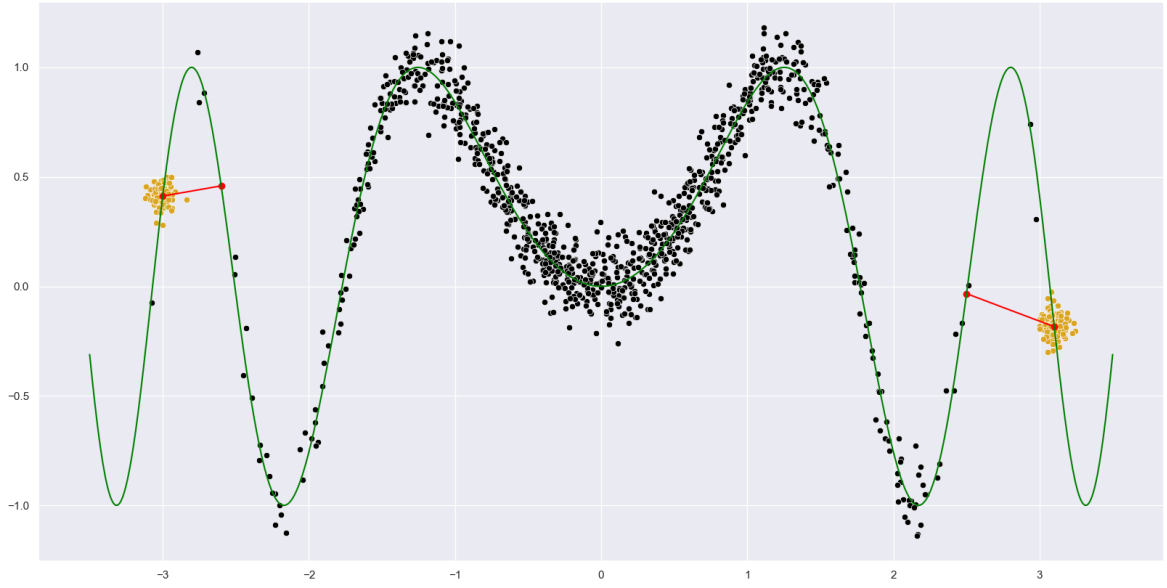


Illustration: et avec la régression ?

- Régression ⇔ variable d'intérêt continue:
  - Définition de valeurs rares ? Lien avec extrêmes ?
  - Mesure du déséquilibre ?
  - Mesure du rééquilibrage ?
  - Génération de la variable d'intérêt ?



Conclusion et perspectives

Conclusions :

- Métriques et méthodes standards inadaptées face au déséquilibre
- Dégradation des performances prédictives (et explicatives)
- Phénomène impactant l'apprentissage supervisée (mais pas que !)
- Nécessité d'introduire des méthodologies spécifiques (prétraitements, métriques, *in-processing*, etc.)

Limites et perspectives :

- Données mixtes
- Prise en compte des corrélations entre les variables (représentativité des données)
- Métriques d'optimisation et de performances à adapter

Merci pour votre attention !

Questions

Remerciements :

Chaire ACTIONS
CARDIF
GROUPE BNP PARIBAS

Institut des ACTUAIRES

INSTITUT Louis Bachelier

DIALog
Digital insurance and long term risk
Chaire d'Excellence

CNP assurances