Digital insurance and long term risk
Chaire d'Excellence

(D)IGITAL (I)NSURANCE (A)ND (LO)N(G)-TERM RISKS

Chaire de Recherche DIALOG - 2020-2025 - CNP Assurances

WebCoffee

Imbalanced Data : Comment le déséquilibre de données peut impacter les performances des modèles ?

19 septembre 2024

La chaire DIALog

Axes

- Axe 1 : Amélioration des process de gestion et de pilotage du risque technique en assurance Vie et Non Vie.
- Axe 2 : Étude de la notion de valeur-client, dans un contexte de transformation digitale du secteur de l'assurance Vie et non Vie.
- Axe 3 : Étude des impacts futurs liés à l'évolution des facteurs environnementaux en assurance Vie et en assurance Non-Vie.
- Thématiques scientifiques associées : IA et risque dynamique, IA et interprétabilité, IA et traitement de données atypiques

Prochain événement: petit-déjeuner du 07/11 ;)

Sommaire

Introduction

Contexte

Quelques éléments contextuels

- **L'importance de la gestion des risques**
  - Fragilité des systèmes $\Longrightarrow$ Réglementation
  - **La modélisation des événements rares et extrèmes**
    Stress-tests : événements climatiques, pandémies, marché financier, risques biométriques, etc.
  - Nouveaux risques : cyber, etc.
- **L'ère de la Data et de l'Apprentissage Statistique**
  - Expansion & évolution du *Machine Learning* (et son appétence)
  - Données accessibles, riches et variées $\equiv$ bases du *Machine Learning*

Cadre et problématique

Les données

Différents types: **structurées**, non structurées, temporelles, textuelles, géospatiales, etc.

L'apprentissage...

Supervisé (classification, **régression**) vs non-supervisé (clustering, etc.) vs auto-supervisé

... à partir de données déséquilibrées

Jeu de données dont la distribution de la variable d'intérêt est "déséquilibrée" ?

Exemples : maladies rares, fraude, résiliation, catastrophes, etc.

Exemple introductif

Imbalanced Classification : Modélisation du taux de résiliation*

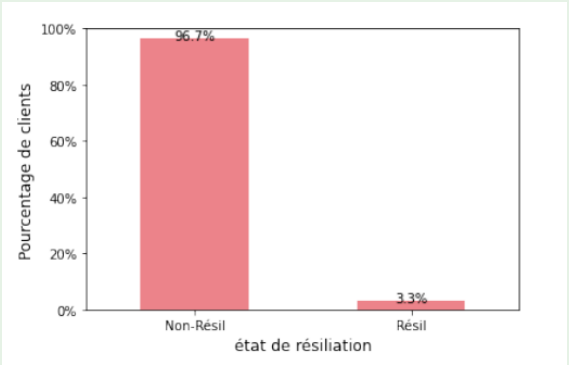


TABLE 1 – Récapitulation résultats sur la base déséquilibrée

Modèle	Accuracy	Précision	Recall	Score-F1	ROC AUC
Régression logistique	0.931	0.106	0.163	0.129	0.618
Arbre de décision	0.804	0.058	0.347	0.100	0.572
Forêt aléatoire	0.699	0.056	0.541	0.101	0.635
Gradient Boosting	0.883	0.076	0.245	0.116	0.620
Machine à vecteurs de support	0.877	0.077	0.265	0.119	0.639

*mémoire de Qingxia WANG

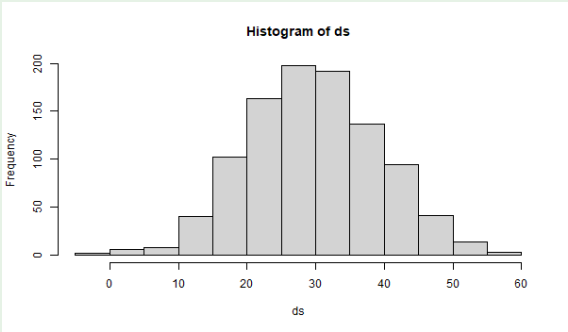
Définition

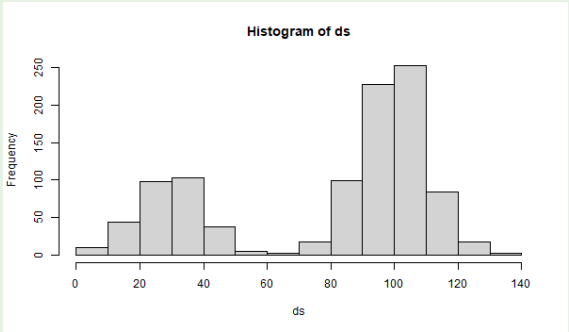
Apprentissage statistique et biais de déséquilibre

- Modèles d'apprentissage standards :
  - Importance uniforme des observations
  - Métriques globales et moyennes
  - Impact échantillon déséquilibré / non-représentatif
- Paradoxe : importance des valeurs rares pour l'étude

Notions de déséquilibre

- Déséquilibre $\Longleftrightarrow$ Rareté ?
- Lien avec la théorie (statistiques) des valeurs extrêmes ?

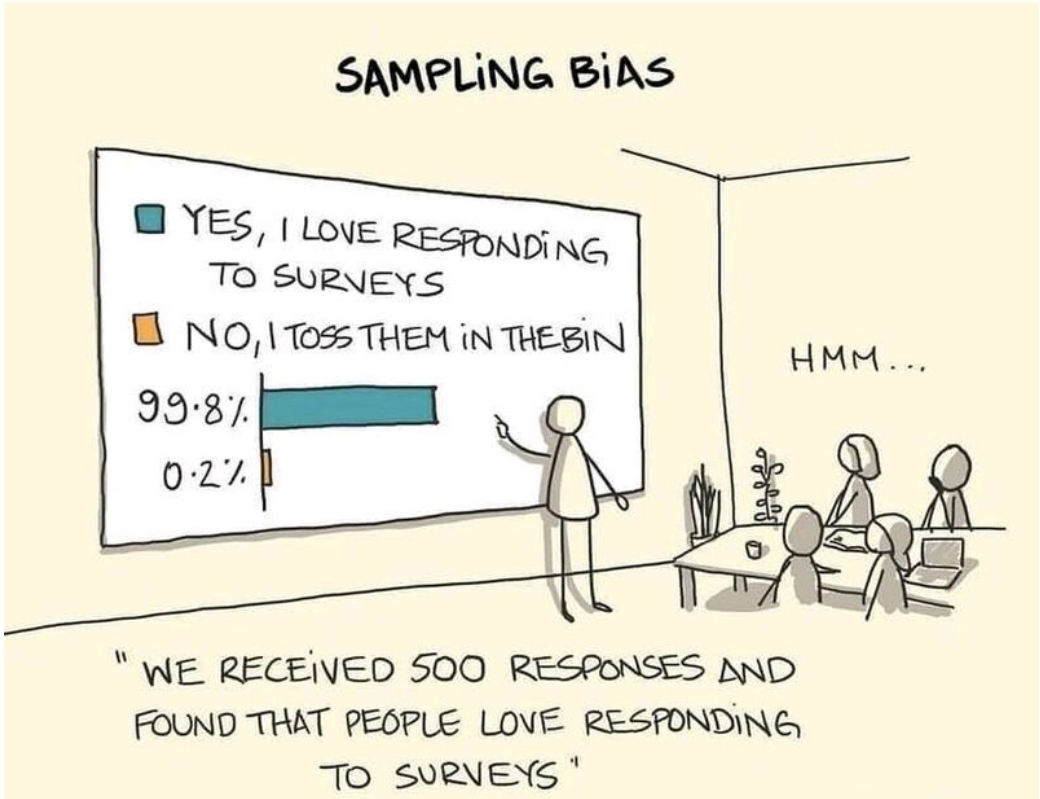




Définition

Source d'un déséquilibre ?

- Intrinsèque ≡ répartition naturelle de la variable
- Extrinsèque ≡ facteur(s) externe(s)
  e.g. biais de sélection $\Longrightarrow$ théorie statistiques des sondages



Définition

Notations

- Jeu de données $\mathbf{x} = (\boldsymbol{x}_j)_{j=1,\cdot,p} = (\boldsymbol{x}_i)_{i=1,\cdot,n} = (x_{ij})_{i=1,\cdot,n;j=1,\cdot,p} \in \mathcal{X}$
  n observations et p variables
- $y = (y_i)_{i=1,\cdot,n} \in \mathcal{Y}$ la variable d'intérêt associée

Illustration fil rouge

Telematics ([12])

- Mesurer et collecte de données de conduite d'une sous-population d'assurés : vitesse, accélération, freinage, virages et distances
- Estimation plus fine des primes d'assurance (comportement de conduite) $\equiv$ asymétrie d'information $\equiv$ personnalisation de primes

Imbalanced Features

Intuition

Toujours Y ! Pourquoi pas X ?

Intuition

L'*Imbalanced Features*

Apprentissage statistique à partir de $\boldsymbol{x}_\ell = (x_{\ell i})_{i=1,\cdot,n} \in \mathcal{X}_\ell \subset \mathcal{X}$ déséquilibrée(s)

$\Longrightarrow$ Comment identifier les valeurs rares ? le déséquilibre ?

$\Longrightarrow$ Comment anticiper cette situation ?

Contextes

- Supervisé (e.g. régression, classification) : covariables
- Non-supervisé (e.g. clustering, réduction de dimension) : caractéristiques
- Auto-supervisé (e.g. autoencoder) : caractéristiques

Illustration

Mesures télématiques

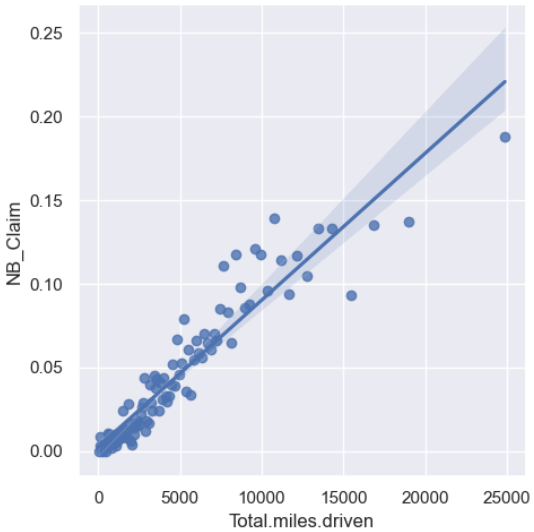
- Boîtiers $\Longrightarrow$ Exclusion des véhicules non conformes (sans prise, trop anciens : véhicule de collection ?)
- Application mobile $\Longrightarrow$ Exclusion des personnes sans smartphone (personnes âgées ?)
- Exclusions naturelles :
  - Assurés "pudiques"
  - Assurés "inattentifs"
  - Conducteurs "imprudents"
  - Conducteurs "grands rouleurs"

$\Longrightarrow$ **Représentativité de la sous-population acceptant le dispositif ?**

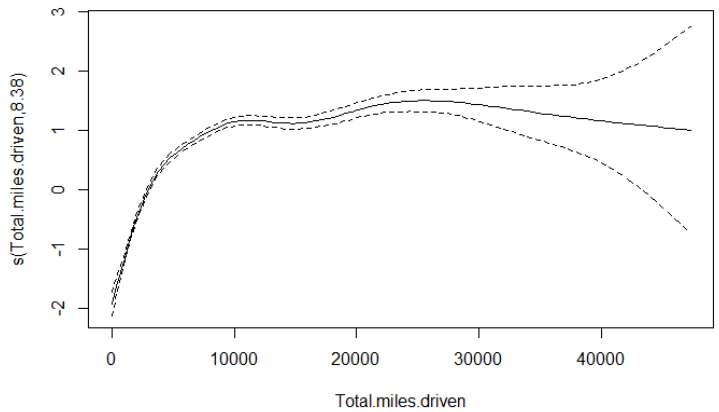
Illustration

Mesures télématiques

- Objectif : modéliser/prédire la fréquence de sinistre
- Fréquence : $\nearrow$ naturellement avec kms $\equiv$ exposition au risque des conducteurs



(a) Nuage de points de kms et freq



(b) Spline de regression de kms (GAM-ZIP)

Illustration

Problématique

Risque : Observer un échantillon avec peu de "grands rouleurs" ($> 10k$ kms) $\longrightarrow$ Extrapoler le lien "kms - fréquence" observé sur les "petits rouleurs" pour estimer la sinistralité des "grands rouleurs"

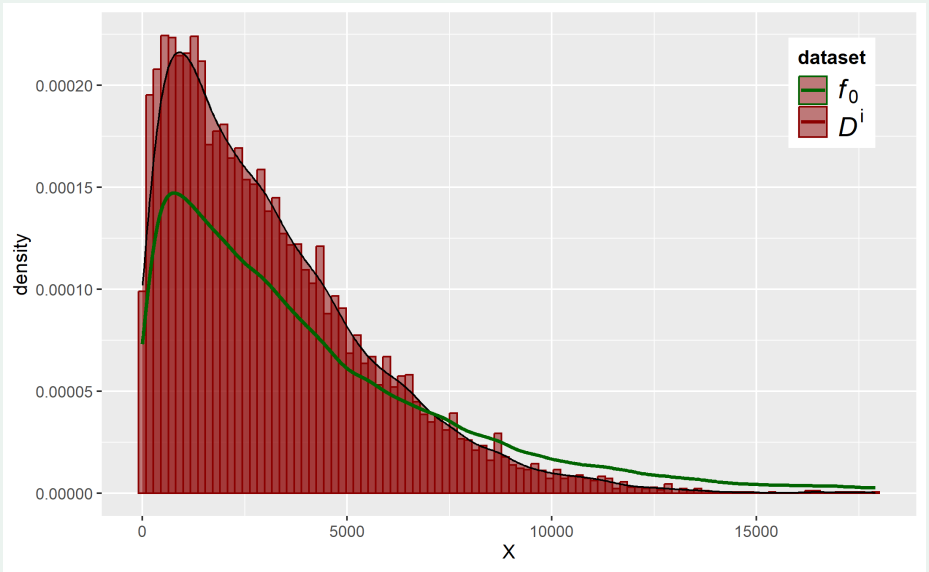


Figure: Histogramme de kms dans l'échantillon ($\widehat{f}$ en rouge) vs la distribution de la population (f_0 en vert)

