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

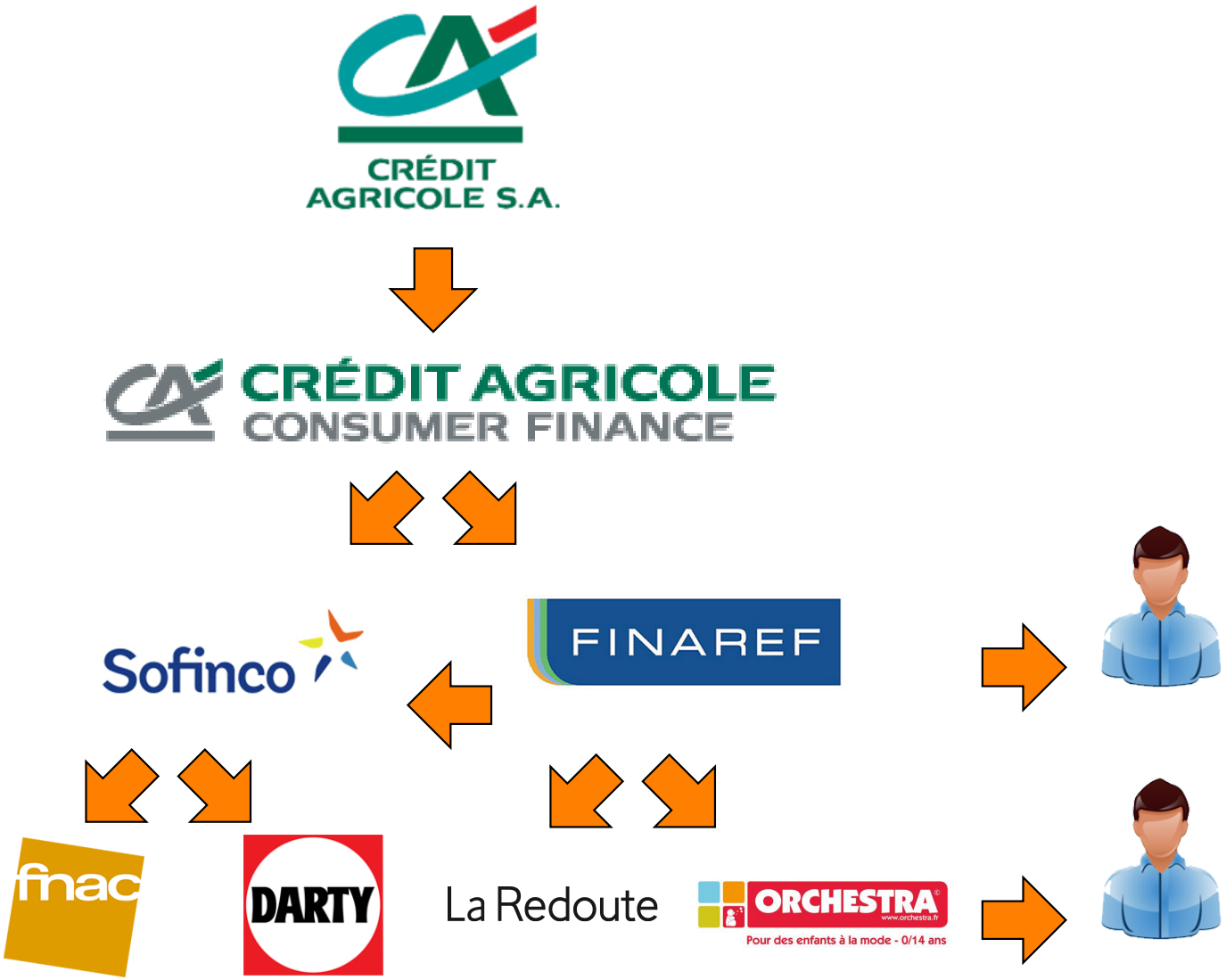
Modal's days 2017

19 janvier 2017

Sommaire

Contexte

Contexte : Entreprise



Contexte : Produits et demandes de crédit I

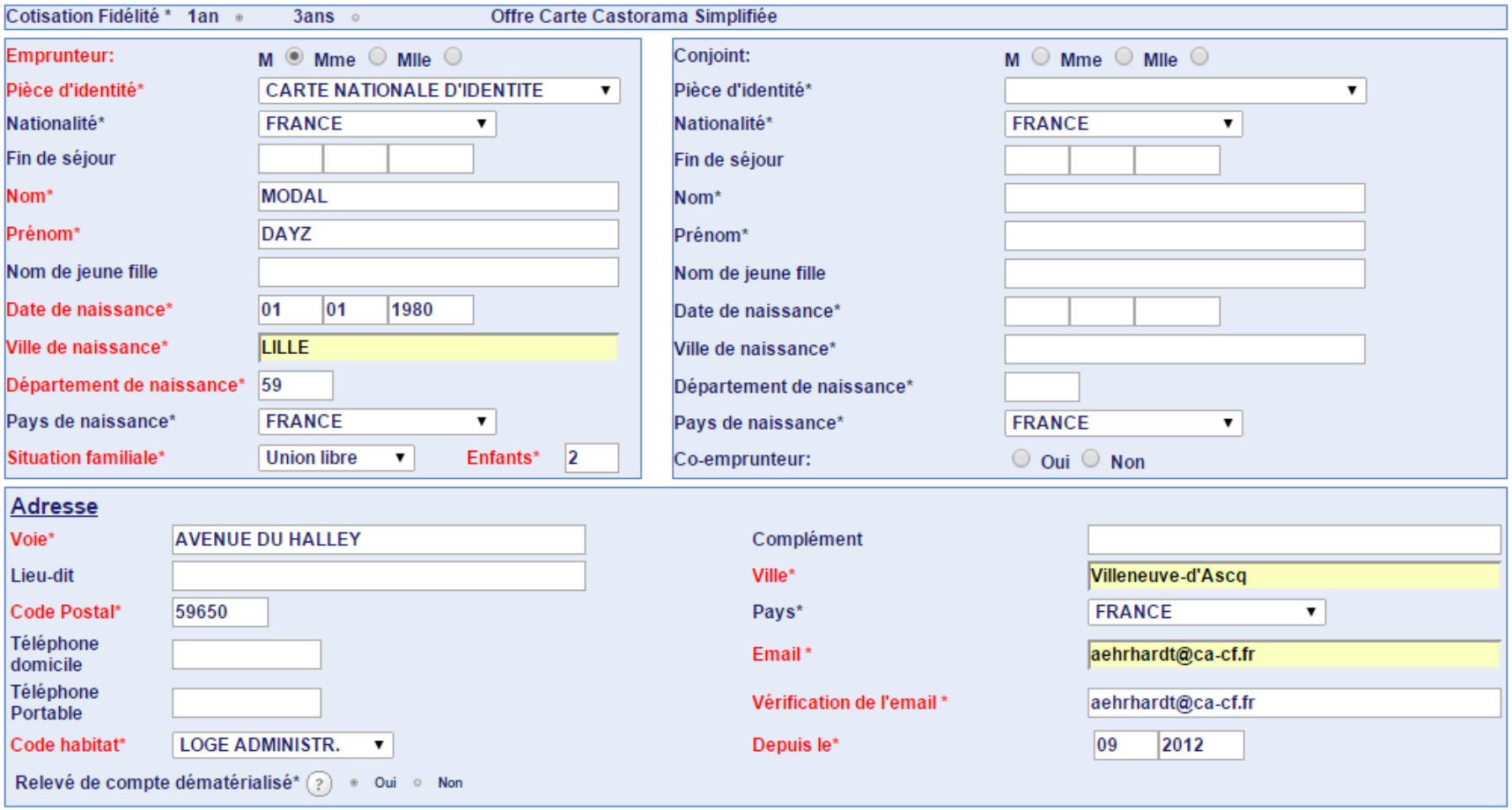
Quelques chiffres :

- Encours gérés (2015) : 71 milliards € (26 milliards € en France)
- Produit Net Bancaire : 2 milliards €
- Résultat : +380 M€
- **Coût du risque** France : 200 M€

Produits :

- Facilités de paiement (exemple : en 3 fois)
- Prêt personnel (mensualité + échéance)
- Carte de crédit (revolving)

Contexte : Produits et demandes de crédit II



Cotisation Fidélité * 1an 3ans Offre Carte Castorama Simplifiée

Emprunteur:	M Mme Mlle
Pièce d'identité*	CARTE NATIONALE D'IDENTITE
Nationalité*	FRANCE
Fin de séjour	
Nom*	MODAL
Prénom*	DAYZ
Nom de jeune fille	
Date de naissance*	01 01 1980
Ville de naissance*	LILLE
Département de naissance*	59
Pays de naissance*	FRANCE
Situation familiale*	Union libre — Enfants* 2

