

Credit Scoring : biais d'échantillon ou *réintégration des refusés*

Adrien Ehrhardt

Christophe Biernacki, Vincent Vandewalle,
Philippe Heinrich, Sébastien Beben

Crédit Agricole Consumer Finance
INRIA Lille - Nord-Europe

Rencontres des jeunes statisticiens 2017 - Porquerolles

3-7 avril 2017



1 Contexte

- Entreprise
- Système d'acceptation
- Credit Scoring

2 Réintégration des refusés

- Par l'exemple
- Formalisation du problème
- Réinterprétation des méthodes
- Résultats expérimentaux

3 Conclusion

Contexte

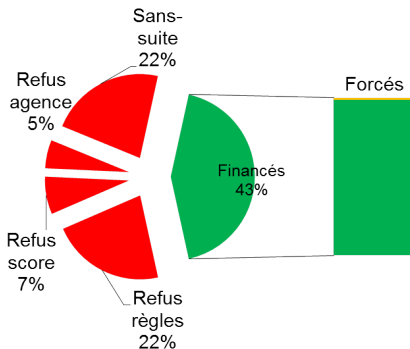


La Redoute



Contexte : Système d'acceptation

% Effectifs



X : vecteur a. des caractéristiques

Y dans $\{0, 1\}$: remboursement

Z dans $\{f, nf\}$: v. a. de financement

$$\exists \theta \in \mathbb{R}^{d+1} \text{ s.t. } \forall x, \ln \left(\frac{p_\theta(1|x)}{p_\theta(0|x)} \right) = \theta \cdot x$$

n clients financés ($Z = f$)

m clients non financés ($Z = nf$)

\mathbf{x} : *features* observées des clients

\mathbf{y} : remboursement observé

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^f \\ \mathbf{x}^{nf} \end{pmatrix}; \mathbf{y} = \begin{pmatrix} \mathbf{y}^f \\ \mathbf{y}^{nf} \end{pmatrix}$$

$$\underbrace{\ell(\theta; \mathbf{x}, \mathbf{y})}_{\text{vraisemblance complète}} = \left(\sum_{i=1}^n + \sum_{i=n+1}^{n+m} \right) \ln(p_\theta(y_i|x_i)) = \underbrace{\ell(\theta; \mathbf{x}^f, \mathbf{y}^f)}_{\text{vraisemblance observée}} + \ell(\theta; \mathbf{x}^{nf}, \mathbf{y}^{nf})$$

Quel intérêt à utiliser \mathbf{x}^{nf} ?

Quel risque à n'utiliser que $(\mathbf{x}^f, \mathbf{y}^f)$?

Réintégration des refusés

Réintégration des refusés : Exemple Fuzzy Augmentation I

Décrit par [Nguyen, 2016].

$$\begin{matrix} \mathbf{y}^f \\ \mathbf{y}^{nf} \end{matrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \text{NA} \\ \vdots \\ \text{NA} \end{pmatrix}$$

$$\begin{matrix} \mathbf{x}^f \\ \mathbf{x}^{nf} \end{matrix} \begin{pmatrix} x_1^1 & \cdots & x_1^d \\ \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^d \\ x_{n+1}^1 & \cdots & x_{n+1}^d \\ \vdots & \vdots & \vdots \\ x_{n+m}^1 & \cdots & x_{n+m}^d \end{pmatrix}$$

Abandon de \mathbf{x}^{nf} et construction de $\hat{\theta}^{\text{f}}$.

$$\begin{array}{c} \mathbf{y}^{\text{f}} \\ \mathbf{y}^{\text{nf}} \end{array} \left(\begin{array}{c} y_1 \\ \vdots \\ y_n \\ \text{NA} \\ \vdots \\ \text{NA} \end{array} \right)$$

$$\begin{array}{c} \mathbf{x}^{\text{f}} \\ \mathbf{x}^{\text{nf}} \end{array} \left(\begin{array}{ccc} x_1^1 & \cdots & x_1^d \\ \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^d \\ x_{n+1}^1 & \cdots & x_{n+1}^d \\ \vdots & \vdots & \vdots \\ x_{n+m}^1 & \cdots & x_{n+m}^d \end{array} \right)$$