Contribution

L'*Imbalanced Covariates*

Problème de régression à partir de $(\boldsymbol{x}, y) \in (\mathcal{X}, \mathbb{R})$ avec
$x_\ell = (x_{\ell i})_{i=1,\cdot,n} \in \mathcal{X}_\ell \subset \mathcal{X}$ déséquilibrée(s)

Remarque

Imbalanced Covariate $\neq$ **Imbalanced Regression**

Comparaison des distributions : échantillon vs population (si disponible)

$\Longrightarrow$ Mesure du déséquilibre $\equiv$ le biais d'échantillonnage

$\Longrightarrow$ Redressement vers une distribution cible $\Longleftrightarrow$ théorie de sondage

Contribution

Notations

- Distribution de référence : densité f_0, F.d.R F_0, probabilité $\mathbb{P}_0$
- Distribution observée : densité estimée $\widehat{f}$, probabilité $\widehat{\mathbb{P}}$

Définition formelle

Problème d'**Imbalanced Covariate**, (α, β)-déséquilibré, portant sur une (ou plusieurs) covariable(s) $x_\ell \in \mathcal{X}_\ell$: $\exists \Psi \subset \mathcal{X}_\ell$ tel que :

$$\mathbb{P}_0(x \in \Psi) \geq \beta : \left| \frac{\widehat{\mathbb{P}}(x \in \Psi)}{\mathbb{P}_0(x \in \Psi)} - 1 \right| > \alpha$$

Échantillon significativement différent d'une population de référence pour au moins une partie significative du support de $\mathcal{X}_\ell$

Contribution

Weighted Resampling

Bootstrap pondéré afin de réduire l'écart entre f_0 et $\widehat{f}$:

1. Poids de tirage $\omega_i = \frac{f_0(x_{\ell i})}{\widehat{f}_{e_n}(x_{\ell i})}$

   Avec $\widehat{f}_{e_n} = \max(\widehat{f}, e_n)$ tel que e_n tel que $e_n \underset{n\to\infty}{\to} 0$
2. Normalisation $q_i = \frac{\omega_i}{\sum_j \omega_j}$
3. Construction d'un échantillon $(\mathbf{x}^*, y^*)$: tirage des $(\mathbf{x}_i, y_i)$ selon q_i : $\mathbb{P}(\mathbf{x}^* = x_i) = q_i$

Contribution

Proposition

Distribution observée dans l'échantillon redressé : F.d.R estimée F^*
Sous certaines conditions [a] :

$$\forall x_\ell \in \mathcal{X}_\ell : F^*(x_\ell) \underset{n\to\infty}{\to} F_0(x_\ell)$$

[a] **(C)** dans le chapitre 2; Le support de F contient le support de F_0

Extensions

- Variable qualitative
- Variables quantitatives multivariée (densité estimée multivariée)

Limites

En pratique : Risque de surapprentissage

Contribution

Data Augmentation - Weighted Resampling

Génération de données synthétiques $\Longleftrightarrow$ 1) éviter le phénomène de surapprentissage 2) étendre le support des observations

1. (Weighted Resampling)
2. Augmentation de données
3. Weighted Resampling

Proposition

Sous certaines conditions [a] *:*

$$\forall x_\ell \in \mathcal{X}_\ell : F^*(x_\ell) \underset{n\to\infty}{\to} F_0(x_\ell)$$

[a] **(C), (C'), (C")** dans le chapitre 2; Le support de F contient le support de F_0

Contribution

Générateur de données synthétiques

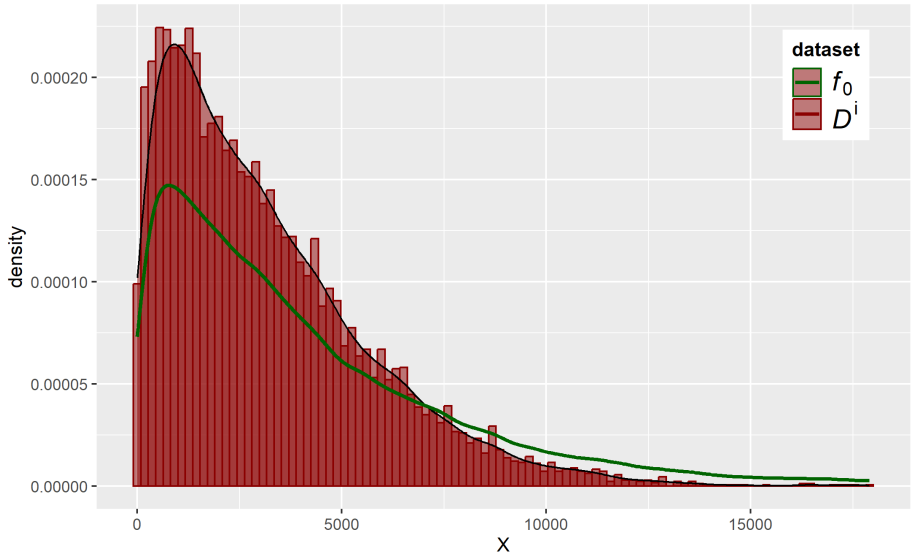
- Approches par perturbation : *Smoothed Bootstrap*, *Gaussian Noise*, *ROSE*, *etc.*
- Approches par interpolation : *SMOTE*, *etc.*
- Modèles à variable latente : *GMM*, *FA*, *ICA*, *etc.*
- Approche par copule
- Approches Deep Learning : *GAN*, *VAE*, *etc.*
- Approches Machine Learning : *Random Forest (Synthpop)*, *etc.*

Génération globale ou par cluster (GMM)

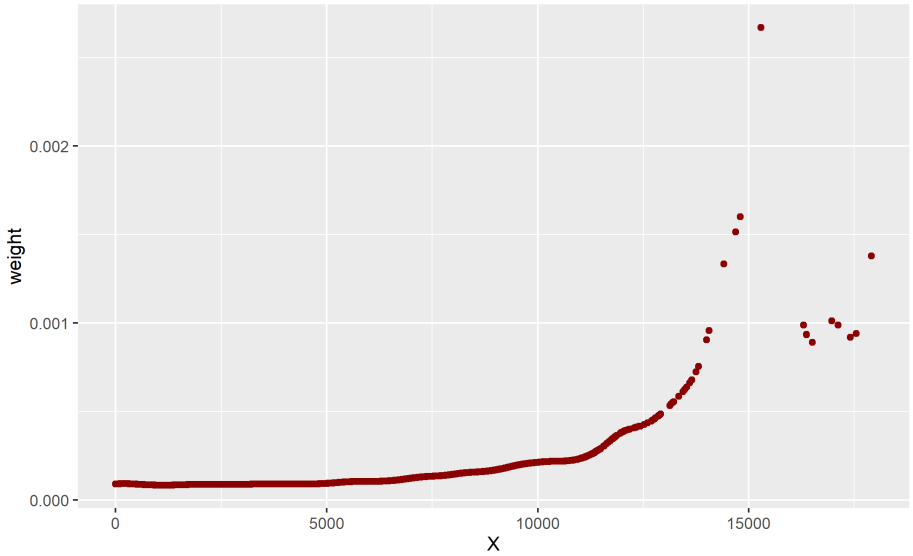
Remarques

Analyse par rapport au problème posé : comparaison des performances et limites, orientation des travaux

Application

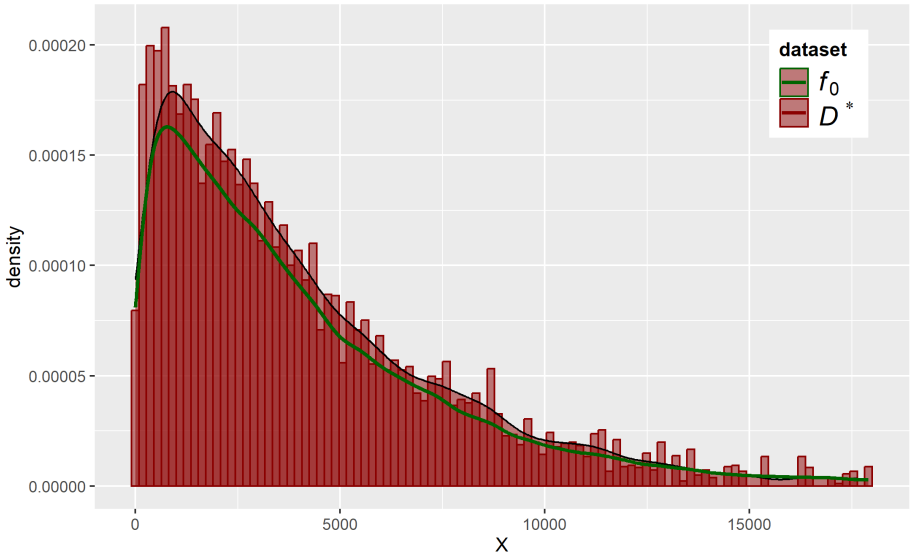


(a) Distribution de kms : observée ($\widehat{f}$ en rouge) vs cible (f_0 en vert)

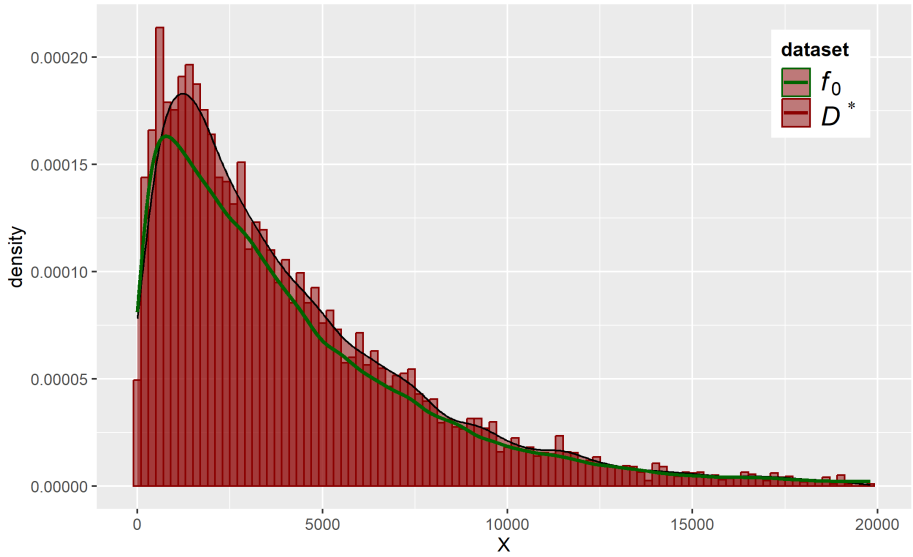


(b) Poids des observations obtenus avec le Weighted Resampling (ω_i)

Application



(a) Weighted Resampling (WR)



(b) Gaussian Noise + WR

Application

RMSE x 100 (% imbalanced sample)			
Dataset\Learning algo.	GAM - Poisson	Random Forest	GAM - ZIPoisson
Balanced Sample	2,0	10,7	2,0
Imbalanced Sample	6,2	14,1	10,9
WR	2,6 (-58,5%)	13,3 (-5,5%)	2,5 (-76,8%)
RF - GMM	2,4 (-61,8%)	13,8 (-2,1%)	2,5 (-77,0%)
GN - GMM	2,4 (-60,6%)	13,7 (-3,0%)	2,5 (-77,1%)
GN	3,7 (-39,9%)	13,0 (-7,9%)	3,7 (-65,6%)
KDE	4,1 (-33,7%)	13,6 (-3,8%)	5,3 (-51,5%)
ROSE	4,4 (-29,5%)	12,6 (-11,0%)	4,4 (-59,2%)

Table: RMSE sur la prédiction du jeu test avec un modèle additif généralisé Poisson, une forêt aléatoire et un modèle additif généralisé - Poisson inflatée en zéro

Discussion

Limites

- WR : Risque de surapprentissage
- DA-WR : Identification et reproduction des corrélations
- DA-WR : Traitement des variables qualitatives
- DA-WR : Convergence dépendante du générateur

Imbalanced Regression

Intuition

Définitions

L'*Imbalanced Regression*

Apprentissage supervisé sur des données (structurées) avec
$y = (y_i)_{i=1,\cdot,n} \in \mathcal{Y} \subset \mathbb{R}^n$ déséquilibrée
Comment mesurer ce déséquilibre ? Comment identifier les valeurs rares ?

Première définition formelle

[11] : Approche *Utility-Based Learning*: $\exists$ une fonction d'utilité
$\phi : \phi(y) : \mathcal{Y} \longrightarrow [0, 1]$ et $\equiv \phi(y)$ n'est pas uniforme sur $\mathcal{Y}$

Identification des valeurs rares

[13] : Binarisation des valeurs de y à l'aide d'un seuil t_r appliqué à $\phi(y)$:

$$D_R : \phi(y) > t_r \text{ et } D_N : \phi(y) \leq t_r$$

$\Longrightarrow$ Adaptation des métriques et solutions de l'*Imbalanced Classification*

Travaux existants

Approche UBL : Comment définir ϕ ? t_r ?

