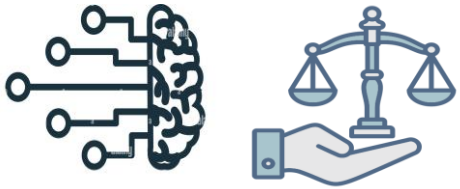
Présentation de mémoire

23 / 10 / 2025

Équité des modèles en assurance : mesure et mitigation des discriminations



Marc Jodel SIMO

A partir du mémoire encadré par :

Caroline HILLAIRET
Tutrice Académique

Boris NOUMEDEM
Tuteur Professionnel

PLAN

01 Pourquoi s'intéresser à l'équité en assurance ?

02 Équité de modèles – plusieurs définitions

03 Mesure, diagnostic et mitigation – les outils

04 Application à la tarification en assurance RC automobile

05 Conclusion

Equité en assurance, Pourquoi s'y intéresser?...

Utilisation de plus en plus importante des modèles de ML&IA dans l'industrie de l'assurance.

- Evaluation des risques et tarification
- Détection de la fraude
- Modélisation de l'élasticité
- Analyse des conversations téléphoniques et retours clients
- Traitement automatisé des documents
- LLM et applications d'IA générative
- ...

- … L'équité n'est pas intégrée automatiquement dans les modèles d'IA
- **Mais c'est un sujet important pour les raisons suivantes :**

Exigences réglementaires

Directive européenne sur légalité de genre (2012)

Les assureurs ne sont pas autorisés à utiliser le genre comme variable de tarification dans les produits d'assurance.

L'AI-Act (2024) | Lignes directrices éthiques pour une IA digne de confiance (2019)

Une IA digne de confiance doit être: légale, éthique et robuste.

Risque Réputationnel pour les assureurs

Perception d'un traitement injuste

Des pratiques jugées inéquitables peuvent nuire gravement à la reputation et affaiblir la confiance des clients et du public.

Une responsabilité collective

L'absence de vigilance sur l'équité mm pour une seule entité peut entrainer des retombées médiatiques négatives pour l'ensemble d'un groupe.

Quelques illustrations de discriminations attribuables aux modèles IA

Institut des ACTUAIRES

Larson et al.(2016), sur l'outil Compas :

« Les accusés noirs étaient souvent considérés comme présentant un risque de récidive plus élevé qu'ils ne l'étaient en réalité » (le taux de faux négatifs était deux fois plus élevé que les autres).

Risque de récidive dans les tribunaux américains

Recrutement automatisé chez Amazon

Budd LP et al. (2021):

L'algorithme choisit préférentiellement les hommes aux femmes, même si les noms ont été supprimés des CV. Amazon a entraîné l'algorithme à partir de ses CV existants, où les hommes sont surreprésentés..

Mercier Fr (2018)

« En 2018, il était possible d'économiser 1100 $ sur votre prime d'assurance auto simplement en changeant de genre. » (Canada)

Objectifs de l'étude

Comprendre

Fournir un cadre pour le questionnement et le bon choix d'objectifs en matière d'équité de modèles

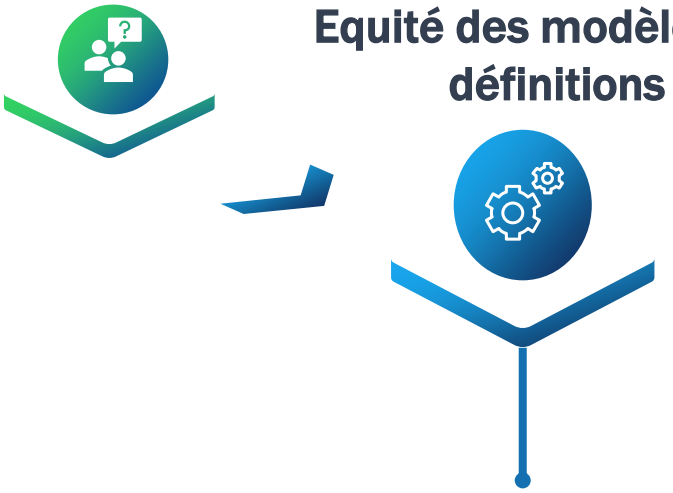
Evaluer

Evaluer l'équité de modèles prédictifs utilisés en assurance (ex: pricing, claim scoring, …)

Corriger

Réduire, lorsque nécessaire, les discriminations induites par les modèles.

Introduction



Equité des modèles, les définitions

Qu'est-ce que l'équité ?

équité – Oxford dictionary

Traitement ou comportement impartial et juste, sans **favoritisme** ni **discrimination**.

discrimination – Oxford dictionary

Action de **traiter un groupe particulier de personnes diffféremment,** notamment en raison de caractéristiques telles-que **l'âge, le genre, l'origine ethnique, …**

20 definitions de l'équité

Equité de groupe — appliquer les mêmes règles à tous les groupes

Equité individuelle — à individus comparables prédictions comparables

Définir l'(in)équité des modèles, le cadre

Configuration du machine learning dans le contexte de l'équité

- $\boldsymbol{X}$ l'ensemble de variables "***légitimes***"
- $\mathbf{D}$ **l'attribut sensible/protégée**
- $\mathbf{Y}$ la variable d'intérêt observée
- $\hat{Y}$ les prédictions

Attributs sensibles ou protégés

Attributs, propriétés ou traits qui, selon la loi*, ne peuvent faire l'objet d'aucune discrimination.

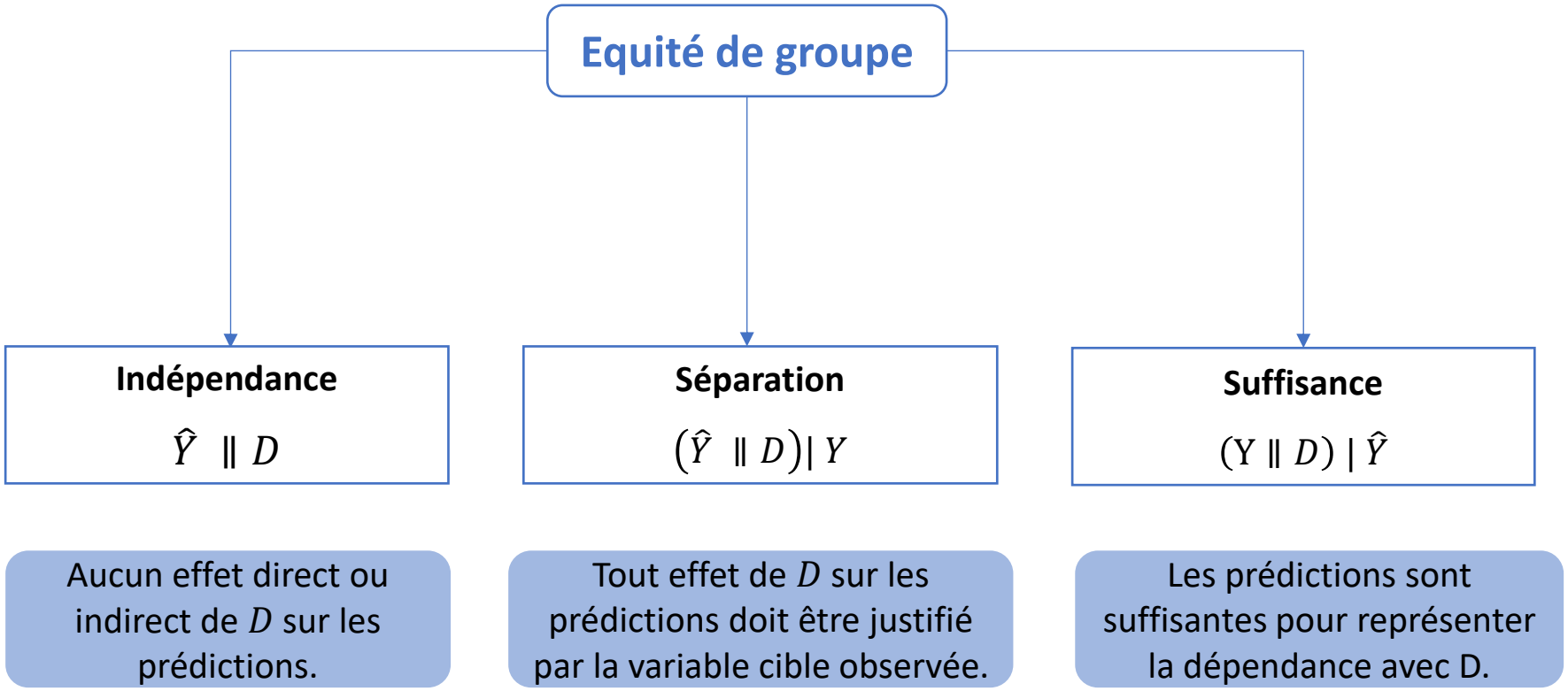
Exemples :