Conjoint:	M Mme Mlle
Pièce d'identité*	
Nationalité*	FRANCE
Fin de séjour	
Nom*	
Prénom*	
Nom de jeune fille	
Date de naissance*	
Ville de naissance*	
Département de naissance*	
Pays de naissance*	FRANCE
Co-emprunteur:	Oui Non

Adresse

Voie*	AVENUE DU HALLEY	Complément	
Lieu-dit		Ville*	Villeneuve-d'Ascq
Code Postal*	59650	Pays*	FRANCE
Téléphone domicile		Email *	aehrhardt@ca-cf.fr
Téléphone Portable		Vérification de l'email *	aehrhardt@ca-cf.fr
Code habitat*	LOGE ADMINISTR.	Depuis le*	09 2012

Relevé de compte dématérialisé* (?) Oui Non

Contexte : Produits et demandes de crédit III

Situation professionnelle emprunteur

Profession*	ARTISTE, PRESSE 35		
Employeur		Depuis*	09 2012
Adresse			
Code Postal		Ville	
Pays	FRANCE		
Téléphone Employeur			

Situation professionnelle conjoint

Profession			
Employeur		Depuis	
Adresse			
Code Postal		Ville	
Pays	FRANCE		
Téléphone Employeur			

Revenus :

Emprunteur :	
Revenus nets mensuels*	1800
Autres revenus nets mensuels*	200
Conjoint :	
Revenus nets mensuels	
Autres revenus nets mensuels	
Prestations familiales*	50
Total	2050,00 €

Charges :

Charges mensuelles d'habitation*	600
Autres charges mensuelles*	150
Autres crédits*	0
Total	750,00 €

Carte facultative emprunteur : ○ Oui ◉ Non

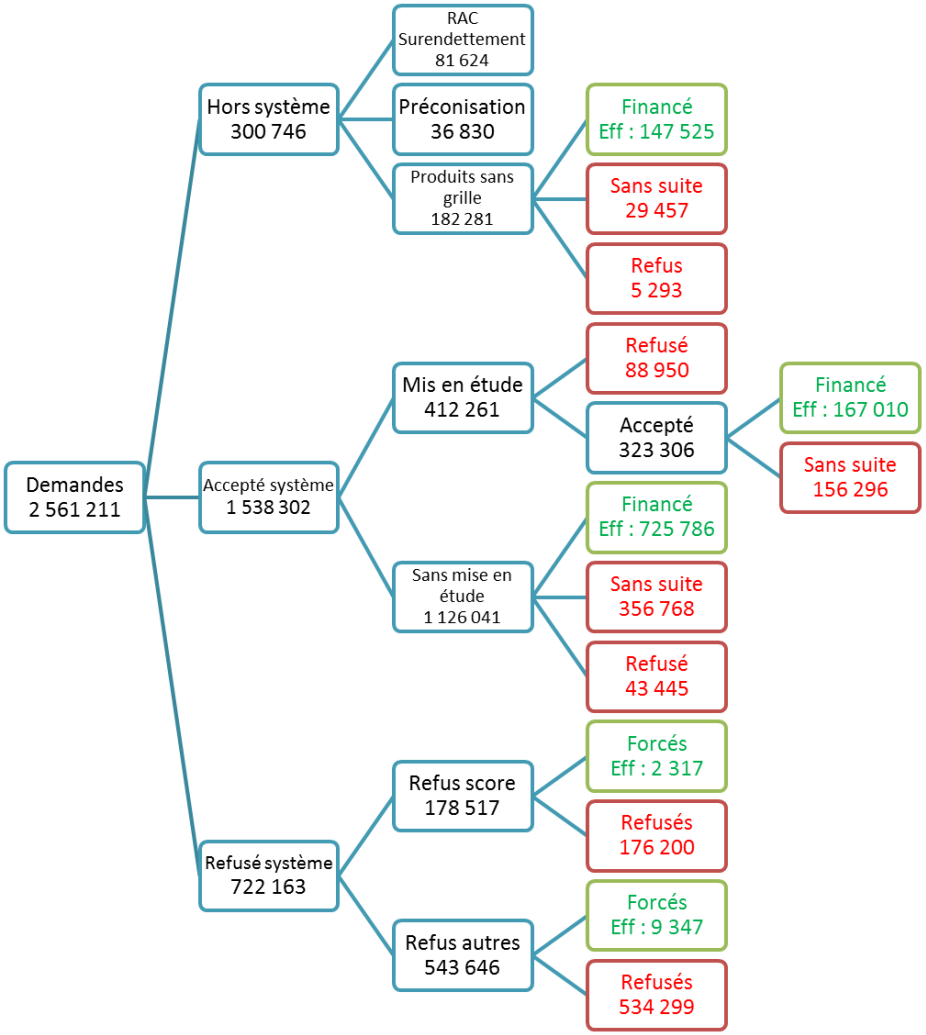
Assurances ☑ DIM CACI T1 — Offre promo ◉ Oui ○ Non