$\Longrightarrow$ Inconvénients : surparamétrisation, binarisation, perte d'information, valeurs rares et extrêmes

Contribution

Définition formelle

Imbalanced $\Longleftrightarrow$ rareté des observations
$\widehat{f}_Y$ un estimateur de la densité de Y observée
Régression $(\alpha, \beta)-$déséquilibrée si $\exists \Psi \subset \mathcal{Y}$ tel que :

$$\int_{\Psi} f_0(y)dy \geq \beta \quad \text{and} \quad \frac{\int_{\Psi} \widehat{f}_Y(y)dy}{\int_{\Psi} f_0(y)dy} < \alpha.$$

Imbalanced Regression $\Longleftrightarrow$ échantillon d'apprentissage significativement différent d'une distribution de référence ou sous-représenté (si $f_0 \sim \mathcal{U}$) pour au moins un sous-ensemble significatif du support de Y
$\Longrightarrow$ Définition englobant les extrêmes et les valeurs rares non extrêmes

Contributions : Réécriture générique des générateurs

Solutions au problème de l'*Imbalanced Learning*

Pre-processing vs in-processing vs post-processing
Générateurs (canoniques) de données synthétiques : interpolation vs perturbation ([2])

Generalized Kernel density Estimate

$$g_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \omega_i K_i(\boldsymbol{x}^*, \boldsymbol{x})$$

avec $(K_i)_{i\in\mathcal{I}}$ une collection de noyaux, $(\omega_i)_{i\in\mathcal{I}}$ une séquence de poids positifs telle que: $\sum_{i\in\mathcal{I}} \omega_i = 1$, et $\mathcal{I}$ représente un sous-ensemble de $\{1, 2, \cdots, n\}$.

Contributions

Réécriture des méthodes par perturbation

$$g_{\boldsymbol{X}^*}^{ROSE}(\boldsymbol{x}^*|\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{H_n}^{ROSE}(\boldsymbol{x}^* - \boldsymbol{x}_i) = \sum_{i=1}^{n} \omega_i \frac{1}{|H_n|^{1/2}} K(H_n^{-1/2}(\mathbf{x}^* - \mathbf{x}_i))$$

$$g_{\boldsymbol{X}^*}^{GN}(\boldsymbol{x}^*|\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{H_n}^{GN}(\boldsymbol{x}^* - \boldsymbol{x}_i) = \sum_{i=1}^{n} \omega_i \frac{1}{|H_n|^{1/2}} K(H_n^{-1/2}(\mathbf{x}^* - \mathbf{x}_i))$$

où $H_n = diag(h_1, ..., h_p)$

Paramétrisation

$\omega_i = \frac{1}{n}$ et $K_i(\boldsymbol{x}^*, \boldsymbol{x}) = K_{H_n}(\boldsymbol{x}^* - \boldsymbol{x}_i)$, i.e. le même noyau Gaussien pour toutes les observations mais différentes matrices de lissage

Contributions

Produit de noyaux

Si $H_n = diag(h_1, ..., h_p)$ alors :

$$g_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i \prod_{j=1}^{p} K_{h_j}(\boldsymbol{x}_j^* - \boldsymbol{x}_{ij})$$

avec $K_{h_j}(u) = (2\pi)^{-1/2} h_j^{-1} e^{-\frac{1}{2h_j{}^2} u^2}$ l'estimateur à noyau Gaussien univarié avec le paramètre de lissage h_j.

$\Longrightarrow$ Inadapté pour les variables discrète, bornés et asymétriques

Contributions

Noyaux non classiques

$$g^{per}_{\boldsymbol{x}^*}(\boldsymbol{x}^*|\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \omega_i \prod_{j=1}^{p} K_{h_j}(\boldsymbol{x}^*_j, \boldsymbol{x}_{ij})$$

avec $K_{h_j}(u, x)$ est un noyau univarié (inversible) adapté à la nature de la $j^{\text{ème}}$ variable et défini à partir de $\boldsymbol{x}$

- Pour une variable discrète : noyau Binomial
- Pour une variable asymétrique positivement (resp. négativement) définie sur $[a, +\infty]$ (resp. $[\infty, b]$) : noyau Gamma
- Pour une variable bornée sur $[a, b]$: noyau Beta & noyau Gaussien tronqué

Contributions

Réécriture des méthodes par interpolation (canoniques)

SMOTE Family

$$g^{int}_{\boldsymbol{x}^*}(\boldsymbol{x}^*|\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \omega_i K_i(\boldsymbol{x}^*, \boldsymbol{x}) = \sum_{i\in\mathcal{I}} \omega_i \sum_{\ell\in\mathcal{J}_i} \pi_{\ell|i}\; g^{int}_{i,\ell}(\boldsymbol{x}^*|\boldsymbol{x})$$

avec $g^{int}_{i,\ell}(\boldsymbol{x}^*|\boldsymbol{x})$ une fonction d'interpolation sur le segment $[\boldsymbol{x}_{i.}, \boldsymbol{x}_{i.}(\ell)]$, $\mathcal{J}_i$ l'ensemble des k-ppv associé à $\boldsymbol{x}_{i.}$ et $\pi_{\ell|i}$ le poids de tirage associé à $\boldsymbol{x}_{i.}(\ell)$

Paramétrisation de SMOTE

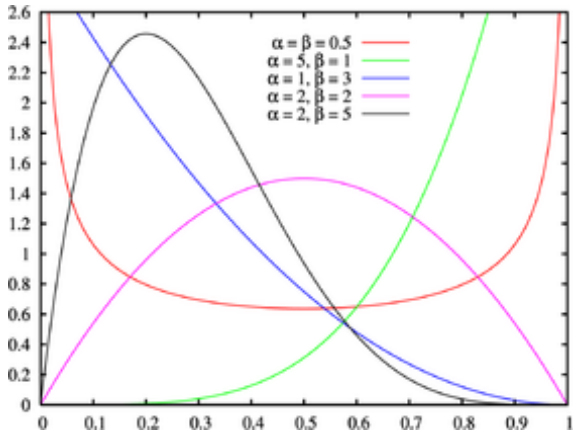
$\mathcal{I} = [0, n]$, $\omega_i = \frac{1}{n}$, $\pi_{\ell|i} = \frac{1}{k}$ et $g^{int}_{i,\ell}(\boldsymbol{x}^*|\boldsymbol{x}) = \frac{\mathbb{1}_{[0;\boldsymbol{x}_{i.}(\ell)-\boldsymbol{x}_{i.}]}(\boldsymbol{x}^*-\boldsymbol{x}_{i.})}{|\boldsymbol{x}_{i.}(\ell)-\boldsymbol{x}_{i.}|}$ distribution uniforme entre les points $\boldsymbol{x}_{i.}$ et $\boldsymbol{x}_{i.}(\ell)$

Densité conditionnelle de SMOTE $\equiv$ mélange de mélanges de distributions uniformes

Contributions

Nearest Neighbors Smoothed Bootstrap

distribution Uniforme $\subset$ distribution Beta



Extended Nearest Neighbors Smoothed Bootstrap

Ajout d'une perturbation Gaussienne : Gaussienne-Beta mélange

$$g^{e-int}(\boldsymbol{x}^{**}|\tilde{\boldsymbol{x}}) = \sum_{i\in\mathcal{I}} \omega_i \sum_{\ell\in\mathcal{J}_i} \pi_{\ell|i}\; g_{i,\ell}^{e-int}(\boldsymbol{x}^{**}|\tilde{\boldsymbol{x}})$$

$$\text{avec } g_{i,\ell}^{e-int}(\boldsymbol{x}^{**}|\tilde{\boldsymbol{x}}) \sim N(\boldsymbol{x}^{*}, \sigma^2) \text{ et } \boldsymbol{x}^{*} \sim g_{i,\ell}^{int}(\boldsymbol{x}^{*}|\tilde{\boldsymbol{x}})$$

Contributions

Propositions pour l'*Imbalanced Regression*

Poids de tirage $\equiv$ inverse du kde de la variable cible Y
$\equiv$ plus une valeur est rare, plus fort sera son poids (non paramétrique)

$$\omega_i := \frac{q_i}{\sum_j q_j}, \text{ with } q_i := \frac{1}{\widehat{f}_y(y_i)^\alpha},$$

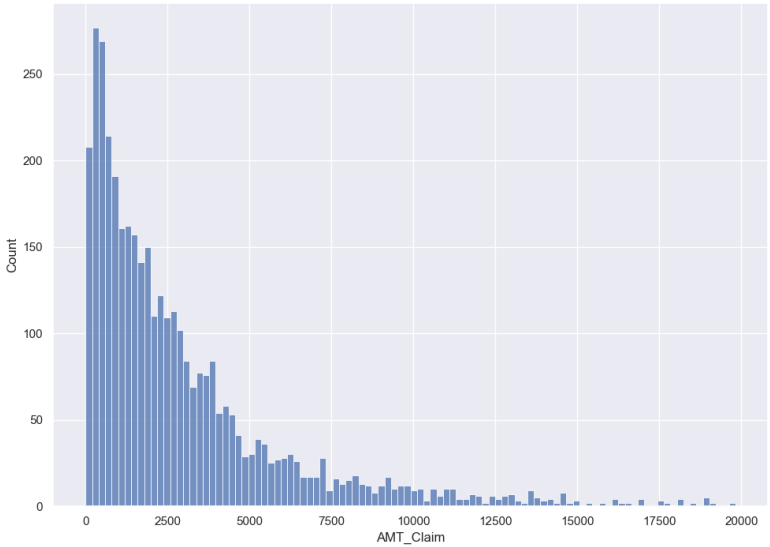
avec $\widehat{f}_y$ un kde et α un hyperparamètre d'échelle

Génération de $Y^*|X^*$

Génération de y^* conditionnellemnt à $\boldsymbol{x}^*$: application d'un Wild-Bootstrap sur la distribution des erreurs de prédiction d'une forêt aléatoire et ajustement par rapport à $\boldsymbol{x}^*$

$$y_s^* := y_s + |\epsilon_k| v_s \times sign(\widehat{y_s} - \widehat{y_s^*})$$

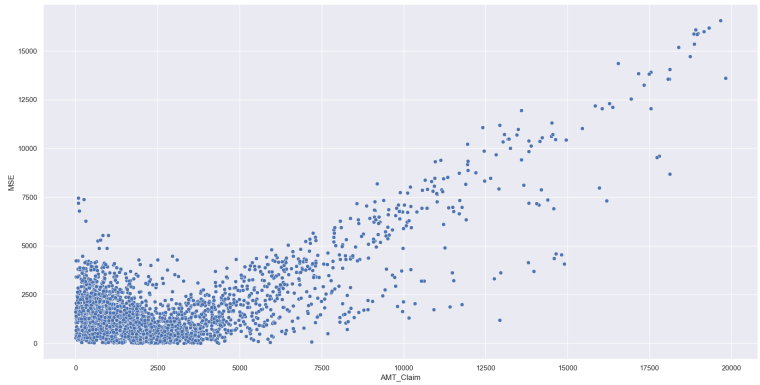
Application



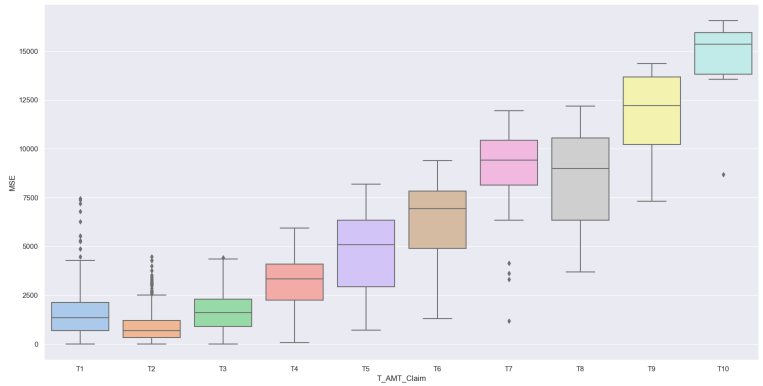
(a) Histogramme de $S \leq 20000$

(b) Boxplot de $S \leq 20000$

Application



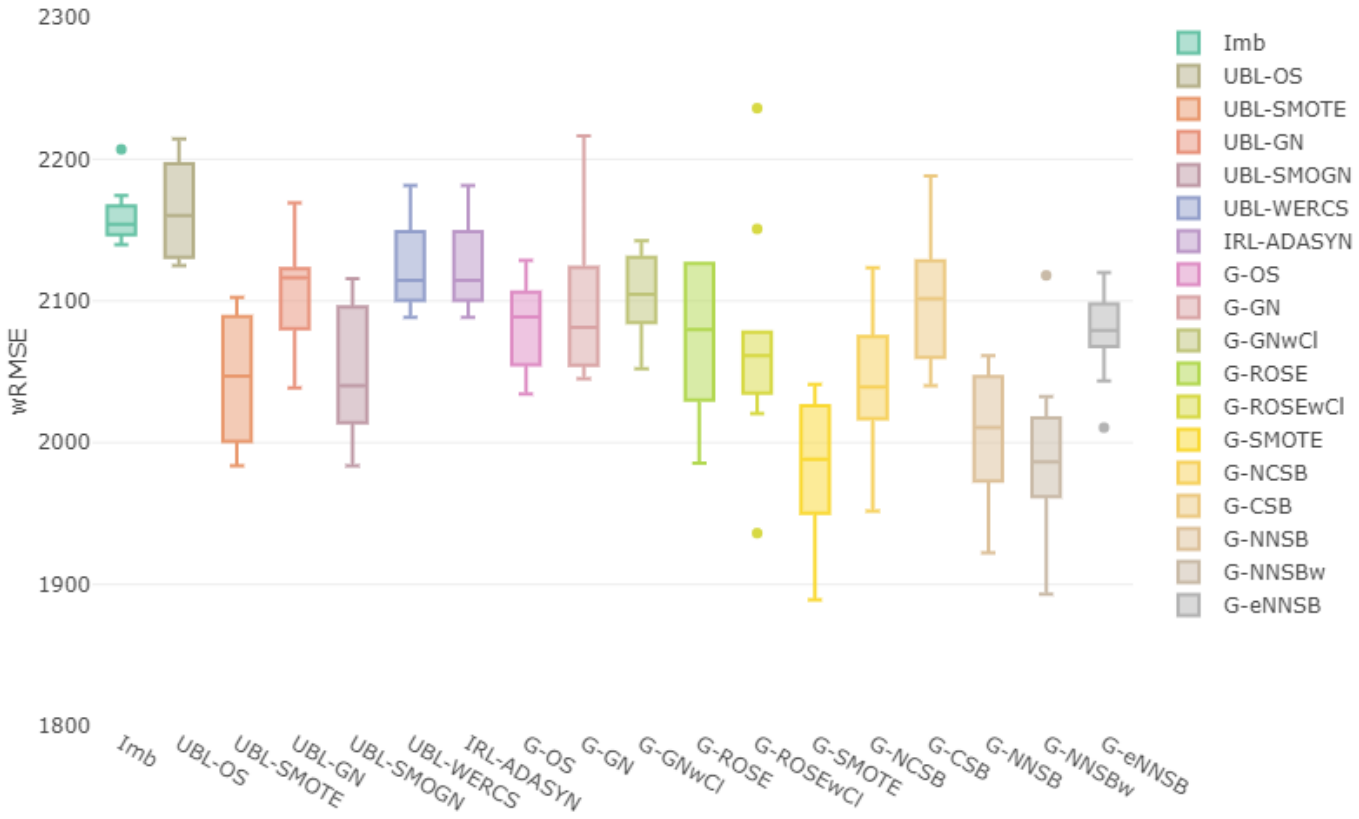
(a) Nuage de points de l'erreur de prédiction