- Âge
- Handicap
- Genre
- Langue maternelle
- Opinions politiques
- Origine, appartenance ethnique
- ….

(Caractéristiques source de biais.)

** Manuel européen du droit de la non-discrimination*

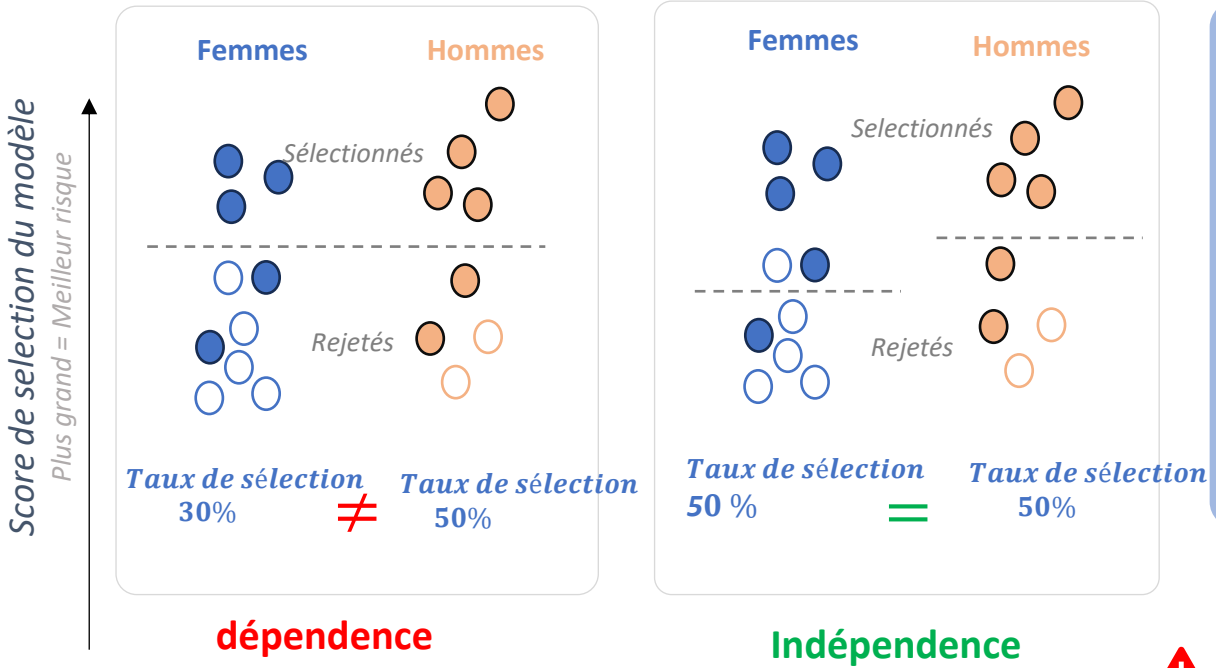
Equité de groupe, 3 grands principes



Le principe d'indépendance : parité démographique

Exemple : **Sélection automatique de risques Auto**
$D =$ **genre**

Parité demographique / statistique
Les chances de sélectionner un risque sont **les mêmes** pour hommes et femmes



Un modèle vérifie la parité démographique si ses prédictions sont indépendantes
de l'attribut sensible: $\hat{Y} \perp D$

Signification pour un classificateur (binaire)

$$P(\hat{Y} = 1 | D = d_1) = P(\hat{Y} = 1 | D = d_2)$$

- Les profils de risque F/H sont-ils **identiques** ?
- Est-ce vraiment équitable de traiter différemment 2 individus ayant le même score ?

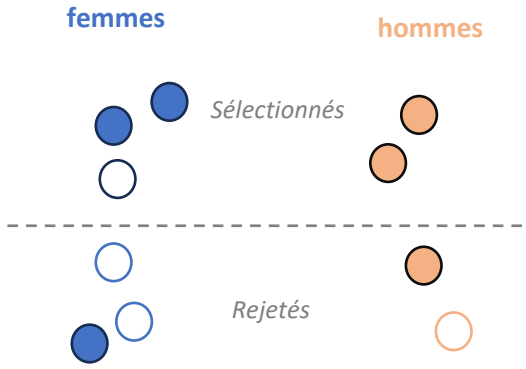
Le principe d'indépendance : parité démographique conditionnelle

Exemple : **Sélection automatique de risques Auto**
D = **genre**

Considérons $\boldsymbol{R}$ le niveau de garantie d'assurance du contrat (partielle vs complète)

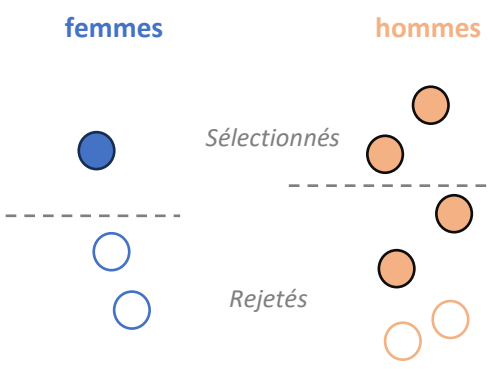
La **parité démographique conditionnelle**
Les chances de sélection des risques auto sont **identiques** pour les hommes et femmes partageant **les mêmes types de garanties**

R = **partielle**



Taux de sélection **50%** = ***Taux de sélection*** **50%**

R = **complète**



Taux de sélection **33%** = Taux de sélection **33%**

La parité **démographique conditionnelle** est vérifiée pour un modèle si les prédictions sont indépendantes de la variable sensible pour les individus partageant **la même variable de contrôle**