Réintégration des refusés : Exemple Fuzzy Augmentation III

Remplacement de \mathbf{y}^{nf} par les proba données par $\hat{\theta}^{\text{f}}$.

$$\begin{array}{c} \mathbf{y}^{\text{f}} \\ \mathbf{y}^{\text{nf}} \end{array} \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ p_{\hat{\theta}^{\text{f}}}(y_{n+1} = 1 | x_{n+1}) \\ \vdots \\ p_{\hat{\theta}^{\text{f}}}(y_{n+m} = 1 | x_{n+m}) \end{pmatrix} \begin{array}{c} \mathbf{x}^{\text{f}} \\ \mathbf{x}^{\text{nf}} \end{array} \begin{pmatrix} x_1^1 & \cdots & x_1^d \\ \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^d \\ x_{n+1}^1 & \cdots & x_{n+1}^d \\ \vdots & \vdots & \vdots \\ x_{n+m}^1 & \cdots & x_{n+m}^d \end{pmatrix}$$

Apprendre $\hat{\theta}^{\text{fuzzy}}$ sur le dataset résultant.

Problème : $\hat{\theta}^{\text{fuzzy}} = \hat{\theta}^{\text{f}}$.

Objet d'intérêt : $p_{\text{vrai}}(y|x)$

Proposition d'un modèle : $p_{\theta}(y|x)$

Problème (sans surprise. . .) : estimer θ

Données :

- 1 Cas idéal : $\mathbf{x}^f, \mathbf{x}^{nf}$ et $\mathbf{y}^f, \mathbf{y}^{nf}$
- 2 Cas CACF : \mathbf{x}^f et \mathbf{y}^f

Réintégration des refusés : Formalisation du problème

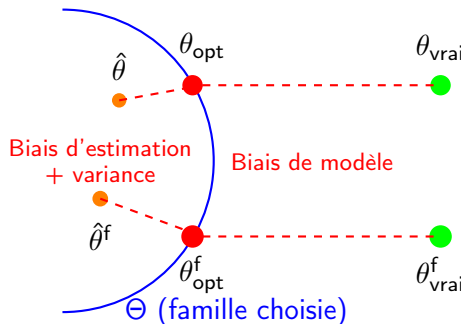
Objet d'intérêt : $p_{\text{vrai}}(y|x)$

Proposition d'un modèle : $p_{\theta}(y|x)$

Problème (sans surprise...) : estimer θ

Données :

- 1 Cas idéal : $\mathbf{x}^f, \mathbf{x}^{nf}$ et $\mathbf{y}^f, \mathbf{y}^{nf}$
- 2 Cas CACF : \mathbf{x}^f et \mathbf{y}^f



Estimateurs :

- 1 Cas idéal : avec toutes les données on obtient $\hat{\theta} = \operatorname{argmax}_{\theta} \ell(\theta; \mathbf{x}, \mathbf{y})$
$$\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n, m \rightarrow \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$$
- 2 Cas réaliste : avec les financés on obtient $\hat{\theta}^f = \operatorname{argmax}_{\theta} \ell(\theta; \mathbf{x}^f, \mathbf{y}^f)$
$$\sqrt{n}(\hat{\theta}^f - \theta_{\text{opt}}^f) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}^f}^f)$$

Réintégration des refusés : Formalisation du problème

Estimateurs :

- 1 Cas idéal : avec toutes les données on obtient $\hat{\theta} = \operatorname{argmax}_{\theta} \ell(\theta; \mathbf{x}, \mathbf{y})$
$$\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n, m \rightarrow \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$$
- 2 Cas réaliste : avec les financés on obtient $\hat{\theta}^f = \operatorname{argmax}_{\theta} \ell(\theta; \mathbf{x}^f, \mathbf{y}^f)$
$$\sqrt{n}(\hat{\theta}^f - \theta_{\text{opt}}^f) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}^f}^f)$$