(b) Boxplot de l'erreur de prédiction

Figure: Analyse de l'erreur de prédiction : en fonction de la charge de sinistre observée : par valeur (nuage de points) ou par segment (boxplot)

Application



(a) Boxplots des wRMSE

Dataset	Telematics
wRMSE gain	**GOLIATH**
G-OS	-3%
G-GN	-3%
G-GNwCl	-3%
G-ROSE	-5%
G-ROSEwCl	-4%
G-SMOTE	-8%
G-NCSB	-6%
G-CSB	-3%
G-NNSB	-7%
G-NNSBw	-8%
G-eNNSB	-4%

(b) Médiane des gains de wRMSE

Discussion

Synthèse

i) Définition formelle de l'*Imbalanced Regression*

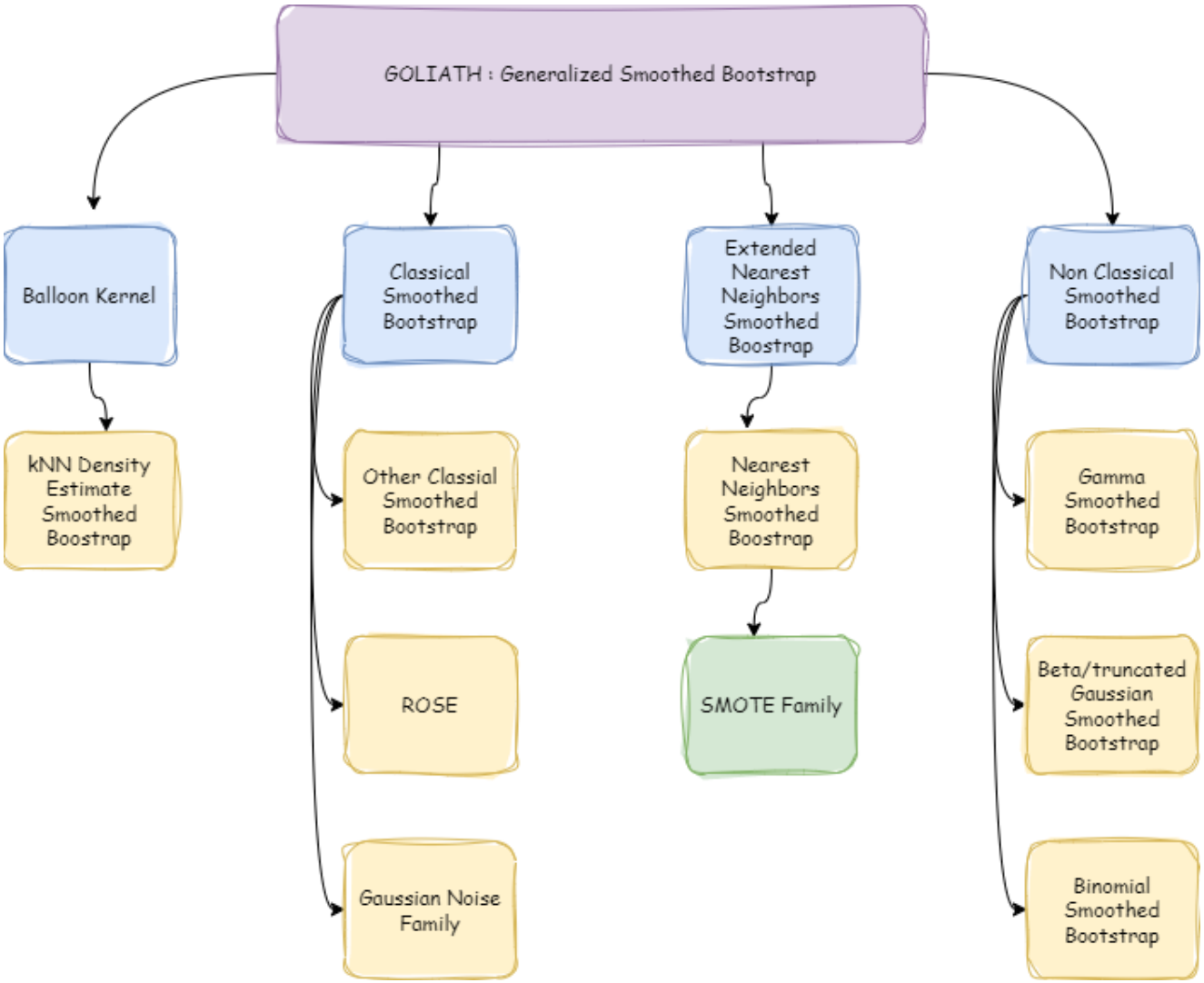
ii) Ecriture statistique globale
- unifiant les familles de génération de données synthétiques
- forme généralisée d'un *Smoothed Bootstrap*

iii) Déduction de nouveaux générateurs
- plus flexibles et mieux adaptées aux données
- améliorant les performances de prédiction (surpassant les compétiteurs)

iv) Méthodologie spécifique au problème de *Imbalanced Regression* :
- Pondération : continue + valeurs et/ou extrêmes
- Génération de $Y^*|X^*$
- Nouvelle manière de générer les données

Discussion

Cartographie de GOLIATH



Discussion

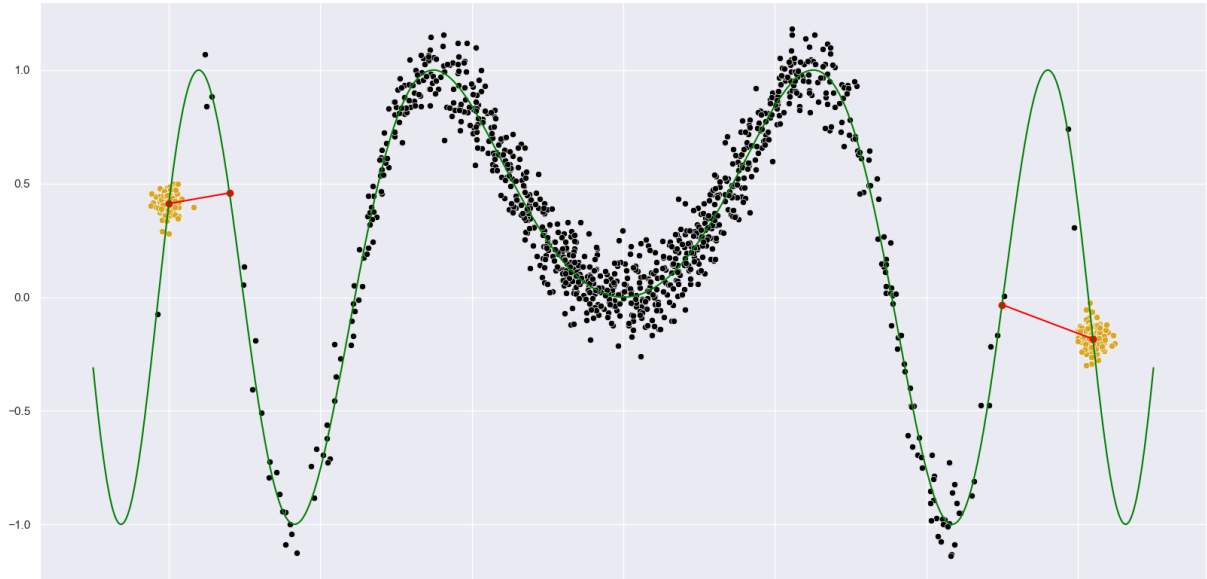
Limites

- Sensibilité à l'aléa
- Performances des nouveaux générateurs dépendantes du jeu de données
- Ne considère pas, nativement, les variables qualitatives
- Ne tient pas ou peu compte des corrélations non linéaires

Illustration

Limites de GOLIATH

- Objectifs
  - Génération de données synthétiques : valeurs rares
  - Préserver les corrélations entre les données
- Limites de GOLIATH
  - Complexité des données : corrélations non linéaires
  - Variables mixtes (qualitatives)



Contribution

Intuition

Changer d'espace de représentation des données, plus propice à la génération

$\Longrightarrow$ Espace latent : corrélations (variables latentes) ; données mixtes

Approches

$\Longrightarrow$ Générer des données, à partir d'un espace latent, tout en cherchant à capter, pour reproduire, les liens non linéaires $\Longrightarrow$ *Variational Autoencoder*

Contribution

Variational Autoencoder

- + : Régularité de l'espace, distributions latentes connues (gaussiennes)
- - : Néglige les valeurs rares : non adapté à l'IR, distributions latentes (im-)posées

Deep Smoothed Bootstrap for Imbalanced Regression

- Fonction de perte spécifique à l'IR
- Génération à partir d'un Smoothed Bootstrap (GOLIATH) dans le nouvel espace de représentation