Coordonnées bancaires ou postales

BIC | IBAN | Titulaire

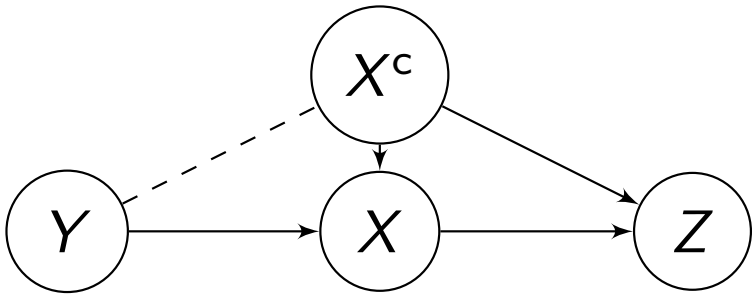
Contexte : Produits et demandes de crédit IV

Contexte : Système d'acceptation



X : vecteur a. des caractéristiques
Y dans $\{0, 1\}$: remboursement
A_S dans $\{\mathrm{a}, \bar{\mathrm{a}}\}$: $f(X)$
A_C dans $\{\mathrm{a}, \bar{\mathrm{a}}\}$: $g(X, X^{\mathrm{c}})$
Z dans $\{\mathrm{f}, \mathrm{nf}\}$: v. a. de financement

On oublie A_S et A_C dans la suite.



Contexte : Credit Scoring

$$\exists \theta = (\theta_0, \ldots, \theta_d) \in \mathbb{R}^{d+1} \text{ s.t. } \forall x, \ln\left(\frac{P(Y=1|x)}{P(Y=0|x)}\right) = \theta_0 + \theta_1 x^1 + \ldots + \theta_d x^d$$

n clients financés ($Z = \text{f}$)
m clients non financés ($Z = \text{nf}$)
$\boldsymbol{x}$: *features* observées des clients
$\boldsymbol{y}$: remboursement observé

$\boldsymbol{x}^{\text{f}}$ ($\boldsymbol{x}^{\text{nf}}$) : *features* observées des clients financés (non financés)
$\boldsymbol{y}^{\text{f}}$ ($\boldsymbol{y}^{\text{nf}}$) : remboursement (non) observé des clients financés (non financés)

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

Quel intérêt à utiliser $\boldsymbol{x}^{\text{nf}}$?
Quel risque à n'utiliser que $(\boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})$?

Réintégration des refusés

Réintégration des refusés : Exemple Fuzzy Augmentation I

$$\boldsymbol{y}^{\mathrm{f}} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{y}^{\mathrm{nf}} \begin{pmatrix} \mathrm{NA} \\ \vdots \\ \mathrm{NA} \end{pmatrix}$$

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

Réintégration des refusés : Exemple Fuzzy Augmentation II

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

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

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

Réintégration des refusés : Exemple Fuzzy Augmentation III

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

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

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

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

Réintégration des refusés : Etat de l'art

[Feelders, 2000] : formalisation et approche par données manquantes.

[Viennet et al., 2006, Guizani et al., 2013, Banasik and Crook, 2007, Nguyen, 2016] : description de méthodes empiriques pour utiliser $\boldsymbol{x}^{\text{nf}}$.

[Viennet et al., 2006, Banasik and Crook, 2007] : la réintégration a un intérêt.

[Nguyen, 2016, Guizani et al., 2013] : la réintégration n'a pas d'intérêt.

[Kiefer and Larson, 2006] : étude de cas du mauvais modèle.