Question 1 : propriétés asymptotiques des deux estimateurs

$$(Q1) \quad \theta_{\text{opt}} \stackrel{?}{=} \theta_{\text{opt}}^f$$

$$(Q2) \quad \text{ARE}(\hat{\theta}^f, \hat{\theta}) = \left(\varphi \frac{|\Sigma_{\theta_{\text{opt}}}|}{|\Sigma_{\theta_{\text{opt}}^f}^f|} \right)^{\frac{1}{d+1}}$$

$$(Q2a) \quad \frac{n}{n+m} \xrightarrow[n, m \rightarrow \infty]{a.s.} \varphi$$

$$(Q2b) \quad \Sigma_{\theta_{\text{opt}}} \stackrel{?}{=} \Sigma_{\theta_{\text{opt}}^f}^f$$

- **MCAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z)$
→ Inadapté au Credit Scoring.

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z)$
→ Inadapté au Credit Scoring.
- **MAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z|x)$
→ Le score détermine l'acceptation : $Z = \mathbb{1}_{\{\theta'X > \text{cut}\}}$.

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z)$
→ Inadapté au Credit Scoring.
- **MAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z|x)$
→ Le score détermine l'acceptation : $Z = \mathbb{1}_{\{\theta'X > \text{cut}\}}$.
- **MNAR** : $\exists x, y, z, p_{\text{vrai}}(z|x, y) \neq p_{\text{vrai}}(z|x)$
→ Influence du "feeling" des conseillers X^c .

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z)$
→ Inadapté au Credit Scoring.
- **MAR** : $\forall x, y, z, p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z|x)$
→ Le score détermine l'acceptation : $Z = \mathbb{1}_{\{\theta'X > \text{cut}\}}$.
- **MNAR** : $\exists x, y, z, p_{\text{vrai}}(z|x, y) \neq p_{\text{vrai}}(z|x)$
→ Influence du "feeling" des conseillers X^c .

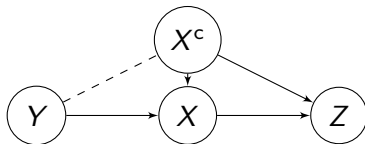


Figure – Schéma des dépendances des variables aléatoires Y , X^c , X et Z

Réintégration des refusés : Bon/mauvais modèle

- **Bon modèle** : $\exists \theta_{\text{vrai}}, p_{\text{vrai}}(y|x) = p_{\theta_{\text{vrai}}}(y|x)$.
→ Données réelles \Rightarrow hypothèse peu probable.
- **Mauvais modèle** : θ_{opt} minimise l'ignorance sur la vraie loi.
→ Utilisation de la régression logistique pour sa robustesse à la misspecification.

Réintégration des refusés : Bon/mauvais modèle

- **Bon modèle** : $\exists \theta_{\text{vrai}}, p_{\text{vrai}}(y|x) = p_{\theta_{\text{vrai}}}(y|x)$.
→ Données réelles \Rightarrow hypothèse peu probable.
- **Mauvais modèle** : θ_{opt} minimise l'ignorance sur la vraie loi.
→ Utilisation de la régression logistique pour sa robustesse à la misspecification.