$(\hat{Y} \perp \boldsymbol{D}) \mid R$

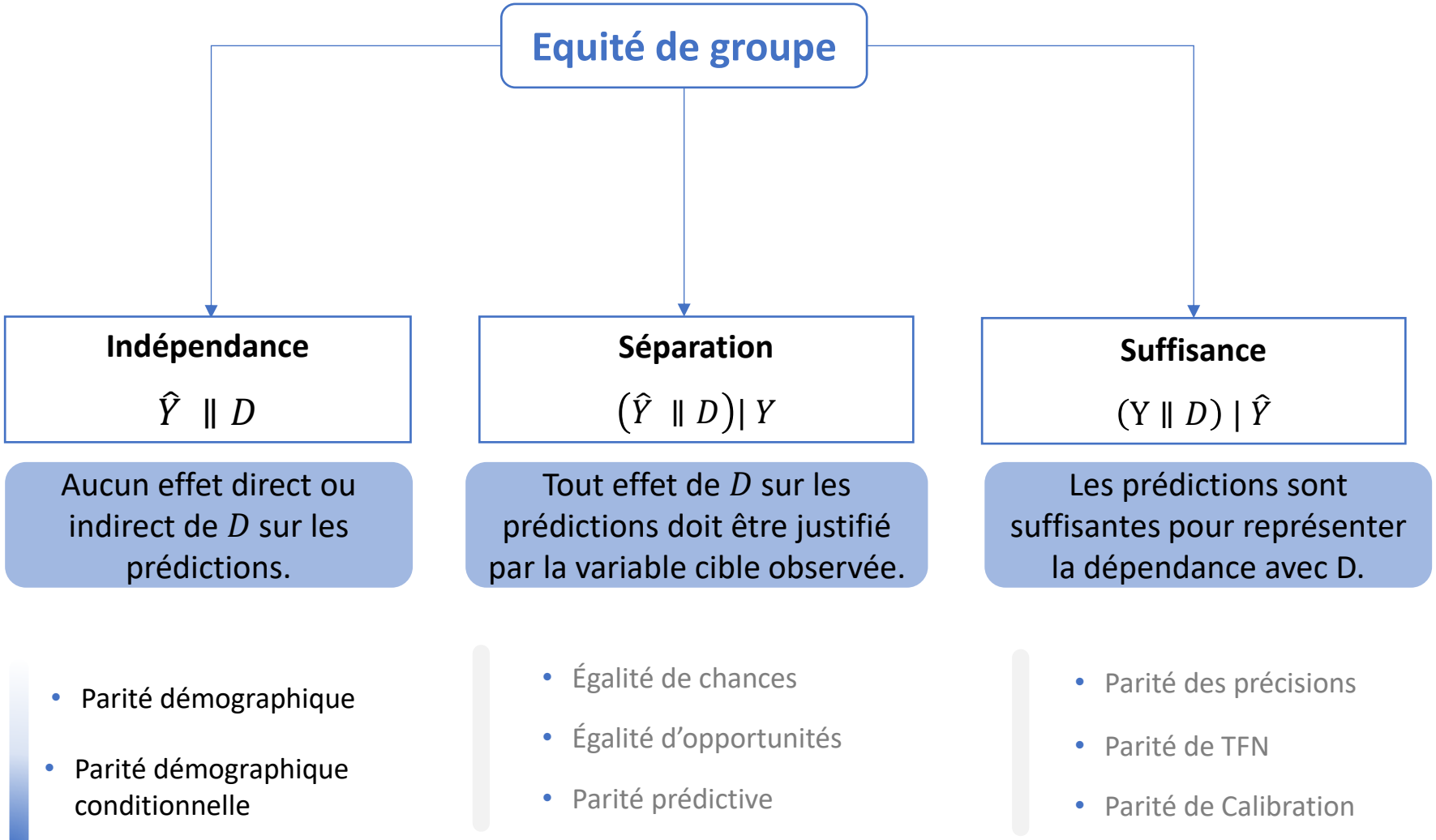
Taux de sélection global $femmes = 4/9$
44%

Taux de selection global hommes $= 4/\mathbf{10}$
40 **%**

- $\boldsymbol{R}$ **est-elle fiable** ie exempte de **biais** ?
- Comment la (les) choisir ? Pourquoi pas le niveau d'expérience de conduite ?

Bons risques Mauvais risques

Equité de groupe, 3 grands principes, plusieurs définitions



Equité individuelle, les grandes lignes

Logique 1 : « Observationnelle »

Traiter chaque individu **selon ses caractéristiques propres, sans effet direct de D.**

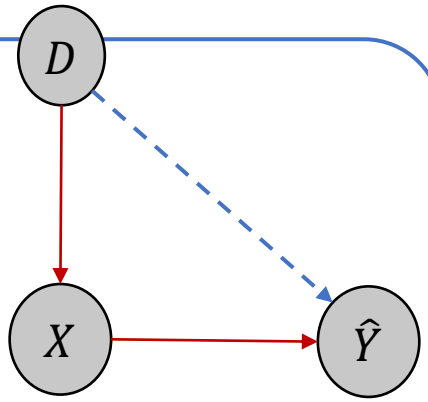
$$\widehat{Y}(X_i, D_i) \approx \widehat{Y}(X_j, D_j) \text{ si } X_i \approx X_j \quad \forall i, j$$

Solution : Exclure D du modèle $\quad \hat{Y} = f(X)$ Simple

Logique 2 : « Contrefactuelle »

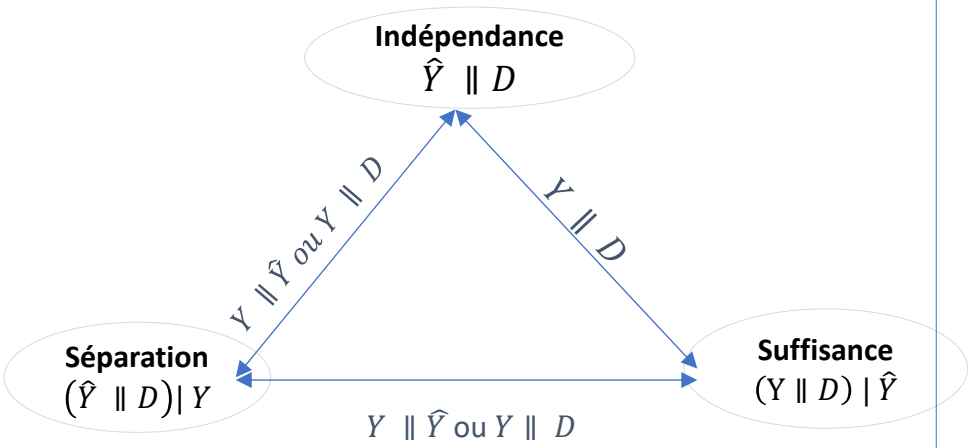
Pour chaque individu, la prédiction doit **rester inchangée si on modifie seulement sa caractéristique D**

$$\widehat{Y}_{D=d}(X_i) = \widehat{Y}_{D=d'}(X_i), \forall\, d, d'$$



Solution : Neutraliser D du *modèle causal* Complexe

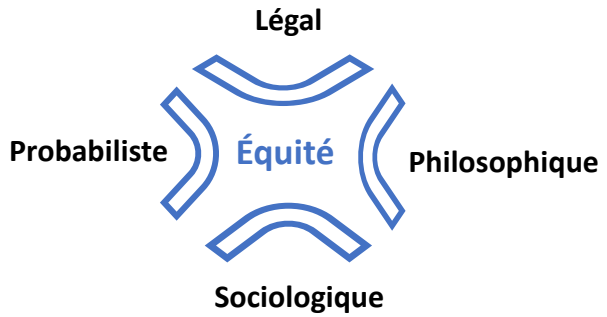
Equité, incompatibilité et enjeux



Les trois principes d'équité de groupe ne sont pas mutuellement compatibles

Arbitrage équité – performance (en général)