Réintégration des refusés : Formalisation du problème

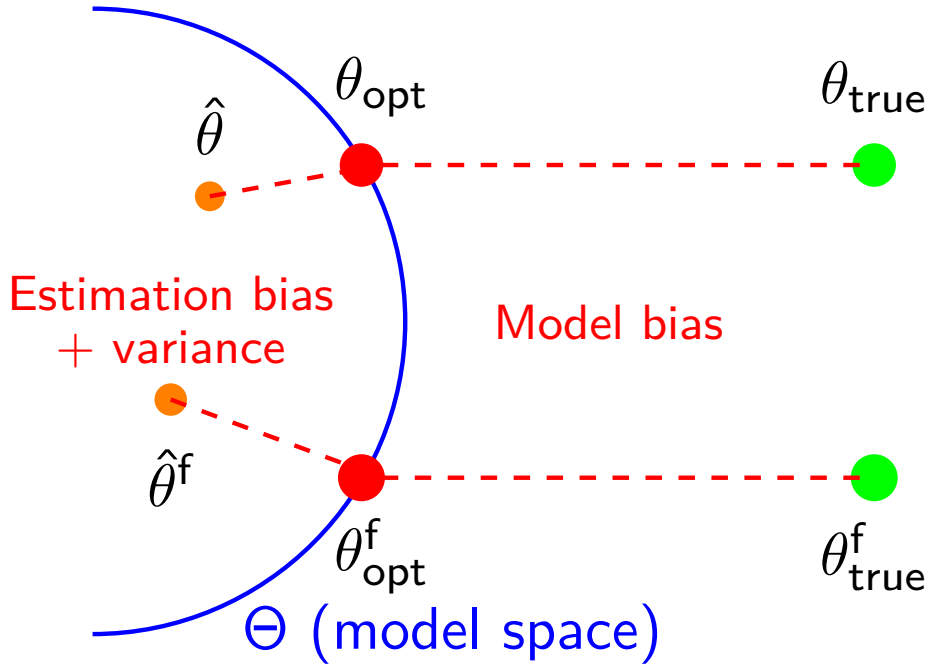
Objet d'intérêt : $p_{\text{true}}(y|x)$

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

Problème (sans surprise...) : estimer θ

Données :

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



Réintégration des refusés : Formalisation du problème

Estimateurs :

1. Cas idéal : $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Cas CACF : $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Question 1 : propriétés asymptotiques des deux estimateurs

$$\boxed{\text{(Q1)}\ \theta_{\text{opt}} \stackrel{?}{=} \theta^{\text{f}}_{\text{opt}}}$$

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

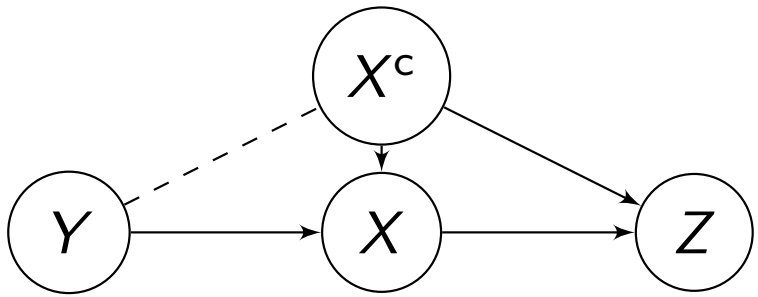
(Q2a) $\dfrac{n}{n+m} \xrightarrow[n,m\to\infty]{a.s.} \varphi$ (Q2b) $\Sigma_{\theta_{\text{opt}}} \stackrel{?}{=} \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}}$

Réponse 1 (toujours sans surprise. . .) : ça dépend.

1. Mécanisme des données manquantes $p_{\text{true}}(z|x)$
2. La vraie loi prédictive $p_{\theta_{\text{true}}}(y|x)$ appartient-elle à Θ ?

Réintégration des refusés : Mécanisme de manque

- **MCAR** : $\forall\, x, y, z,\ p_{\text{true}}(z|x, y) = p_{\text{true}}(z)$
  $\rightarrow$ Inadapté au Credit Scoring.
- **MAR** : $\forall\, x, y, z,\ p_{\text{true}}(z|x, y) = p_{\text{true}}(z|x)$
- **MNAR** : $\exists\, x, y, z,\ p_{\text{true}}(z|x, y) \neq p_{\text{true}}(z|x)$
  $\rightarrow$ Influence du "feeling" des conseillers X^c.