Contribution

Architecture

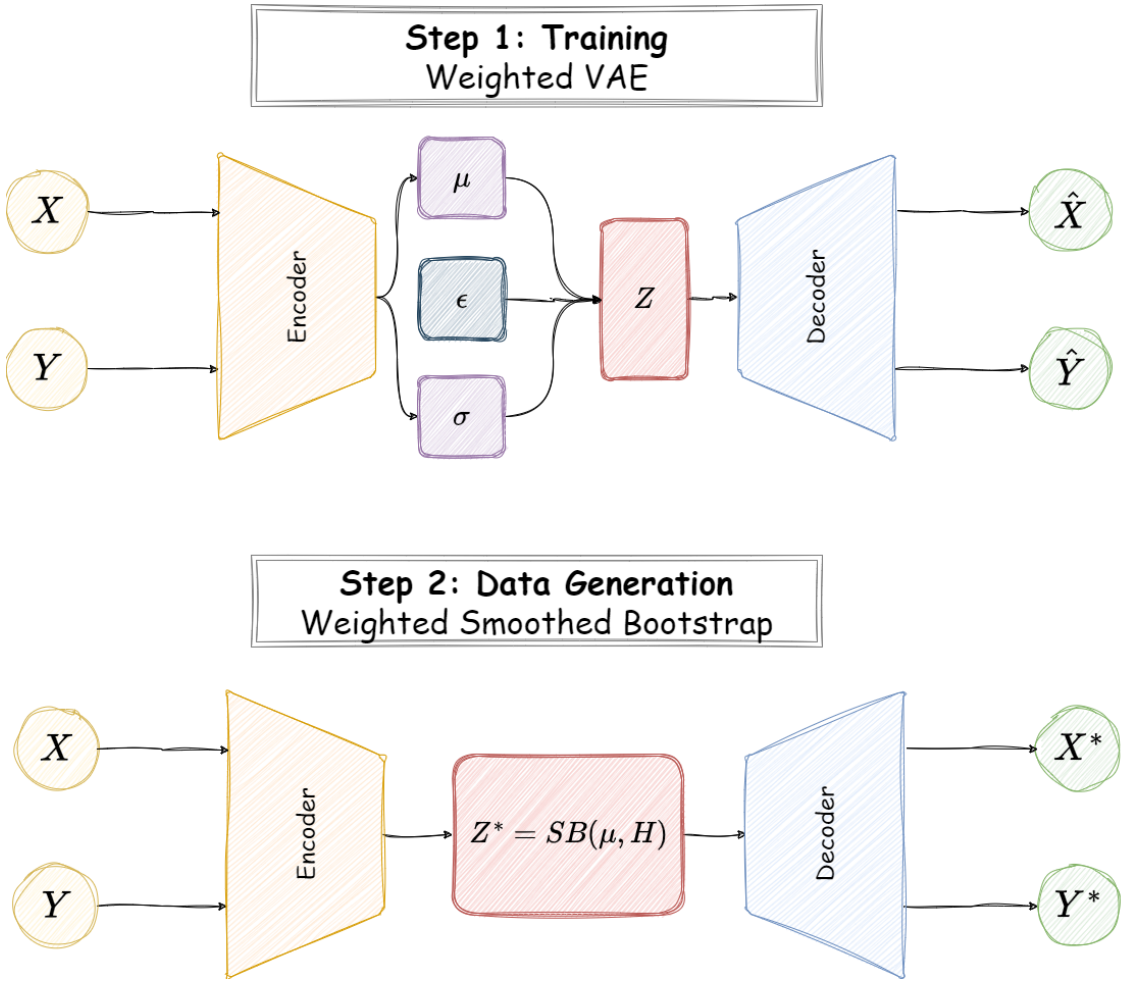


Figure: DAVID Algorithm

Contribution

Entraînement

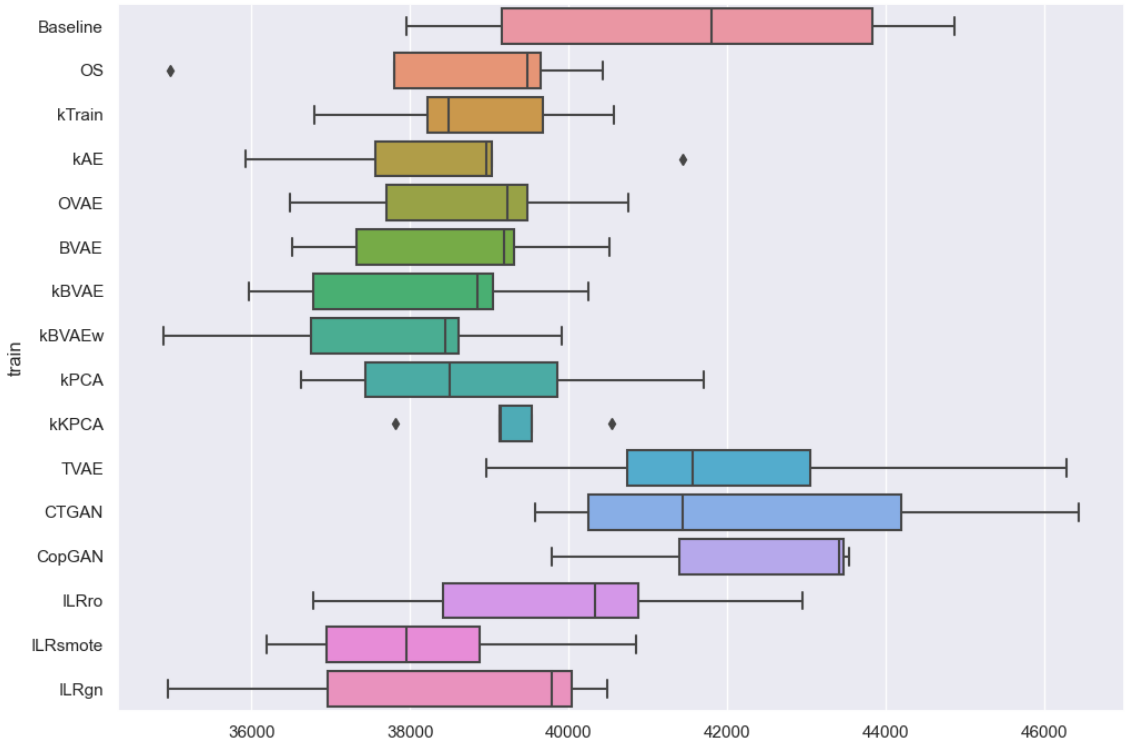
$$\mathcal{L}(\theta, \phi, \boldsymbol{x}, y) = \beta_x \mathbb{E}_q[log\ p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta_{KL} D_{KL}(q_\theta(z|\boldsymbol{x})||p_\theta(\boldsymbol{z})) + \frac{\beta_y}{\widehat{f}(Y)^\alpha} \mathbb{E}_q[log\ p_\theta(y|\boldsymbol{z})]$$

Génération

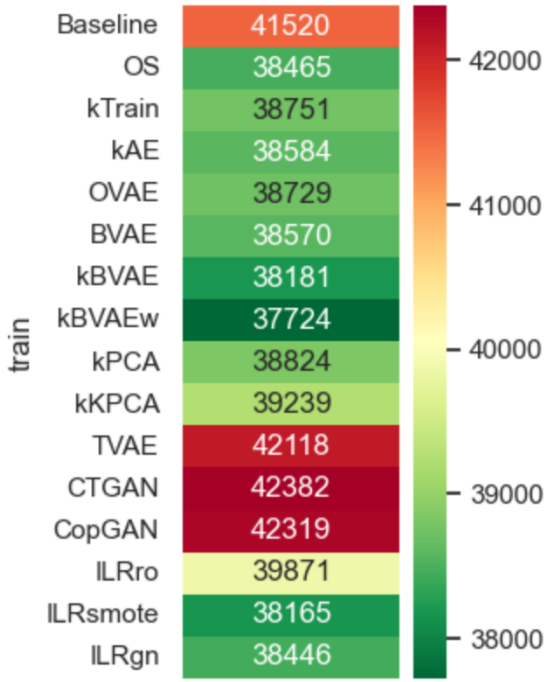
$$g_{z^*}(z^*|\mu) = \sum_{i=1}^{n} \omega_i K_{H_n}(z^* - \mu_i)$$

$$\omega_i := \frac{1}{\hat{f}_Y(y_i)^\alpha}$$

Application



(a) Boxplots des wMSE



(b) Moyenne des wMSE

Discussion

Synthèse

- Potentiel du *Representation Learning* pour données tabulaires
- Nouvelle approche pour la mesure de corrélation non linéaires multivariée
- Nouvelle approche, non paramétrique, pour la construction de *Disentangled Autoencoders*
- Correction des défauts des approches standards et du VAE (en IR)

Limites

- Difficulté de la calibration (VAE/RNA)
- Sensibilité aux paramètres / hyperparamètres / architecture
- Temps de calcul

Conclusion

Conclusion

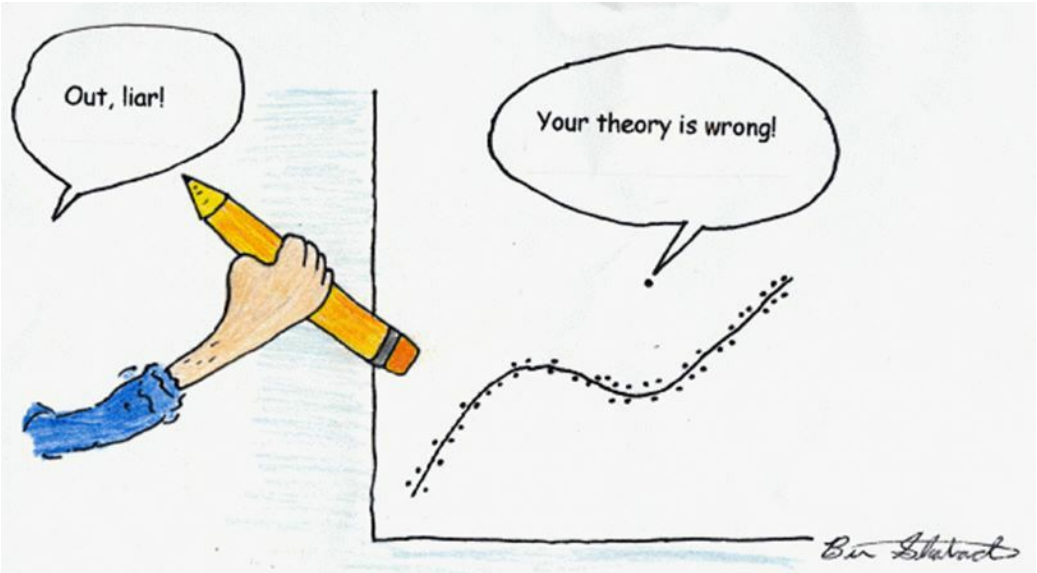
Synthèse

- Impact du déséquilibre sur l'apprentissage, non limité à la variable d'intérêt
- Modélisation des valeurs rares (complémentarité TVE)
- Définitions formelles
- *Imbalanced Y* : pondération continue, mesure du déséquilibre, génération cohérente
- Extension naturelle à la classification, au cadre non supervisé et la génération de données synthétiques

Conclusion

Perspectives & limites

- Données mixtes
- Fonction de pondération
- Apprentissage de représentation latente
- Données synthétique vs observations
- Valeurs rares vs bruit



Digital insurance and long term risk
Chaire d'Excellence

(D)IGITAL (I)NSURANCE (A)ND (LO)N(G)-TERM RISKS

Chaire de Recherche DIALOG - 2020-2025 - CNP Assurances

WebCoffee

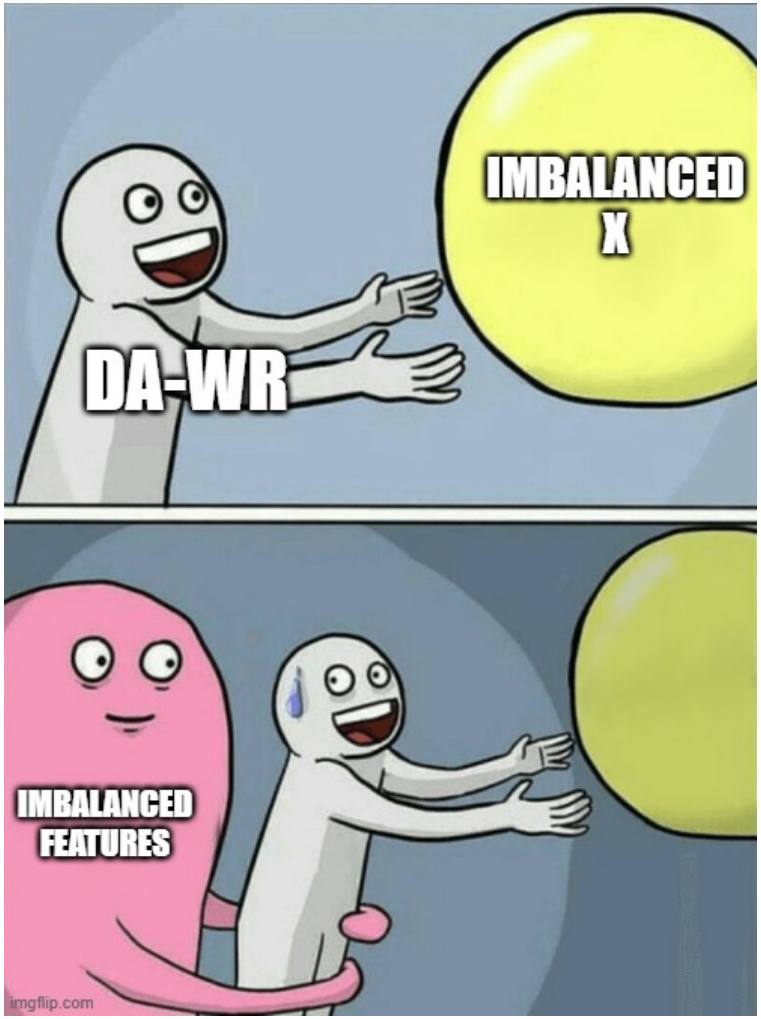
Imbalanced Data : Comment le déséquilibre de données peut impacter les performances des modèles ?

19 septembre 2024

Annexes

Intuition

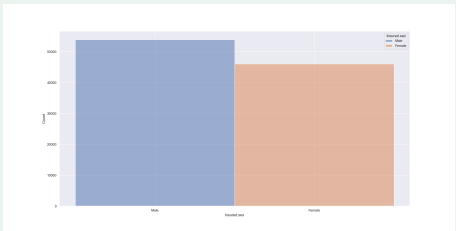
Problème plus large ?



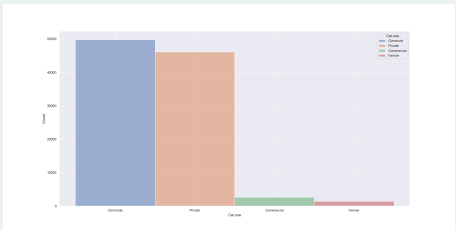
Illustration

Caractéristiques de la sous-population

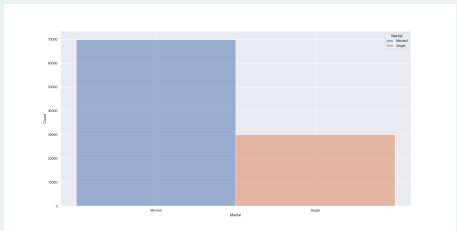
Analyse des distributions des variables qualitatives: "Sexe", "Region", "Statut marital" et "Usage" (deux modalités avec une fréquence d'observation de 2.6% et 1.4%).



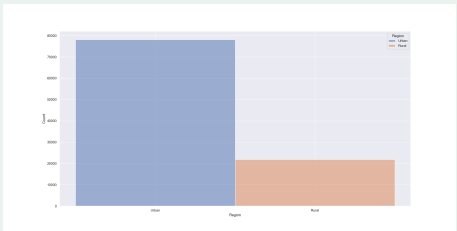
(a) Répartition de "Sexe"



(b) Répartition de "Usage"



(c) Répartition de "Statut marital"



(d) Répartition de 'Region'

Déséquilibre de répartition $\equiv$ déséquilibre d'apprentissage ?

Illustration

Problématique de l'*Imbalanced Learning*

Approches classiques :

- Métrique à optimiser $\equiv$ minimisation d'un critère d'erreur
- Poids des observations : uniformes

$\Longrightarrow$ Un ensemble d'observations "majoritaires" plus influent qu'un ensemble d'observations "minoritaires"

Autoencoder

Approche multi-supervisée / auto-supervisée (*self-supervised*)

- Encodeur : plonge les observations dans un espace latent
- Espace latent : nouvel espace de représentation des données
- Décodeur : reconstruit les observations à partir de l'espace latent

Illustration

Cadre d'application

Cadre d'application: données tabulaires mixtes
Autoencoder sur données mixtes $\Longleftrightarrow$ *One-Hot Encoding (OHE)* + critère d'optimisation standard *Mean Squarred Error (MSE)* multivariée

Intuition

Impact d'un déséquilibre de caractéristiques sur l'apprentissage d'un autoencodeur ? $\Longrightarrow$ sur la construction de l'espace latent associé ?