Nécessité de faire un choix

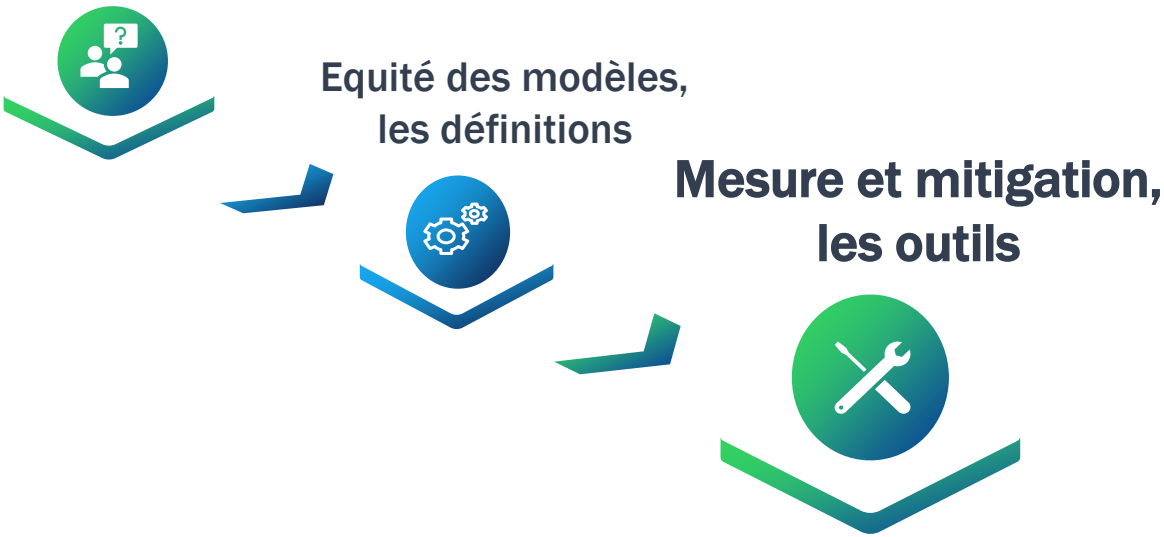


Equité, un concept multi-facettes

Tous concernés :

- Cadrage & Gouvernance,
- Dev & validation
- …
- Utilisation, pilotage

Introduction



Mesurer l'équité, les outils…

une fois le principe/l'objectif défini

Institut des ACTUAIRES

Équité de groupe

Indépendance
$\hat{Y} \perp\!\!\!\perp D$

Parité démographique

Tests statistiques

Test de Welch

But: Comparaison des moyennes prédites de 2 populations (adaptation du test de Student).

Métrique : $t = \frac{\hat{Y}_{n_1} - \hat{Y}_{n_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, s_1^2, s_2^2 variances empiriques

Limites : - Hypothèse de normalité des données
- Ne compare que les moyennes
- D binaire uniquement

Mesurer l'équité, les outils…

une fois le principe/l'objectif défini

Institut des ACTUAIRES

Équité de groupe

Indépendance
$\hat{Y} \perp\!\!\!\perp D$

Parité démographique

Tests statistiques

Test de Kolmogorov-Smirnov

But: Comparaison de 2 distributions de prédictions.

Métrique : $\mathrm{KS} = Sup_t \left| F_{n_1}(t) - F_{n_2}(t) \right|$

Limites : - Utilisable pour D binaire uniquement

- Significativité sensible à la taille de l'échantillon
($n = n_1 + n_2$ très grand → écarts faibles détectés comme significatifs)

Valeur critique : $C(\alpha)\sqrt{\frac{n_1+n_2}{n_1.n_2}}$

Mesurer l'équité, les outils…

une fois le principe/l'objectif défini

Institut des ACTUAIRES

Équité de groupe

Indépendance
$\hat{Y} \parallel D$

Parité démographique

Analyse de sensibilité 

Variance Expliquée

But : Part de variabilité des prédictions expliquée par la variable sensible.

Métrique : $EV = \frac{Var\left(\mathbb{E}\left(\hat{Y}\middle|D\right)\right)}{Var(\hat{Y})}$

Limites : - Ne capte que les effets linéaires
- Sensible à la taille relative des groupes → dilution des différences mesurées si un groupe est relativement petit

Équité de groupe

Indépendance
$\hat{Y} \parallel D$

Parité démographique

Analyse de sensibilité

Information mutuelle

But : quantifie la réduction d'incertitude sur les prédictions apportée par la connaissance de la variable sensible D.

Métrique : $*\ \mathrm{I}(\hat{\mathrm{Y}};\mathrm{D}) = h(\hat{Y}) - \sum_d P(D=d)\, h(\hat{Y}|D=d)$
$*\ h(\hat{Y}) = -\int P(\hat{y}) log P(\hat{y}) d\hat{y}$.

Limite : - Sensible à la taille relative des groupes

Diagnostic: les outils d'identification des proxys

Proxys : Variables explicatives fortement corrélées à la variable sensible

Avant modélisation

Outils statistiques précédents +

Test de Khi-deux et V-Cramer

But : Mesure l'association statistique entre 2 variables catégorielles en mesurant l'écart entre les distributions observées et attendues.

Métrique : $\chi^2 = \sum \frac{(O-E)^2}{E}$, $V = \sqrt{\frac{\chi^2}{n.\,min(k-1, r-1)}}$

k, r nb de modalités des 2 variables

O, E effectifs observés et théoriques du tableau de contingence

Limites : - Sensibilité aux nombres de catégories

Diagnostic: les outils d'identification des proxys

Proxys : Variables explicatives fortement corrélées à la variable sensible

Pendant & après modélisation

Importance par permutation

<u>Principe</u>**:** Impact de chaque variable X_j sur le niveau d'une métrique d'équité M.

$$\boldsymbol{\Delta M_j = M_0 - \mathbb{E}(M_{permut_j})}$$

+

Valeur de Shapley

<u>Principe</u> **:** Mesure la contribution individuelle d'une variable aux prédictions en moyenne.

<u>Utilisation</u> : Identifier si un proxy influence fortement le modèle.

Méthodes de mitigation – corriger l'inéquité

Principe d'équité recherché : **l'indépendance** plus précisément la **parité démographique stricte**

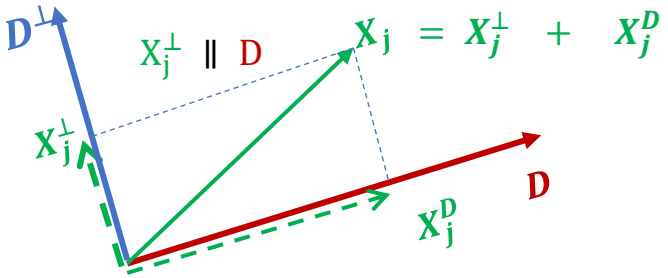
Avant modélisation

Orthogonalisation

Principe: Décorréler chacune des variables explicatives X_j ***non sensibles*** de la variable sensible D.

Méthode:
- Résidus de la régression de X_j sur D. $X_j^{\perp} = X_j - \mathbb{E}(X_j|D)$
- Entrainer le modèle sur : $X_j^* = \alpha X_j + (1-\alpha)X_j^{\perp}, \alpha \in [0,1]$

Limite : Interprétation des données transformées



Méthodes de mitigation – corriger l'inéquité

Après modélisation

Transport optimal

Principe: Transformer les prédictions en variables aléatoires indépendantes de D.