Ignorabilité :
Les paramètres de $p_{\text{true}}(z|x, y)$ ne dépendent pas fonctionnellement de θ_{true}.

Réintégration des refusés : Bon/mauvais modèle

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

$p_\theta(y|x,z)$ \ $p_{\text{true}}(z|x)$	MCAR	MAR	MNAR
True	$\theta^{\text{f}}_{\text{opt}} = \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} = \Sigma_{\theta_{\text{opt}}}$	$\theta^{\text{f}}_{\text{opt}} = \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$	$\theta^{\text{f}}_{\text{opt}} \neq \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$
Misspecified		$\theta^{\text{f}}_{\text{opt}} \neq \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$	

Réintégration des refusés : Quelle case en Credit Scoring ?

Mécanisme de manque :

- MCAR : la sélection des clients n'est pas aléatoire !
- MAR : vrai si X détermine Z (cas du score)
- MNAR : vrai si conseillers $=$ oracles

On ne peut pas tester MNAR.

Bon modèle : non.
Mais toutes les variables sont discrétisées : on peut approcher n'importe quelle relation fonctionnelle entre la variable cible et chaque variable continue.

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

Question 2 : Comment "sauter" dans une meilleure case ?

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 $\boldsymbol{x}^{\mathrm{nf}}$.

Case 1 : MAR et bon modèle
Idée : Améliorer la vitesse de convergence
Proposition : Modèle génératif avec données manquantes
Case 2 : MAR et mauvais modèle
Idée : "Dé-biaiser" le modèle en corrigeant $p_\theta(y|x)$
Proposition : Méthode "Augmentation" proche de l'Importance Sampling
Case 3 : MNAR
Idée : "Dé-biaiser" le modèle en corrigeant $p_\theta(y|x,z)$
Proposition : Méthode "Parcelling" : hypothèses invérifiables

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

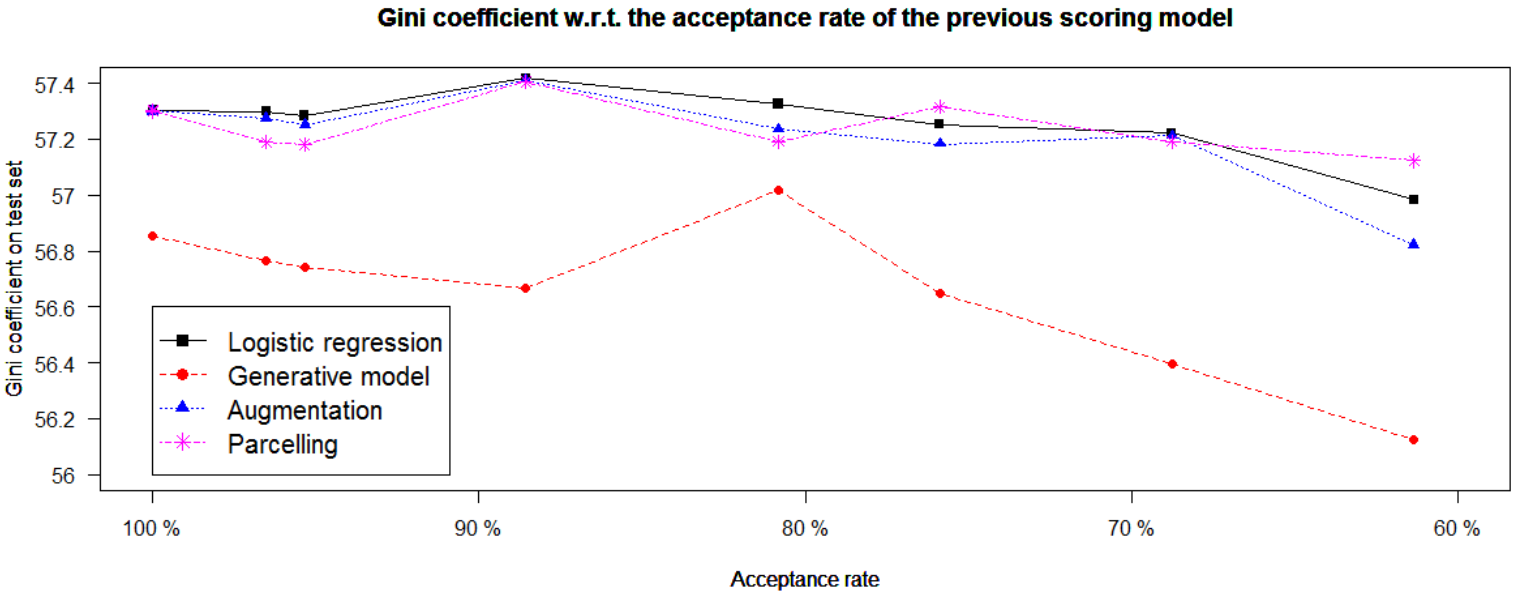
Certaines méthodes sont ré-interprétables comme des hypothèses partielles sur la loi jointe.