- Déséquilibre d'influence intra-qualitative : entre les modalités
- Déséquilibre d'influence inter-qualitatives : entre les qualitatives
- Déséquilibre d'influence entre variables qualitatives vs variables quantitatives

Contribution

Intuition

Fonction de perte à minimiser : mesure d'erreur de reconstruction ≡ "distance" entre l'entrée X et la sortie $\widehat{X}$
MSE en multi-supervisé = moyenne arithmétique des MSE univariées :

$$MSE(X, \widehat{X}) = \frac{1}{\pi}\left(\sum_{k \in K_n} MSE(X_k, \widehat{X}_k) + \sum_{q \in Q} \sum_{k_q \in K_q} MSE(X_{k_q}, \widehat{X}_{k_q})\right)$$

$$\text{avec : } MSE(X_{k_q}, \widehat{X}_{k_q}) = f_{k_q}^n MSE(1, \widehat{X}_{k_q}) + (1 - f_{k_q}^n) MSE(0, \widehat{X}_{k_q})$$

$\Longrightarrow MSE(1, \widehat{X}_{k_q})$ proportionnel à $f_{k_q}^n \equiv$ contribution à la MSE totale

Contribution

Problématique

Déséquilibre des modalités

MSE multi-supervisée $\equiv$ minimisation de $MSE(X_{k_q}, \widehat{X}_{k_q})$

Proposition

Pour X_{k_q} variable binaire :

$$MSE(X_{k_q}, \widehat{X}_{k_q}) \equiv 1 - accuracy(X_{k_q}, \widehat{X}_{k_q})$$

$\Longrightarrow$ *minimisation de la MSE binaire $\equiv$ maximisation de la précision*

Contribution

Problématique

Déséquilibre des variables qualitatives

Modalités $\equiv$ variables binaires à part entière $\equiv$ modalités équipondérée $\Longrightarrow$ influence sur la MSE $\nearrow$ #modalités $\Longrightarrow$ déséquilibre de l'apprentissage

Problématique

Déséquilibre des variables mixtes

poids(modalité) = poids(variable numérique)

Contribution

Balanced MSE

$$bMSE := \frac{1}{\pi} \sum_{k \in K_n} \sum_{i=1}^{n} MSE(X_k, \widehat{X}_k) + \frac{1}{\pi} \sum_{q \in Q} \sum_{k_q \in K_q} \left(\frac{1}{2 p_q f_{k_q}^n} \times \sum_{\substack{i=1 \\ x_{ik_q}=1}}^{n} MSE(1, \widehat{X}_{k_q}) + \frac{1}{2 p_q (1 - f_{k_q}^n)} \times \sum_{\substack{i=1 \\ x_{ik_q}=0}}^{n} MSE(0, \widehat{X}_{k_q}) \right)$$

Permet d'équilibrer l'influence :

i) entre les modalités d'une même variable,

ii) entre les variables qualitatives et

iii) entre les variables quantitatives et qualitatives (sous la condition $MSE(X_k, \widehat{X}_k) \leq 1$)

Contribution

Remarques

- Rééquilibrage entre les modalités // Analyse Factorielle de Données Mixtes [9]
- Minimisation de la *bMSE* pour une modalité $\equiv$ maximisation de sa précision équilibrée (*balanced accuracy*) définie par [8] :

$$BalAcc(X_{k_q}, \widehat{X}_{k_q}) := \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

- Déséquilibre de l'influence observée en pratique

Application

Contexte d'application

- Contexte supervisé
  - Classification binaire
  - Classification multi-classe
  - Régression
- Contexte non-supervisé
  - Réduction de dimension
  - Classification non supervisée (*Clustering*)
- Contexte génératif
  - *Variational Autoencoder*

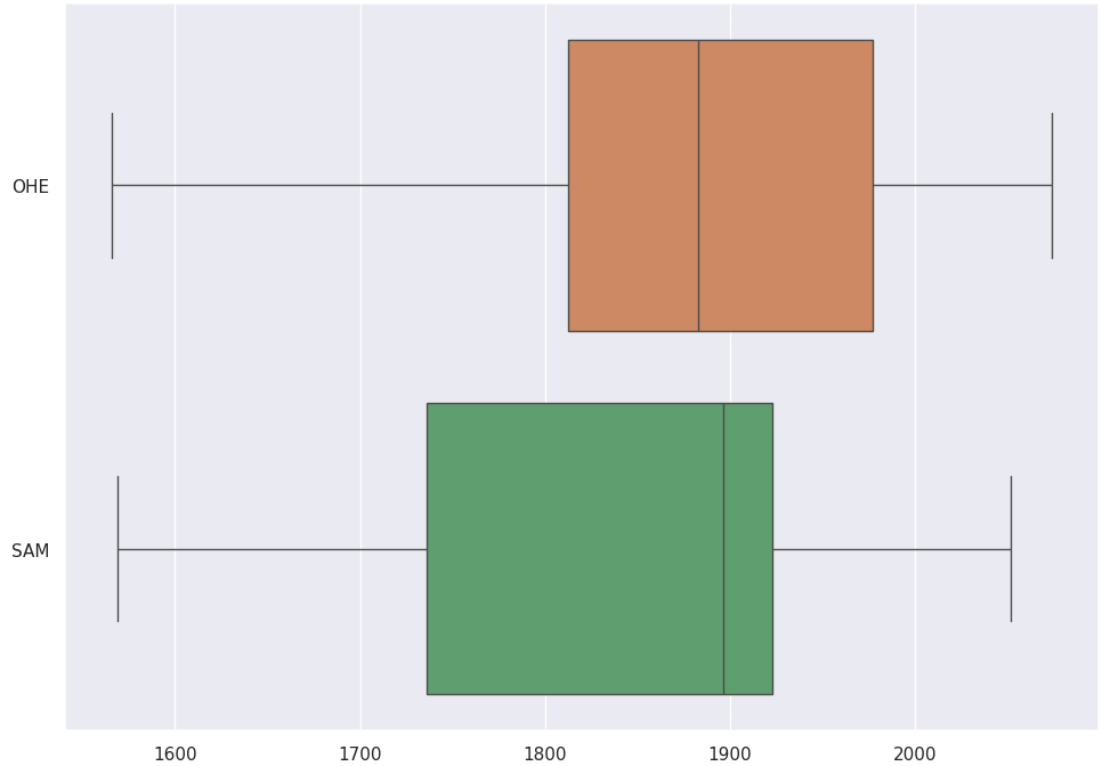
Application

Résultats de l'illustration

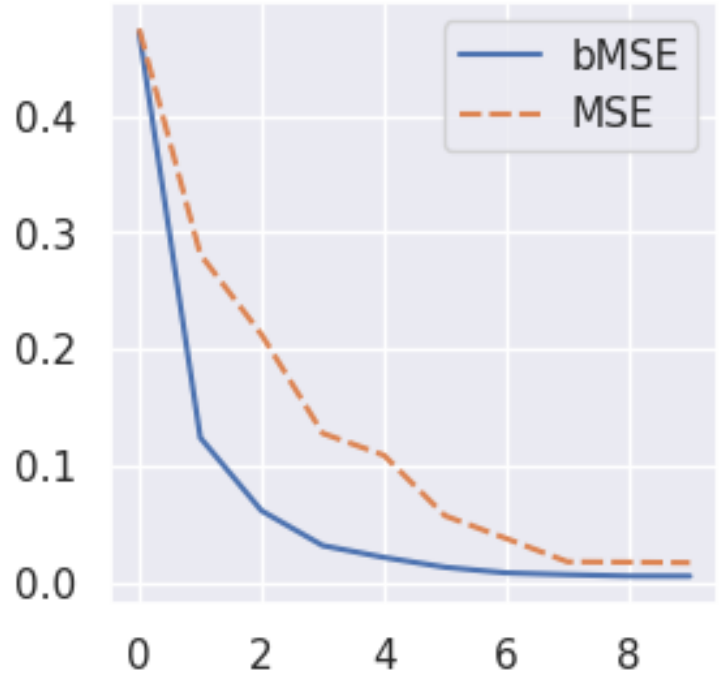
Train	MAE (mean)	MAE (std)	MC (mean)	MC (std)
Baseline	170134.1	10108.7	0.0	0.0
MSE	226013.2	26731.7	25134.4	633.1
bMSE	**206478.3**	14219.0	**24090.9**	534.5

Table: Résultats de prédiction (*autoML*) de Y du jeu test, à partir du jeu reconstruit : 10 "train-test" + analyse de la reconstruction des corrélations mixtes

Application



(a) Erreur de prédiction de la charge de sinistre à partir de l'espace latent d'un autoencodeur. Comparaison de l'approche MSE standard (orange) et de la *Balanced MSE* (vert)



(b) Courbe d'apprentissage (*MSEM*). Comparaison de l'approche MSE standard (orange) et de la *Balanced MSE* (bleu)

Discussion

Limites

- Résultats dépendants de l'architecture du réseau, hyperparamètres, complexité et volumétrie des données
- Convergence si complexité du réseau et nombre d'itération suffisants
- Variable quantitative ?

Définition Imbalanced Classification

Notations

- Jeu de données $\mathbf{x} = (\boldsymbol{x}_j)_{j=1,\cdot,p} = (\boldsymbol{x}_i)_{i=1,\cdot,n} = (x_{ij})_{i=1,\cdot,n;j=1,\cdot,p} \in \mathcal{X}$ n observations et p variables
- $y = (y_i)_{i=1,\cdot,n} \in \mathcal{Y}$ la variable d'intérêt associée

L'*Imbalanced Classification* (binaire)

- Apprentissage supervisé sur des données (structurées)
- variable d'intérêt, binaire, présente un déséquilibre important i.e. est asymétrique

$$Y = f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$$

$$\#\mathcal{D}_R \ll \#\mathcal{D}_N$$

avec : $\mathcal{D}_R = \{(\mathbf{x}_i, y_i) : y_i = 1\}$ et $\mathcal{D}_N = \{(\mathbf{x}_i, y_i) : y_i = 0\}$