Méthode: - Choix d'une distribution cible F^+ à partir de $F_{d_1}(\hat{y}), F_{d_2}(\hat{y})$

- Transport des prédictions sur cette distribution :

$$\hat{y}^* = \widehat{F^{+-1}} \circ \widehat{F_{d_j}}(\hat{y})$$

Problème : Comment choisir ?

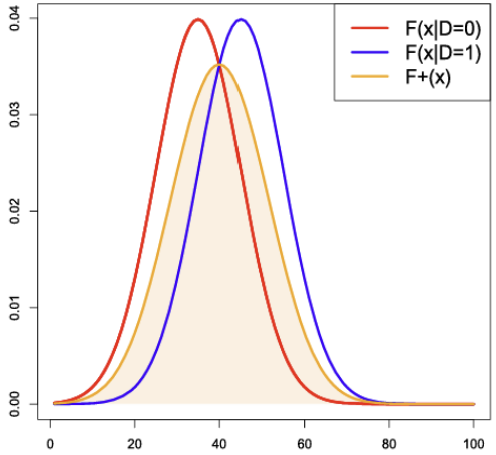
Barycentre de Wasserstein

But: Eviter la question du choix de distribution

Méthode:

- Calcul du *Transport* dans le groupe opposé

- Moyenne pondérée des deux résultats



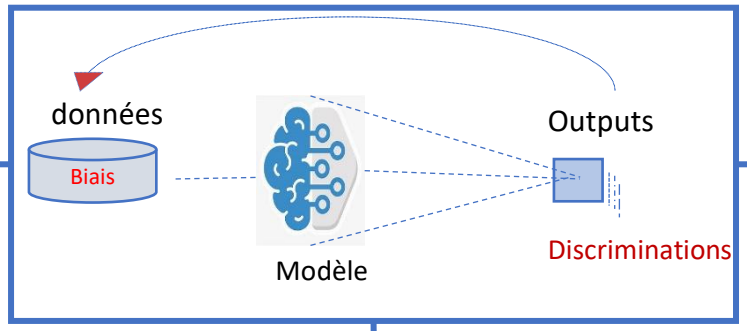
D'autres méthodes de mitigation existent

… utilisation dépendant du critère recherché

Objectif : Réduire une forme d'iniquité dans un modèle d'apprentissage supervisé

Utilisée pour :

- Indépendance
- Séparation
- Suffisance



Méthodes pré-modélisation

Correction des sources de discriminations dans les données.

- **Exclusion de la variable sensible** ●
- **Othogonalisation** ●

Méthodes in-modelling

Réduction de l'inéquité durant la calibration du modèle.

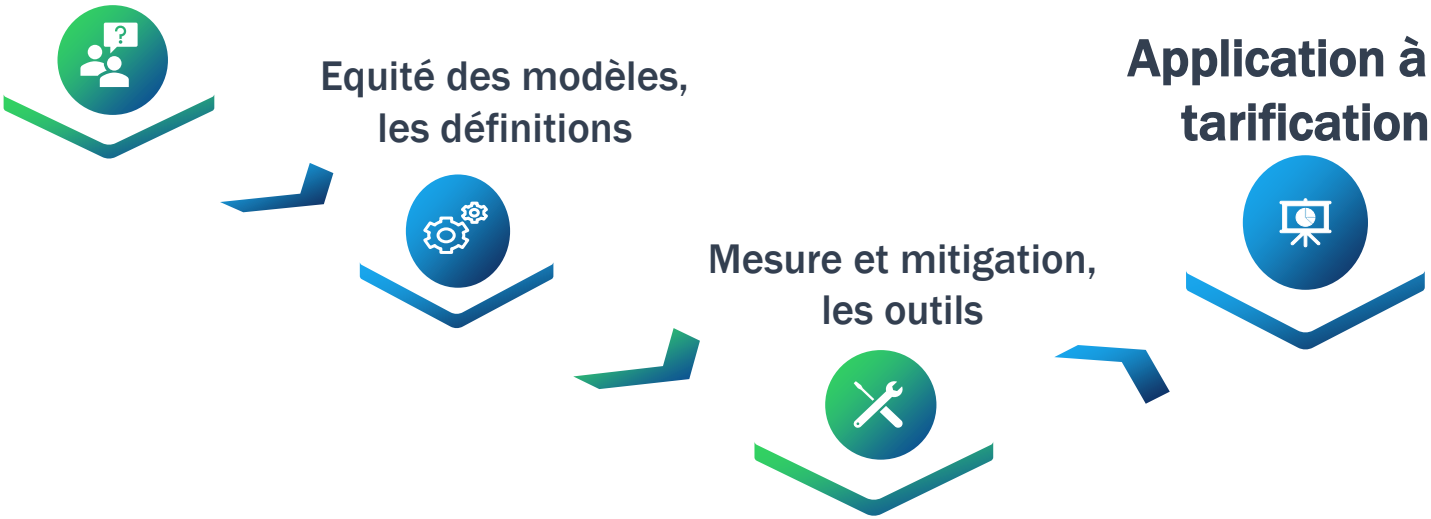
- **Ajout d'une contrainte d'équité au problème d'optimisation** ● ● ●
- **Adversarial debiasing** ● ● ●

Méthodes post-modélisation

Corriger les discriminations dans les outputs du modèle

- **Optimisation par seuils** ● ● ●
- **Barycentres de Wasserstein** ● ● ●

Introduction



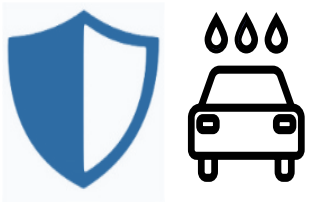
Les données d'application

❑ Le produit : assurance responsabilité civile (RC) automobile partic.

Indemnisation de tiers pour les dommages matériels et corporels causés par le véhicule assuré, conformément aux limites et conditions du contrat.

❑ Les garanties : 3 types de couverture

- **Couverture minimale** : uniquement la garantie rc
- **Omnium partielle** : garantie RC + vol, incendie, bris de glace, heurt animal et forces de la nature
- **Omnium complète** : garanties de l'omnium partielle + garantie « dégâts matériels ». → assurance auto « tous risques »



❑ La base de données d'application

- Environ **163 212 contrats;**
- **11 caractéristiques** relatives au **contrat** (type de couverture, …), à **l'assuré** (âge, genre, …), au **véhicule** assuré (puissance, âge, …);
- Source : package R '**CASdatasets**' de Chistophe Dutang, Arthur Charpentier ;
- Critère principal de choix : Disponibilité (quasi-) complète d'une variable sensible (Taux de valeurs manquantes négligeable)

❑ La segmentation des données

Apprentissage (80%)	Test (20%)
Calibration des modèles Détection des ***proxys*** potentiels	- Test et choix des modèles - Evaluation de l'indépendance

Aperçu du process global & premières informations sur le portefeuille

Process de construction et d'évaluation

01 Préparation et description des données

- Examen de valeurs manquantes
- Description unidimensionnelle des variables

02 Analyse multidimensionnelle

- Corrélation de Pearson
- Khi-deux, V de Cramer, Information mutuelle, ANOVA
- Proxys potentiels, – sensible : **Genre**

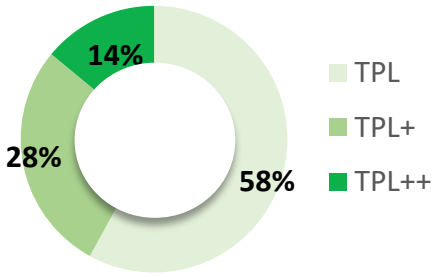
03 Modélisation et évaluation

- Hyper paramétrisation, pénalisation
- Estimation des GLM vs Light GLM
- Évaluation de l'équité.