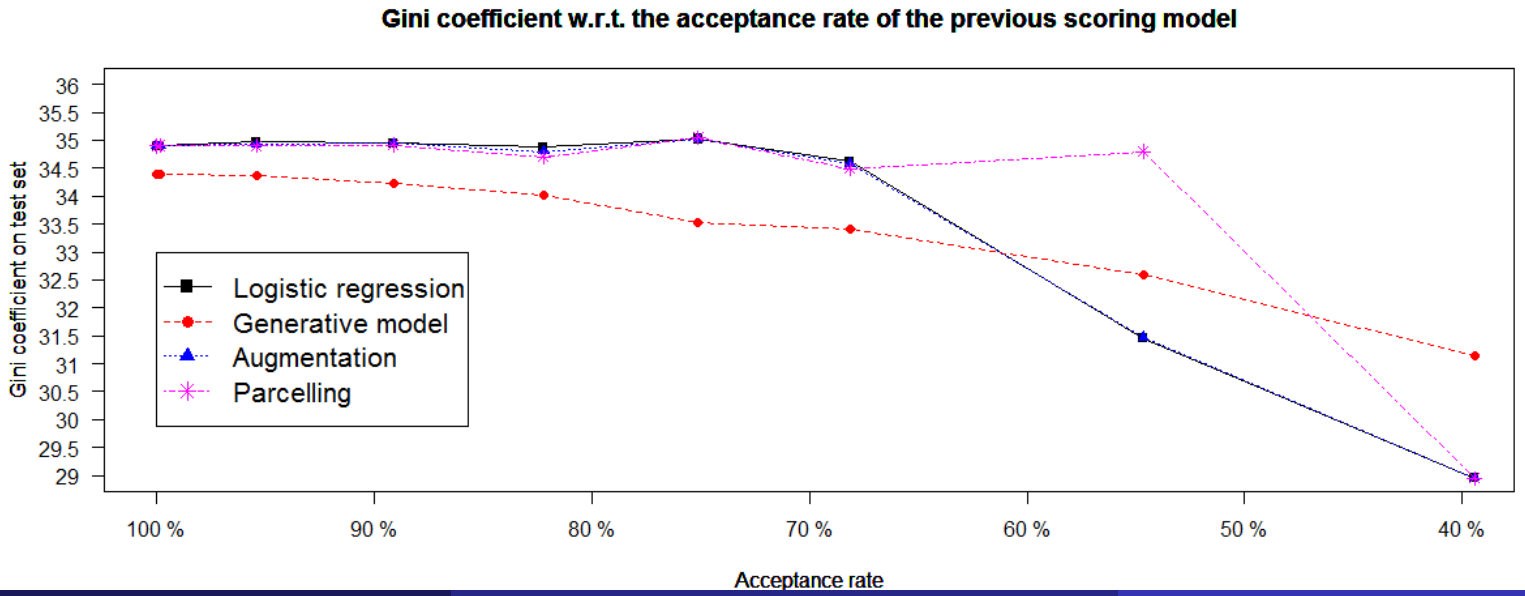
6 méthodes testées :

1. 2 méthodes conduisent à $\theta^{\mathrm{f}}_{\mathrm{opt}}$.
2. 1 méthode utilisant un algorithme sans garantie de convergence.
3. Modèle génératif : biais de modèle important.
4. Augmentation (MAR/mauvais modèle) : hypothèses non satisfaites.
5. Parcelling (MNAR) : introduction d'*a priori*.

Réponse 2 : arbitrer entre (pas de réintégration) et (méthode Toto de réintégration), c'est comparer deux "prises de risque" inquantifiables.