Résultats DA-WR

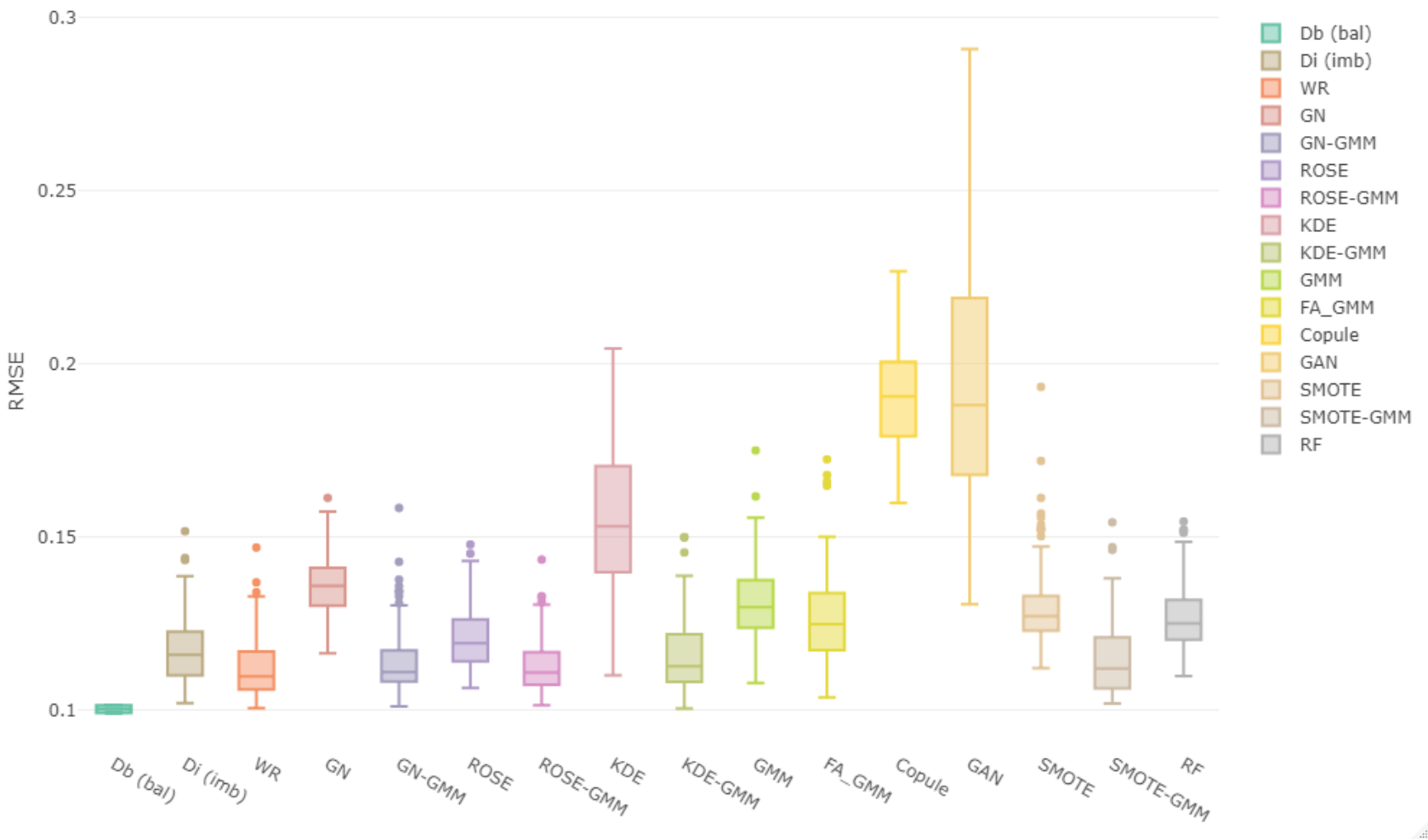


Figure: RMSE boxplots on test datasets prediction with GAM

Résultats DA-WR

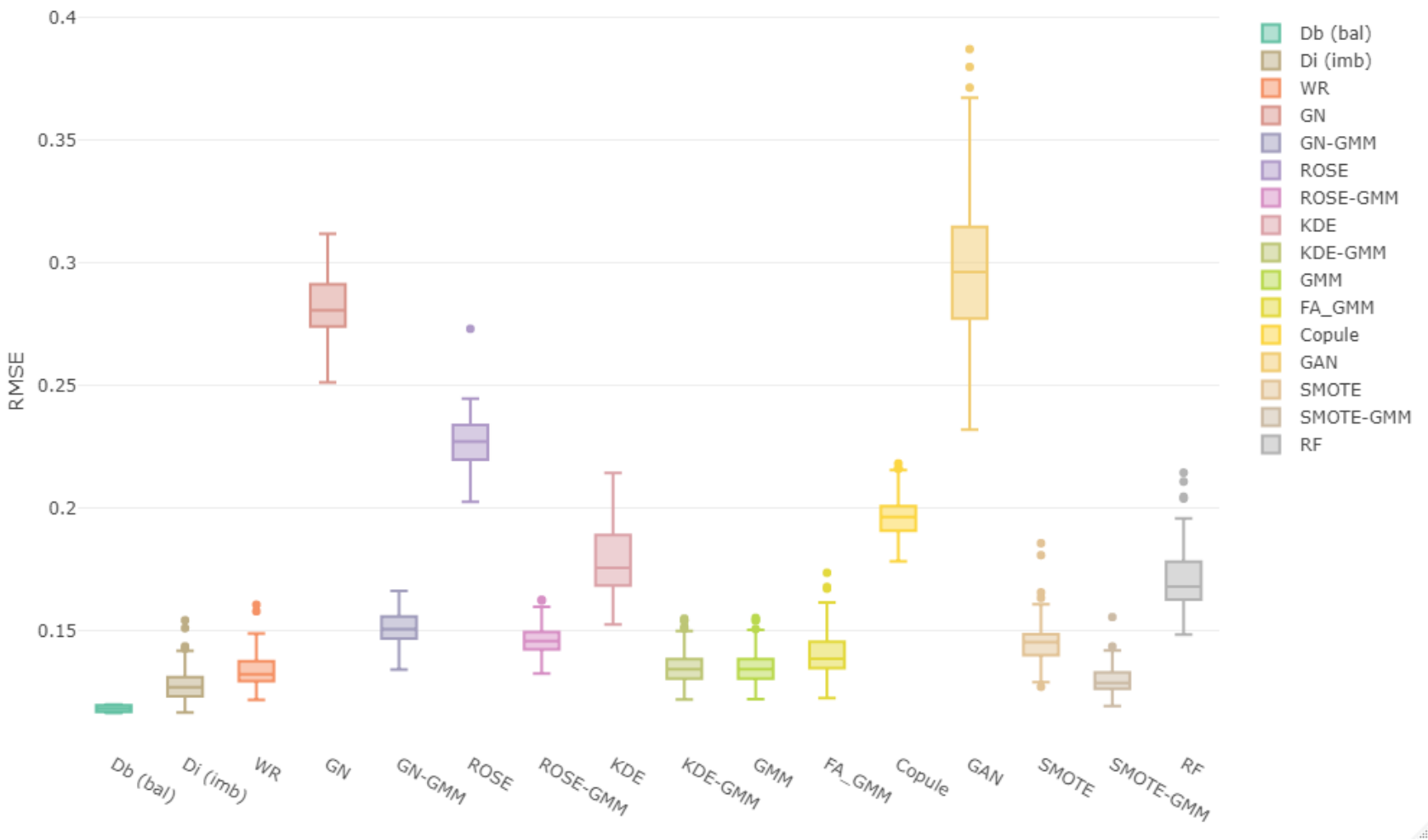


Figure: RMSE boxplots on test datasets prediction with RF

Résultats DA-WR

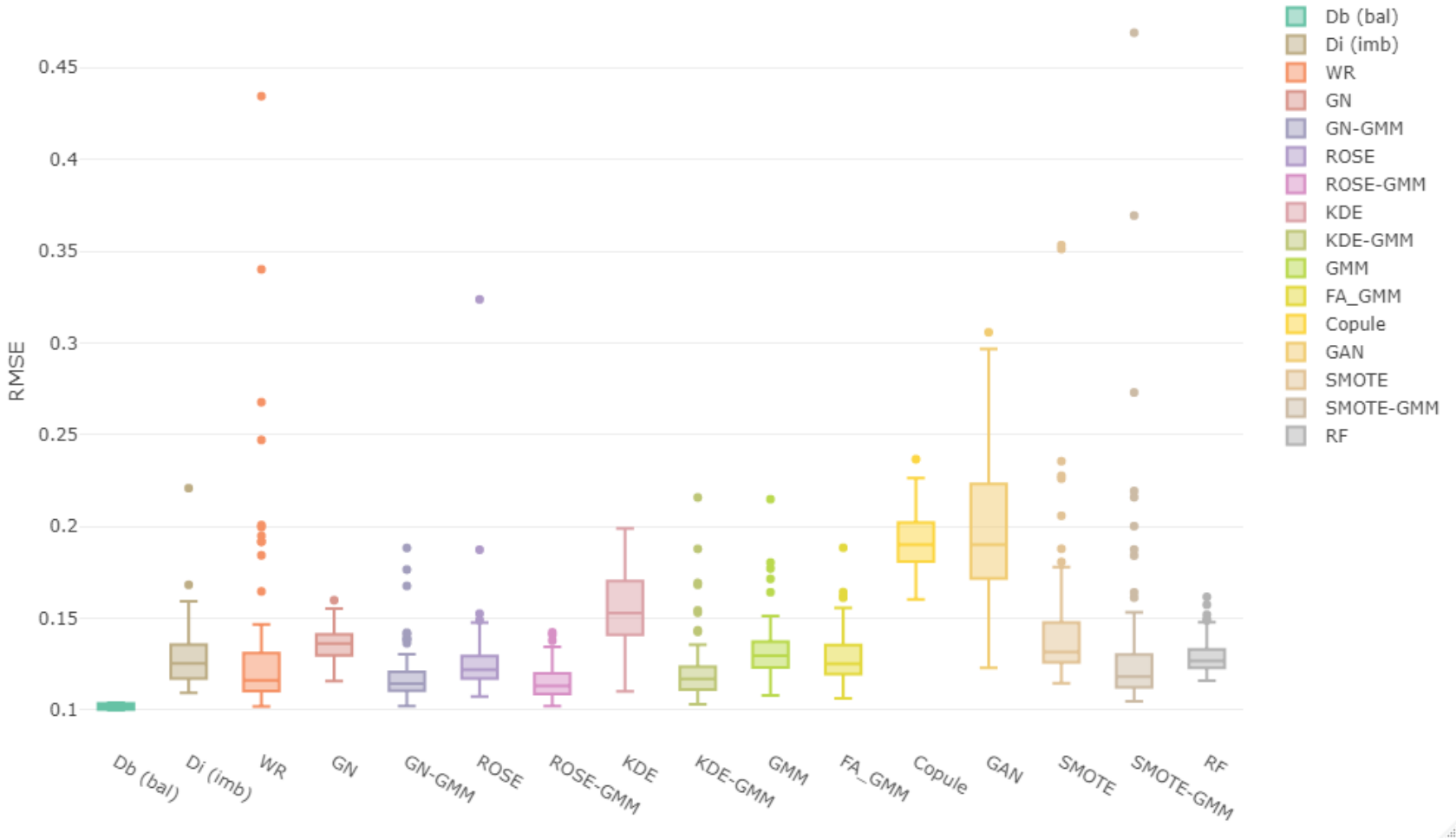
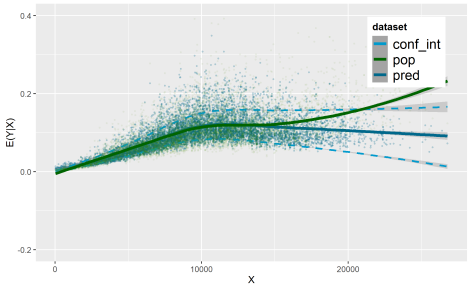
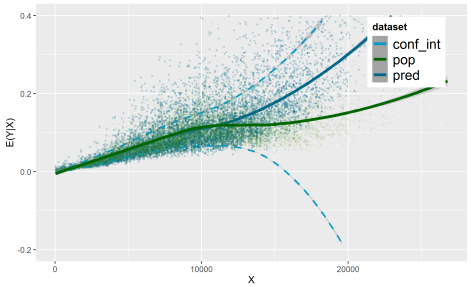


Figure: RMSE boxplots on test datasets prediction with MARS

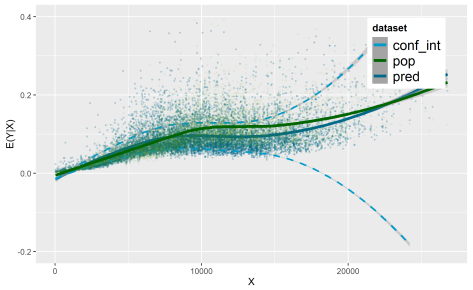
Résultats DA-WR



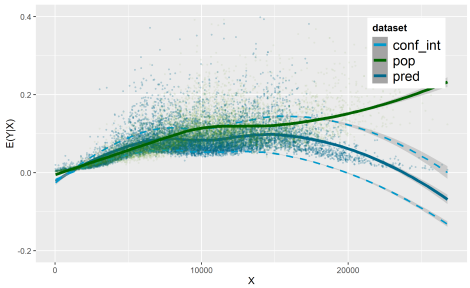
(a) $\mathcal{D}^b$



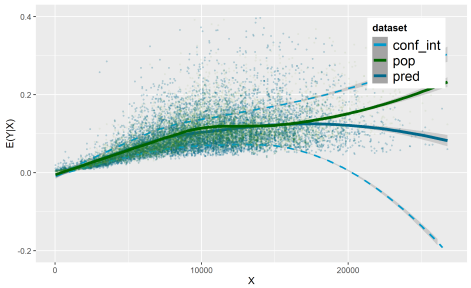
(b) $\mathcal{D}^i$



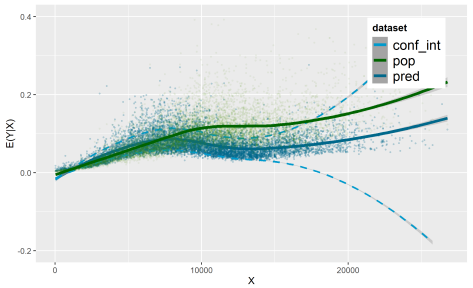
(c) WR



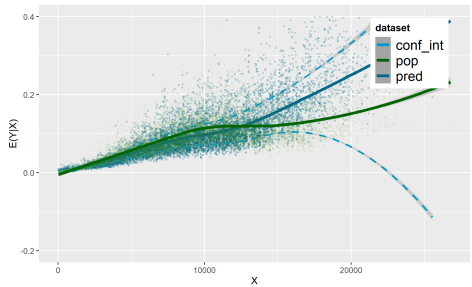
(d) GN



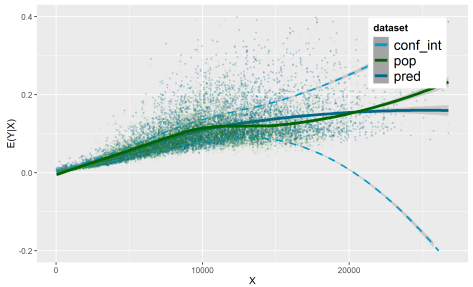
(e) GN - GMM



(f) ROSE



(g) KDE



(h) RF - GMM

Contribution ISS

Notations

- X : dataset constitué de p variables et n observations
- p_{K_c} (resp. p_{K_n}) nombre de variables catégorielles après OHE i.e. variables binaires issues des variables qualitatives (resp. quantitatives)
- K_n l'ensemble des variables numériques
- $\mathcal{Q}$ l'ensemble des variables qualitatives
- Pour une variable $q \in \mathcal{Q}$: K_q le sous-ensemble des variables binaires correspondant à ses modalités
- $f_{k_q}^n$ la proportion observée de la modalité k_q dans l'échantillon
- $\epsilon_{ik} = X_{ik} - \widehat{X}_{ik}$ l'erreur pour la variable quantitative k, et
- $\epsilon_{iq} = X_{iq} - \widehat{X}_{iq}$ l'erreur pour la variable catégorique q

IR : Travaux existants

[11]: Estimation de ϕ à partir des indicateurs de la distribution (boxplot): points de contrôle (point d'inflexion : $\phi'(y) = 0$) + interpolation (cubic Hermite)