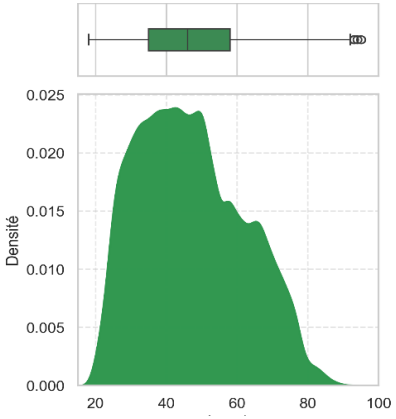
04 Correction / Réentrainement

- Identification des proxys
- Orthogonalisation des variables ou transport des résultats

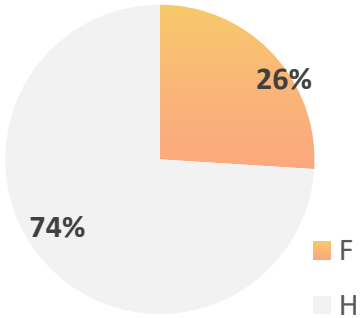
Quelques caractéristiques



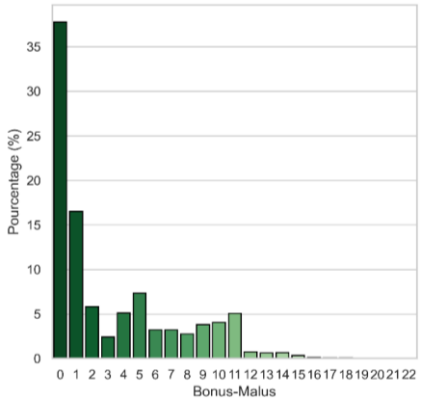
Répartition des garanties choisies



Distribution de l'âge des assurés



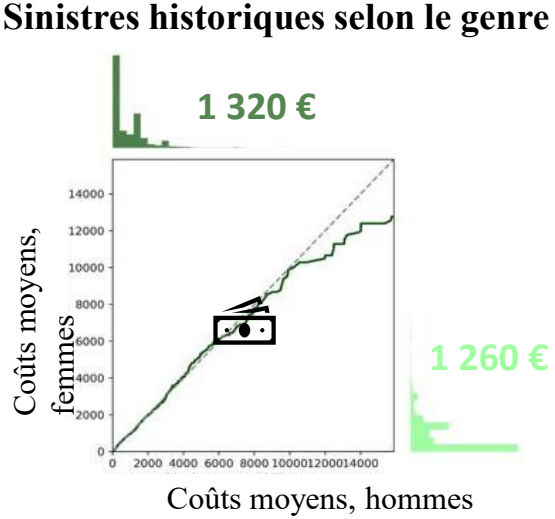
Distribution du genre des assurés

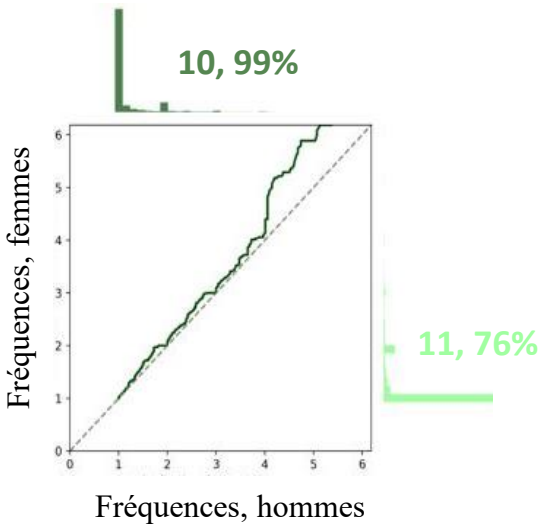


Répartition des scores de conduite

Analyses bidimensionnelles: disparités historiques ? proxys ?

Sinistres historiques selon le genre

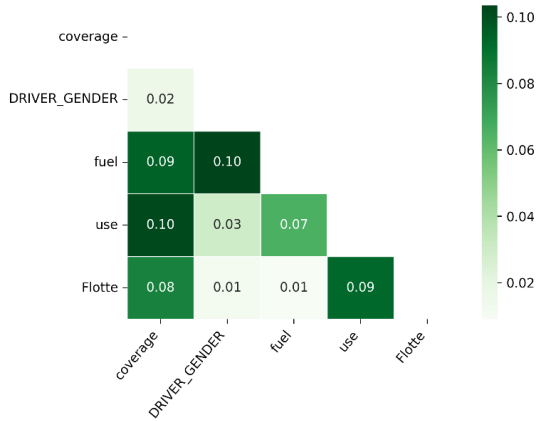




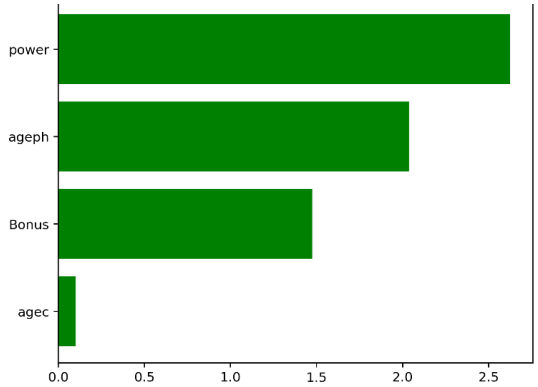
- Disparités légères, visibles principalement en queue de distribution
- Statistiquement non significatives

Y'a-t-il des proxys potentiels ?

V de Cramer



Variance expliquée par le genre



- Liaisons **toutes** statistiquement significatives
- Associations globalement de **faible intensité**

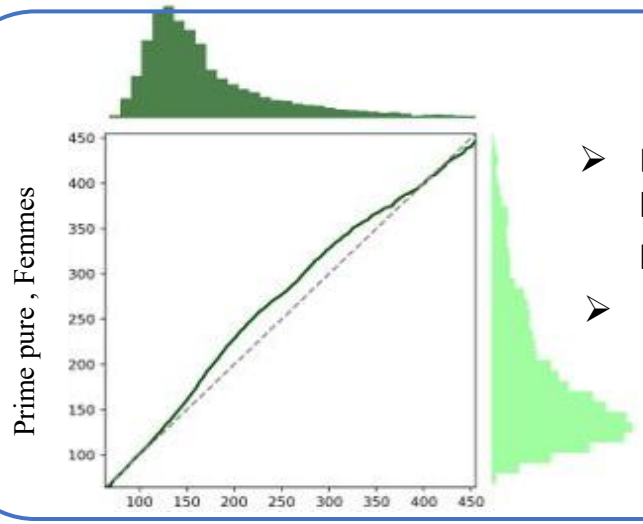
Le modèle : GLM

Performance – explicabilité – facilité de déploiement

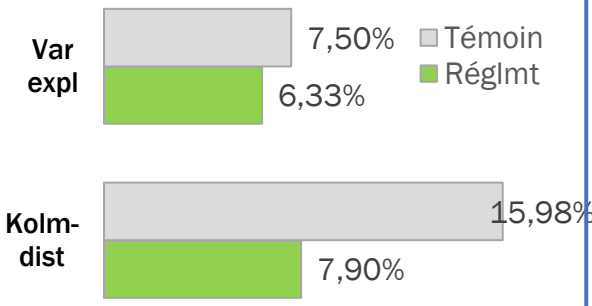
Entrainement dans 2 configurations: **Sans** et **avec** la variable sensible (« témoin »)

	Coût		Fréquence	
Modèle	Témoin	Réglemenatire	Témoin	Réglementaire
GLM				
Indice de Gini	10,19%	10,36%	22,28%	22,36%
Déviance expliquée	0,71%	0,83%	3,00%	2,98%

→ Le genre ne semble pas déterminant pour les performances.