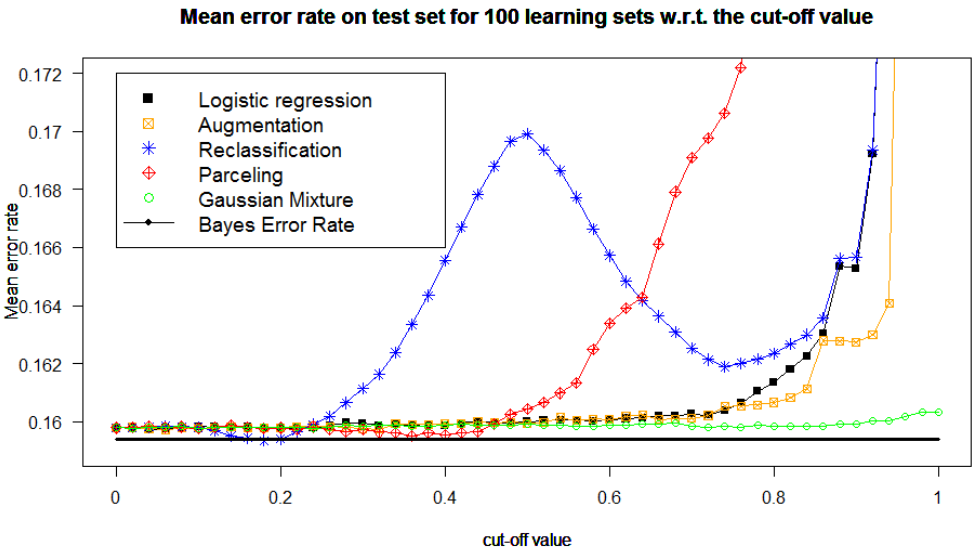
Réintégration des refusés : Résultats

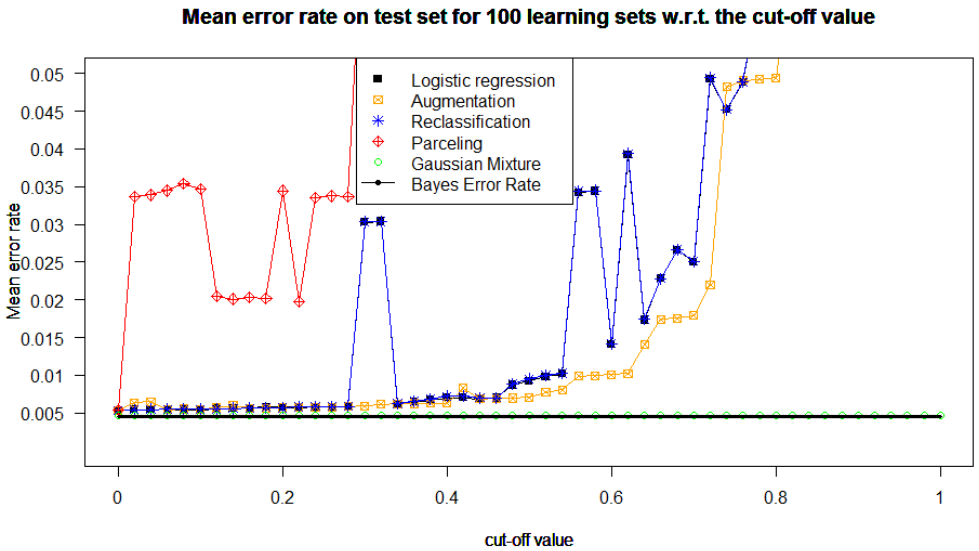




Réintégration des refusés : Augmentation de la dimension

Gains hypothétiques de l'utilisation de $\mathbf{x}^{nf}$ faibles





Grands gains potentiels de l'augmentation du nombre de prédicteurs :

- On agit sur le modèle.
- On abaisse le taux d'erreur de Bayes (classes séparables).

Conclusion

Conclusion

- Le choix d'une méthode relève de la croyance.
- La formalisation a permis de clotûrer un débat resté empirique jusqu'alors.
- Article en préparation
- Sujet en cours : discrétisation

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.


Feelders, A. (2000).
Credit scoring and reject inference with mixture models.
International Journal of Intelligent Systems in Accounting, Finance & Management, 9(1):1–8.


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.


Kiefer, N. M. and Larson, C. E. (2006).
Specification and informational issues in credit scoring.
Available at SSRN 956628.


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.