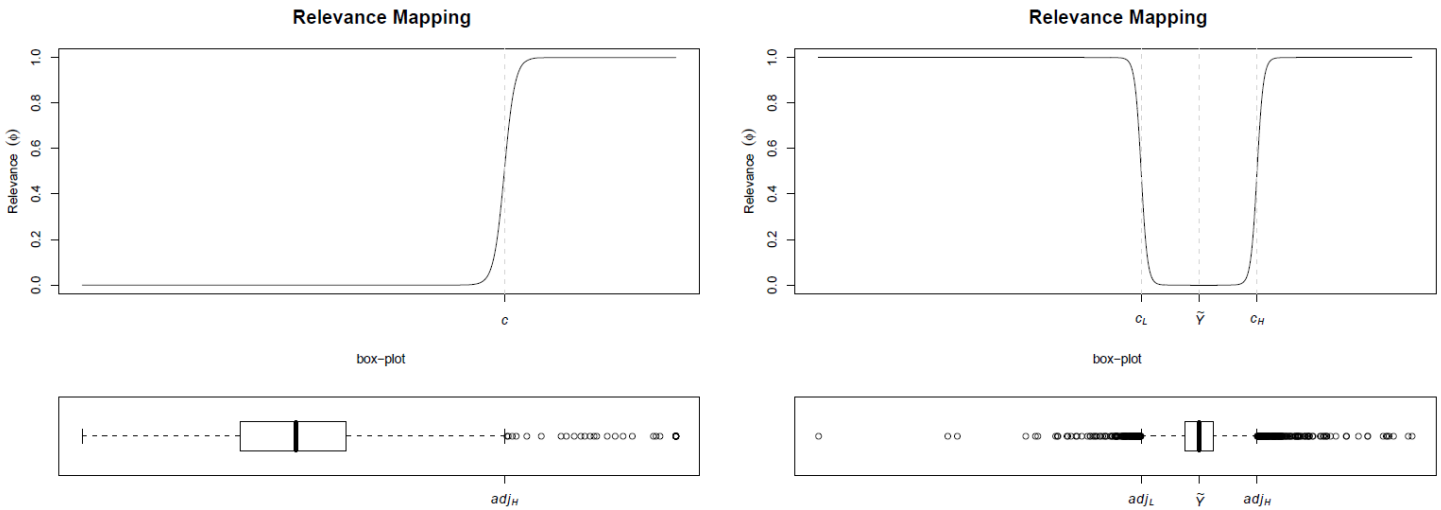
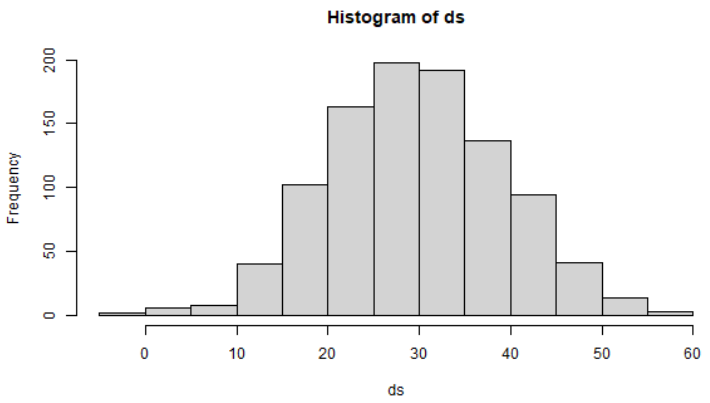
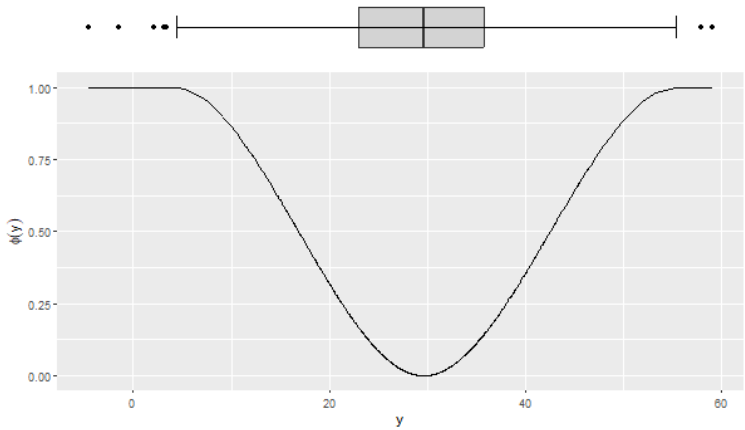


Figure: Deux exemples de fonction d'utilité générée avec l'approche UBL

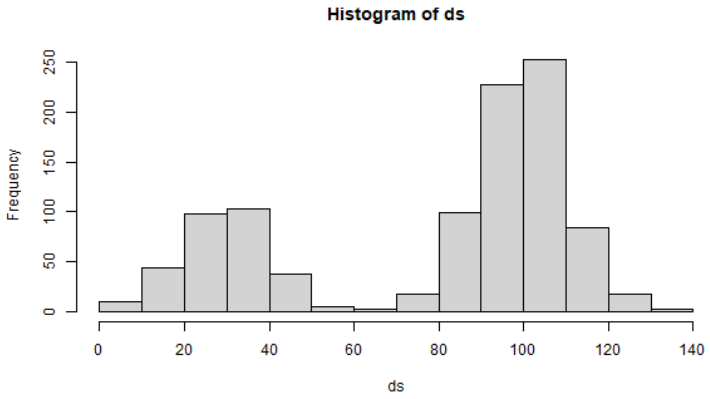
IR : Travaux existants



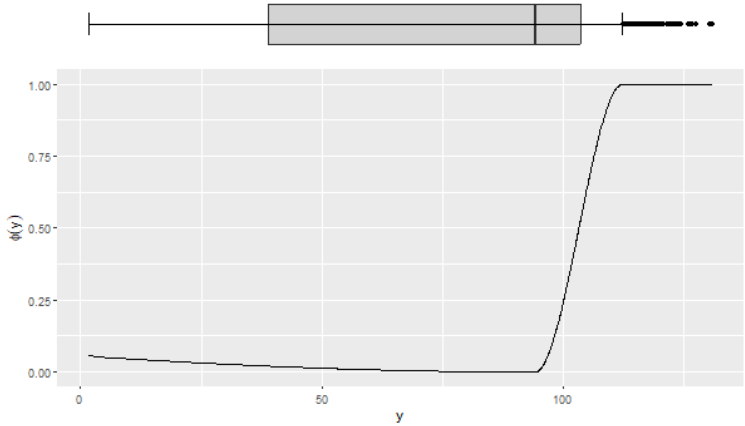
(a) Histogramme d'une distribution unimodal



(b) Fonction ϕ de la distribution unimodal



(c) Histogramme d'une distribution bimodal



(d) Fonction ϕ de la distribution bimodal

Approches par perturbation (canoniques) existantes

Gaussian Noise

Gaussian Noise ([6]) : génération par perturbation gaussienne d'une graine tirée aléatoirement $\boldsymbol{x}_i$:

$$g_{\mathbf{x}^*}^{GN}(\mathbf{x}^*|\mathbf{x}, S = i) = \boldsymbol{x}_i + \epsilon \text{ avec } \epsilon \sim \mathcal{N}(0, \sigma_{noise} \times \widehat{\sigma}_{\boldsymbol{x}})$$

ROSE

ROSE ([7]) : génération par *Smoothed Bootstrap* standard, avec matrice de lissage de [1], à partir d'une graine tirée aléatoirement $\boldsymbol{x}_i$:

$$g_{\mathbf{x}^*}^{ROSE}(\mathbf{x}^*|\mathbf{x}, S = i) = K_{H_n}^{ROSE}(\mathbf{x}^* - \mathbf{x}_i)$$

avec K un noyau gaussien

SMOTE - UMM

$$g_{x^*}^{SMOTE}(x_j^*|x, x_{i.}, x_{i.}(\ell)) = \frac{\mathbb{1}_{[\min(x_{ij}, x_{ij}(\ell)), \max(x_{ij}, x_{ij}(\ell))]}(x_j^*)}{|x_{ij}(\ell) - x_{ij}|} \mathbb{1}_{v_{i\ell}=v_{j\ell}; \forall i,j}.$$

$$g_{x^*}^{SMOTE}(x^*|x, x_{i.}) = \frac{1}{k} \sum_{\ell=1}^{k} \frac{\mathbb{1}_{[\min(x_{ij}, x_{ij}(\ell)), \max(x_{ij}, x_{ij}(\ell))]}(x_j^*)}{|x_{ij}(\ell) - x_{ij}|} \mathbb{1}_{v_{i\ell}=v_{j\ell}; \forall i,j}.$$

$$\begin{aligned} g_{x^*}^{SMOTE}(x^*|x) &= \frac{1}{n} \sum_{i=1}^{n} g_{x^*}^{SMOTE}(x^*|x, x_{i.}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{\ell=1}^{k} \frac{\mathbb{1}_{[\min(x_{ij}, x_{ij}(\ell)), \max(x_{ij}, x_{ij}(\ell))]}(x_j^*)}{|x_{ij}(\ell) - x_{ij}|} \mathbb{1}_{v_{i\ell}=v_{j\ell}; \forall i,j} \\ &= \frac{1}{n} \sum_{i=1}^{n} K_i^{SMOTE}(x^*, x) \end{aligned}$$

Noyaux non classiques

Noyaux non classiques

$$g^{per}_{\boldsymbol{X}^*}(\boldsymbol{x}^*|\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \omega_i \prod_{j=1}^{p} K_{h_j}(x^*_j, x_{ij})$$

avec $K_{h_j}(u,x)$ est un noyau univarié (inversible) adapté à la nature de la $j^{\text{ème}}$ variable et défini à partir de $\boldsymbol{x}$

Pour une variable discrète :

Noyau Binomial

$$K_h(u,x) = \frac{(x+1)!}{u!(x+1-u)!}\left(\frac{x+h}{x+1}\right)^u\left(\frac{1-h}{x+1}\right)^{x+1-u}$$

Noyaux non classiques

Pour une variable asymétrique positivement (resp. négativement) définie sur $[a, +\infty]$ (resp. $[\infty, b]$)

Noyau Gamma

$$K_h(u,x) = \frac{u^{(x-a)/h}}{\Gamma(1+(x-a)/h)h^{1+(x-a)/h}} \exp\left(\frac{-u}{h}\right)\mathbb{1}_{[a,+\infty]}(u)$$

$$K_h(u,x) = \frac{u^{-(x-b)/h}}{\Gamma(1-(x-b)/h)h^{1-(x-b)/h}} \exp\left(\frac{-u}{h}\right)\mathbb{1}_{[-\infty,b]}(u)$$

Noyaux non classiques

Pour une variable bornée sur $[a, b]$

Noyau Beta & Noyau Gaussien tronqué

$$K_h(u, x) = \frac{u^{x/h}(1-u)^{(1-x)/h}}{\mathcal{B}(\frac{x}{h}+1, \frac{1-x}{h}+1)} \mathbb{1}_{[0,1]}(u)$$

$$K_h(u, x) = \frac{\alpha}{h\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-x}{h})^2} \mathbb{1}_{[a,b]}(u),$$

$$\alpha := \left(\int_a^b \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{u-x}{h})^2} \right)^{-1}$$

Nearest Neighbors Smoothed Bootstrap

Nearest Neighbors Smoothed Bootstrap

distribution Uniforme $\subset$ distribution Beta

$$g_{i,\ell}^{int}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(\boldsymbol{x}^* - \boldsymbol{x}_{i.})^{\alpha-1}(\boldsymbol{x}_{i.}(\ell) - \boldsymbol{x}^*)^{\beta-1}\mathbb{1}_{[0;\boldsymbol{x}_{i.}(\ell)-\boldsymbol{x}_{i.}]}$$

Extended Nearest Neighbors Smoothed Bootstrap

Ajout d'une perturbation Gaussienne : Gaussienne-Beta mélange

$$g^{e-int}(\boldsymbol{x}^{**}|\tilde{\boldsymbol{x}}) = \sum_{i\in\mathcal{I}} \omega_i \sum_{\ell\in\mathcal{J}_i} \pi_{\ell|i}\; g_{i,\ell}^{e-int}(\boldsymbol{x}^{**}|\tilde{\boldsymbol{x}})$$

$$\text{avec } g_{i,\ell}^{e-int}(\boldsymbol{x}^{**}|\tilde{\boldsymbol{x}}) \sim N(\boldsymbol{x}^*, \sigma^2) \text{ et } \boldsymbol{x}^* \sim g_{i,\ell}^{int}(\boldsymbol{x}^*|\tilde{\boldsymbol{x}})$$

References I

Bowman, A. W., and Azzalini, A.

Applied smoothing techniques for data analysis : the kernel approach with s-plus illustrations.

Journal of the American Statistical Association 94 (1999), 982.

Branco, P., Torgo, L., and Ribeiro, R. P.

A survey of predictive modeling on imbalanced domains.

ACM computing surveys (CSUR) 49, 2 (2016), 1–50.

Gebelein, H.

Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung.

ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik 21, 6 (1941), 364–379.

References II

Grari, V., Lamprier, S., and Detyniecki, M.

Fairness-aware neural rényi minimization for continuous features.

Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20 (7 2020), 2262–2268.

Main track.

Hirschfeld, H. O.

A connection between correlation and contingency.

In *Mathematical Proceedings of the Cambridge Philosophical Society* (1935), vol. 31, Cambridge University Press, pp. 520–524.

Lee, and Sauchi.

Noisy replication in skewed binary classification.

Computational Statistics and Data Analysis 34, 2 (2000), 165–191.

Menardi, and Torelli.

Training and assessing classification rules with imbalanced data.

Data Mining and Knowledge Discovery 28, 1 (2014), 92–122.

References III

Mosley, L. S. D.

A balanced approach to the multi-class imbalance problem.

PhD thesis, Iowa State University, 2013.

Pagès, J.

Analyse factorielle de donnees mixtes: principe et exemple d'application.

Revue de statistique appliquée 52, 4 (2004), 93–111.

Rényi, A.

On measures of dependence.

Acta mathematica hungarica 10, 3-4 (1959), 441–451.

Ribeiro, R. P.

Utility-based regression.

Ph. D. dissertation (2011).

References IV

So, B., Boucher, J.-P., and Valdez, E. A.

Synthetic dataset generation of driver telematics.

Risks 9, 4 (2021), 58.

Torgo, L., and Ribeiro, R.

Utility-based regression.

In *Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings 11* (2007), Springer, pp. 597–604.