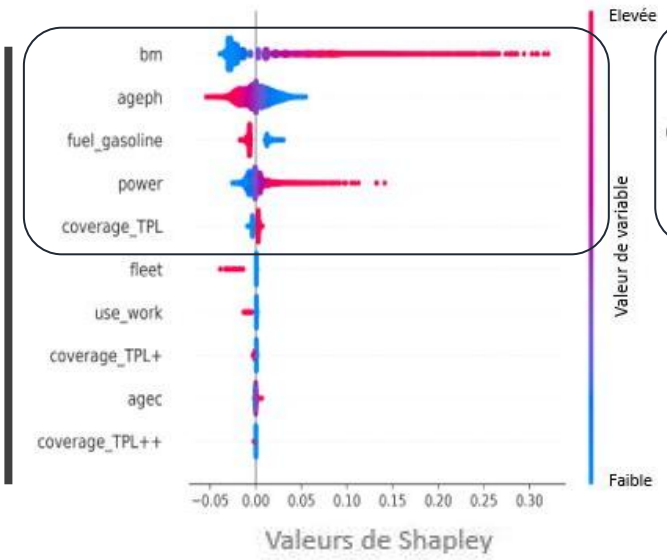


- Exclure la variable sensible réduit l'intensité des disparités de primes pures
- Cette dépendance n'est pas complètement annihilée → **des proxys ? Ou biais de modèles ?**

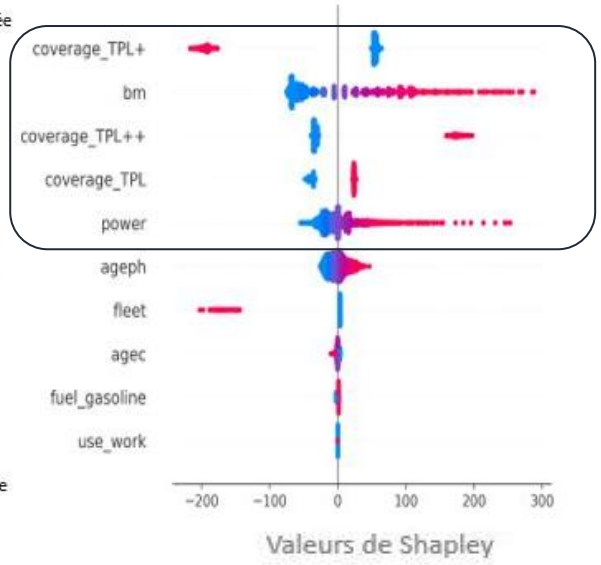


Disparités sur les prédictions – les variables légitimes

➢ **Fréquences**



➢ **Coûts**



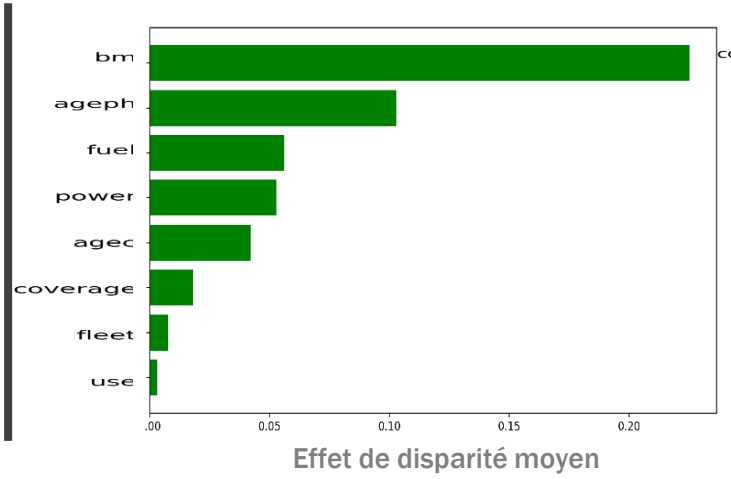
Facteurs de risque les plus déterminants dans les prédictions

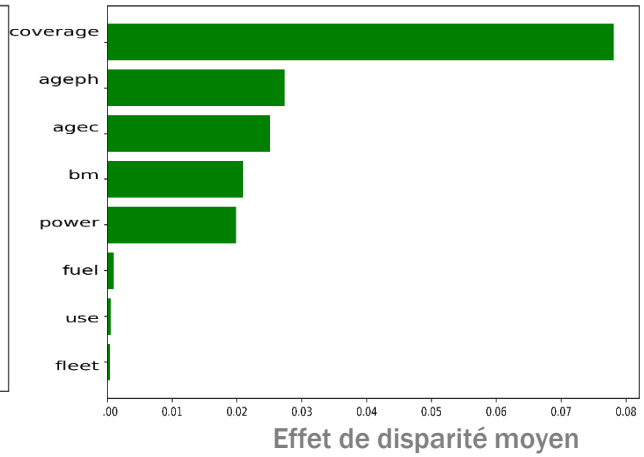
Des fréquences : score de conduite, âge du conducteur;

Des coûts : garantie, score de conduite, puissance.

Principaux facteurs de risque source de disparités dans les prédictions:

Des fréquences & coûts : score de conduite, âge du conducteur et garantie





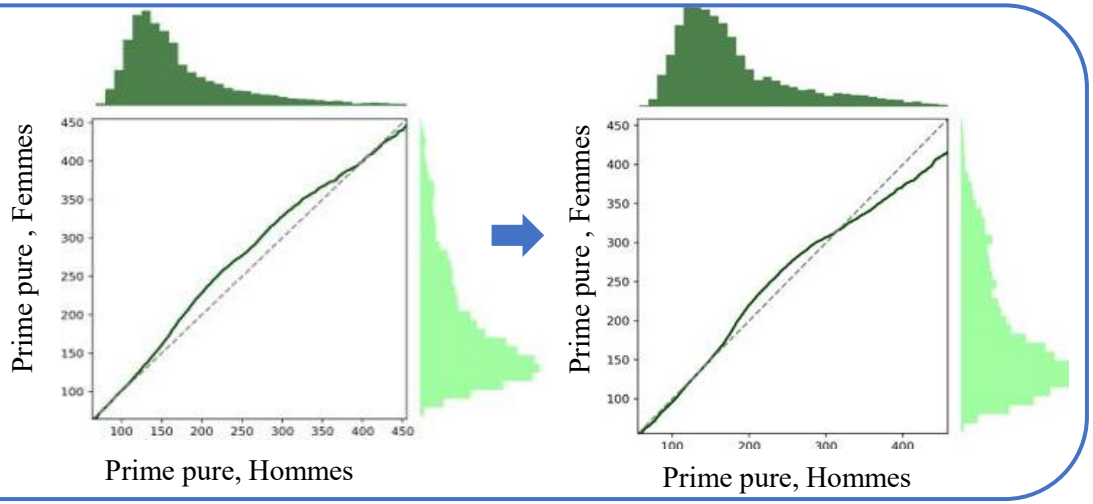
- Les facteurs de risque déterminants sont aussi les plus discriminants.
- Sont-ils vraiment discriminants s'ils représentent le choix/comportement de l'assuré ?