$p_{\theta}(y x, z)$ \ $p_{\text{vrai}}(z x)$	MCAR	MAR	MNAR
Bon	$\theta_{\text{opt}}^f = \theta_{\text{opt}}$	$\theta_{\text{opt}}^f = \theta_{\text{opt}}$ $\Sigma_{\theta_{\text{opt}}^f}^f \neq \Sigma_{\theta_{\text{opt}}}$	$\theta_{\text{opt}}^f \neq \theta_{\text{opt}}$
Mauvais	$\Sigma_{\theta_{\text{opt}}^f}^f = \Sigma_{\theta_{\text{opt}}}$	$\theta_{\text{opt}}^f \neq \theta_{\text{opt}}$ $\Sigma_{\theta_{\text{opt}}^f}^f \neq \Sigma_{\theta_{\text{opt}}}$	$\Sigma_{\theta_{\text{opt}}^f}^f \neq \Sigma_{\theta_{\text{opt}}}$

Table – Réponses à (Q1) et (Q2b) selon le mécanisme des données manquantes et l'hypothèse du bon modèle

Question 2 : Comment améliorer $\hat{\theta}^f$?

Question 2 : Comment améliorer $\hat{\theta}^f$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),

Question 2 : Comment améliorer $\hat{\theta}^f$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),
- Modéliser la sélection (i.e. $p_{\alpha}(z|x, y)$),

Question 2 : Comment améliorer $\hat{\theta}^f$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),
- Modéliser la sélection (i.e. $p_\alpha(z|x, y)$),
- Utiliser \mathbf{x}^{nf} .

Question 2 : Comment améliorer $\hat{\theta}^f$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),
- Modéliser la sélection (i.e. $p_\alpha(z|x, y)$),
- Utiliser \mathbf{x}^{nf} .

Moyen “naturel” de faire les trois : le modèle génératif $p_\theta(y|x)$, $p_\beta(x)$ et $p_\gamma(z|x, y)$ où θ , β et γ sont fonctionnellement dépendants (en général).

$$\hat{\theta}^{\text{gen}} = \operatorname{argmax} \ell(\theta; \mathbf{x}, \mathbf{y}^f) = \ell(\theta; \mathbf{x}^f, \mathbf{y}^f) + \ell(\theta; \mathbf{x}^{nf})$$

généralement obtenu par algorithme EM.

Question 2 : Comment améliorer $\hat{\theta}^f$?

Leviers :

- ~~Changer le modèle (i.e. l'espace Θ)~~ régression logistique,
- Modéliser la sélection (i.e. $p_\alpha(z|x, y)$),
- Utiliser \mathbf{x}^{nf} .

Moyen “naturel” de faire les trois : le modèle génératif $p_\theta(y|x)$, $p_\beta(x)$ et $p_\gamma(z|x, y)$ où θ , β et γ sont fonctionnellement dépendants (en général).

$$\hat{\theta}^{\text{gen}} = \operatorname{argmax} \ell(\theta; \mathbf{x}, \mathbf{y}^f) = \ell(\theta; \mathbf{x}^f, \mathbf{y}^f) + \ell(\theta; \mathbf{x}^{nf})$$

généralement obtenu par algorithme EM.

Question 2 : Comment améliorer $\hat{\theta}^f$?

Leviers :

- Changer le modèle (i.e. l'espace Θ) régression logistique,
- Modéliser la sélection (i.e. $p_\alpha(z|x, y)$) relève de la croyance,
- Utiliser \mathbf{x}^{nf} .

Moyen “naturel” de faire les trois : le modèle génératif $p_\theta(y|x)$, $p_\beta(x)$ et $p_\gamma(z|x, y)$ où θ , β et γ sont fonctionnellement dépendants (en général).

$$\hat{\theta}^{\text{gen}} = \operatorname{argmax} \ell(\theta; \mathbf{x}, \mathbf{y}^f) = \ell(\theta; \mathbf{x}^f, \mathbf{y}^f) + \ell(\theta; \mathbf{x}^{nf})$$

généralement obtenu par algorithme EM.