L'orthogonalisation des variables explicatives

Institut des ACTUAIRES

❑ **Résultat : légère amélioration de la parité démographique des primes pures**

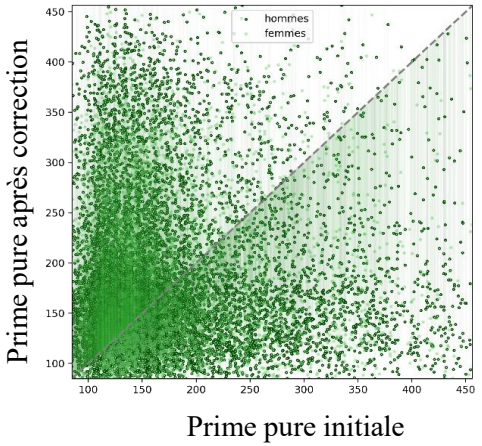
	Distance Kolmogorov	Ecart moyennes
Avant	7,9%	6,4%
Après	5,6%	1,1%



❑ **Des limites tout de même…**

- La différence des prédictions demeure significative
- Signification des nouvelles variables / facteurs de risque
- Acceptation commerciale des nouveaux « tarifs »??

Méthode pas complètement efficace



La méthode des barycentres de Wasserstein (post-modélisation)

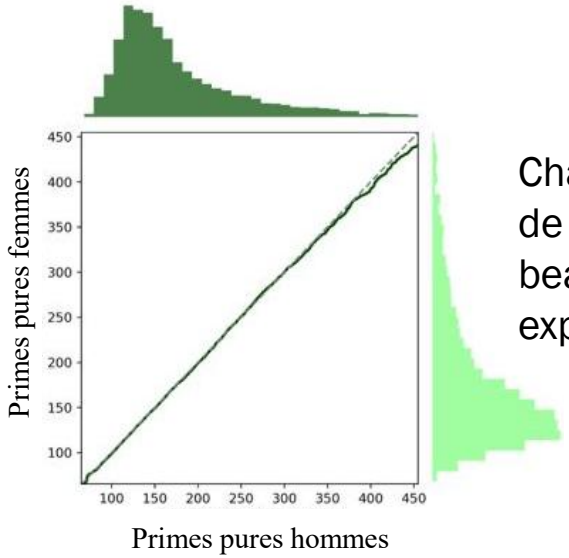
Institut des ACTUAIRES

Rappel méthode

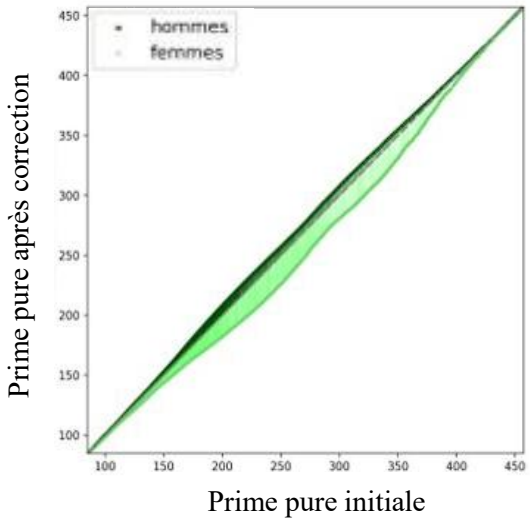
- Estimation de $\widehat{F_{d_1}}(\hat{y}), \widehat{F_{d_2}}(\hat{y})$ et de leurs inverses à partir des prédictions de l'échantillon d'entrainement;
- « Transport » de chaque prédiction $\hat{y}$ dans le groupe opposé : $\hat{y}_{tr} = \widehat{F_{d_1}^{-1}}\left(\widehat{F_{d_2}}(\hat{y})\right)$ (échantillon test);
- Moyenne pondérée des deux : $\hat{y}^* = w_1\hat{y} + w_2\hat{y}_{tr}$

Résultat

Parité démographique quasi-stricte vérifiée



Changements de « primes » beaucoup plus explicables



Introduction

Equité des modèles, les définitions

Mesure et mitigation, les outils

Application à la tarification

Conclusion

Conclusion

Messages clés

- Nécessité de choix pour le principe d'équité à rechercher;
- Besoin de données sur la/les variable sensible;
- Exclure la variable sensible d'un modèle n'assure pas nécessairement l'indépendance;
- Orthogonaliser les variables n'est pas la meilleure solution pour assurer l'indépendance;
- La correction post-modélisation des barycentres peut fournir des résultats intéressants

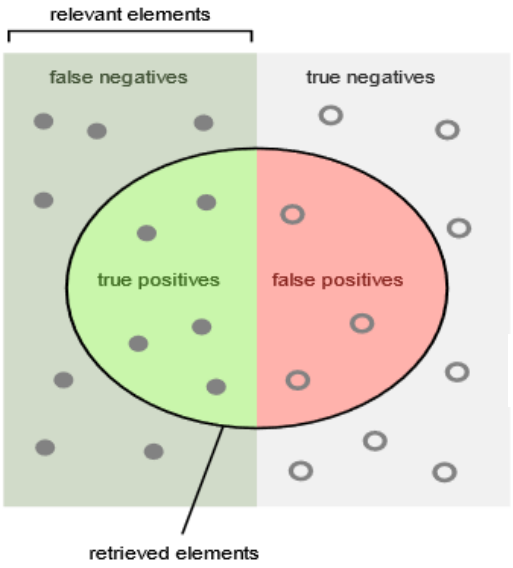
Limites

- × Disparités observées faibles
- × Variable sensible uniquement binaire et unidimensionnelle;
- × Analyses et diagnostics portées uniquement sur le principe d'indépendance;
- × Equité à l'échelle individuelle non examinée;
- × Absence d'analyses business poussées : primes commerciales et effets d'élasticité de la mitigation par exemple

Merci,

Pour votre attention

Useful when **Recall** or **False positive rate** are the most sensitive performances



$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

$$\text{FPR} = \frac{\text{false positives}}{\text{true negatives}}$$

Example where one may prefer :

- ***Recall – equality of opportunity : In a recruitment scenario,***

we want to give the same opportunity - $\hat{Y} = 1$ – to men/women who deserve ($Y = 1$) ;

- ***False positive rate : In a claim scoring,***

False positives (accept non acceptable claims) – ie financial cost – could be more dangerous than false negatives – immediate reputation cost.

In general, more useful when exhaustiveness is the most important property.

- Is the target variable Y trustable / free from bias ?
- What if Y is only observed for some predictions ? (In credit scoring for example)
- What if ***precision*** is the most sensitive metric ?

Formule de la valeur de Shapley

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{j\}) - f(S) \right]$$

- ϕ_j : contribution de la variable j à la prédiction
- F : ensemble de toutes les variables explicatives
- S : sous-ensemble de variables ne contenant pas j
- $f(S)$: prédiction du modèle avec les variables de S
- Le terme $f(S \cup \{j\}) - f(S)$ mesure **le gain de performance** quand on ajoute j.
- Les coefficients de pondération assurent que toutes les combinaisons possibles sont prises en compte de manière équitable.