Reclassification

([Viennet et al., 2006, Banasik and Crook, 2007, Guizani et al., 2013]) :

$$(\hat{\theta}^{\text{CEM}}, \hat{\mathbf{y}}^{\text{nf}}) = \underset{\theta, \mathbf{y}^{\text{nf}}}{\operatorname{argmax}} \ell(\theta; \mathbf{x}, \mathbf{y}^{\text{f}}, \mathbf{y}^{\text{nf}}) = \ell(\theta; \mathbf{x}^{\text{f}}, \mathbf{y}^{\text{f}}) + \sum_{i=n+1}^{m+n} \ln(p_{\theta}(\hat{y}_i | x_i)).$$

où $\hat{y}_i = \operatorname{argmax}_{y_i} p_{\hat{\theta}^{\text{f}}}(y_i | x_i)$.

C'est une tentative de “mimer” le modèle génératif par l'utilisation d'un algorithme Classification-EM.

Problème : produit un estimateur inconsistant, aucune garantie d'amélioration.

Augmentation ([Viennet et al., 2006, Banasik and Crook, 2007, Guizani et al., 2013, Nguyen, 2016]) : critère issu de l'*Importance Sampling* permettant de corriger le biais de modèle **mais** algorithme insatisfaisant.

Twins (méthode interne CACF) : le critère conduit à $\hat{\theta}^{\text{twins}} = \hat{\theta}^f$.

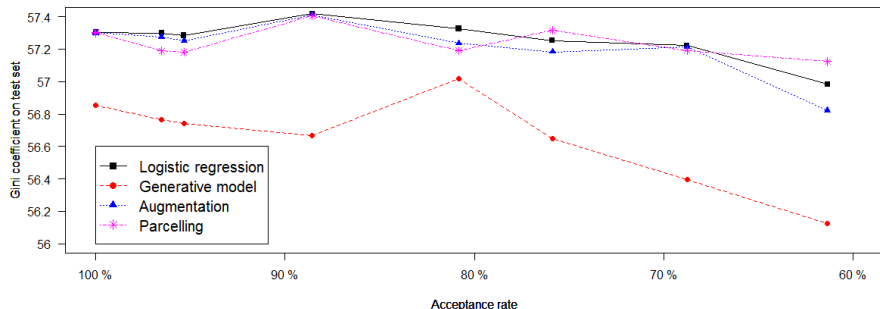
Parcelling

([Viennet et al., 2006, Banasik and Crook, 2007, Guizani et al., 2013]) : hypothèses MNAR invérifiables.

Génératif : biais de modèle important.

Réintégration des refusés : Résultats expérimentaux

Gini coefficient w.r.t. the acceptance rate of the previous scoring model



Conclusion

- **Fuzzy Augmentation, Reclassification** et **Twins** ont été écartées.
- Le choix d'une (autre) méthode relève de la croyance.
- La formalisation a permis de clotûrer un débat resté empirique jusqu'alors.
- Article en préparation.

- **Fuzzy Augmentation, Reclassification** et **Twins** ont été écartées.
- Le choix d'une (autre) méthode relève de la croyance.
- La formalisation a permis de clotûrer un débat resté empirique jusqu'alors.
- Article en préparation.
- **Sujet en cours** permettant de limiter le biais de modèle : discrétisation de variables continues, regroupement de modalités pour variables qualitatives et croisements automatiques de variables.
- **Sujet à venir** : augmentation du nombre de prédicteurs permettant d'abaisser l'erreur de Bayes.

Merci pour votre attention !

Questions ?



Banasik, J. and Crook, J. (2007).

Reject inference, augmentation, and sample selection.

[European Journal of Operational Research](#), 183(3) :1582–1594.



Guizani, A., Souissi, B., Ammou, S. B., and Saporta, G. (2013).

Une comparaison de quatre techniques d'inférence des refusés dans le processus d'octroi de crédit.

In [45 emes Journ'ees de statistique](#), page pp.



Nguyen, H. T. (2016).

Reject inference in application scorecards : evidence from France.

Technical report, University of Paris West-Nanterre la Défense, EconomiX.



Viennet, E., Soulié, F. F., and Rognier, B. (2006).

Evaluation de techniques de traitement des refusés pour l'octroi de crédit.

[arXiv preprint cs/0607048](#).