Actulab

Comment rendre l'assurance plus équitable

Rapport technique présenté au jury

Travail réalisé par l'équipe CDA
Mathieu Bazinet,
Simon-Olivier Lépine, Liliane Ouedraogo et Michaël Rioux

Le 23 novembre 2022

Résumé

Nous étudions séparément la fréquence de réclamation avec *equalized odds* et la sévérité des réclamations avec une métrique adaptée à l'atelier d'innovation. Selon le cas, nous pénalisons une fonction de perte donnée pour minimiser un critère d'inéquité choisi, de façon similaire au terme ajouté pour la parcimonie de la régularisation L1 (LASSO). Nous proposons la métrique d'équité **Parité Actulab par Quartiles** (PAQ) ainsi qu'une conjecture portant sur l'équité du produit de modèles prédictifs.

Nous montrons que l'approche proposée permet d'assurer une meilleure équité des modèles. La relation entre la pénalisation sur la performance prédictive des modèles dépend de la métrique choisie. Tandis qu'elle est claire dans le cas d'une régression logistique pénalisée par la disparité moindre, elle est beaucoup moins claire dans le cas d'une régression gamma pénalisée par notre métrique PAQ.

L'ensemble du code utilisé est disponible sur GitHub. [1]. On y retrouve également les diapositives ainsi que ce rapport.

Glossaire des notations utilisées

Y :	Variable binaire de classification
$\hat{Y}$:	Prédiction de Y
M :	Variable continue du montant réclamé
$\hat{M}$:	Prédiction de M
$\{+, -\}$:	Décision positive (+) ou négative (-), e.g. présence de réclamation (+)
A :	Attribut protégé
δ :	Décalage ($\geqslant 0$)
$\mathcal{D}$:	Décalage cumulatif ($\geqslant 0$)
λ :	Facteur de pénalisation
ℓ :	log-vraisemblance
$\boldsymbol{\beta}$:	Coefficients de régression

1. `https://github.com/MathieuBazinet/Actulab`

Introduction

Une prime est raisonnablement et légitimement discriminatoire s'il s'agit d'une estimation actuarielle de la valeur attendue (espérée) de tous les coûts associés à un transfert de risque individuel. Si aucune caractéristique protégée n'influence disproportionnellement le calcul d'une telle prime, alors elle est dite équitable (Statement of Principles Regarding Property and Casualty Insurance Ratemaking).

La prime peut être déterminée à partir des prédictions de fréquence de réclamation ou de celles de la sévérité du montant réclamé. Afin que le modèle utilisé pour faire de telles prédictions soit équitable, il faut minimiser les disparités des prédictions d'un modèle selon les valeurs d'un attribut protégé. La disparité peut se manifester de deux façons. La première est une disparité d'impact, et la deuxième est la disparité de traitement. La première est plus difficile à corriger (Khoury, 2022). La disparité d'impact intervient lorsque l'algorithme traite tous les individus de manière identique, mais que l'effet des décisions est différent selon les valeurs de l'attribut protégé. La disparité de traitement se produit lorsque le modèle traite différemment les individus ayant des valeurs différentes d'un attribut protégé. Nous nous concentrons sur le deuxième cas de disparité.

La détermination d'une prime est basée sur 1) le risque d'un individu et 2) la tarification du risque. Nous nous concentrons sur la prévision du risque d'un individu. Nous cherchons donc à quantifier le risque de façon équitable.

L'estimation d'un modèle prédictif peut avoir un biais (statistique) et un décalage (biais éthique). On veut diminuer le décalage. Cependant, diminuer le niveau de discrimination du modèle peut diminuer ses performances prédictives. Alors que plusieurs définitions de « performances prédictives » existent, il faudra d'abord définir mathématiquement des métriques pour mesurer l'équité (section 2).

1 Sur quels attributs est-il correct de discriminer

Un attribut protégé, ou une caractéristique protégée, ne doit pas être utilisé pour faire les prédictions afin d'éviter la discrimination directe. De plus, il est important de chercher à éviter la discrimination indirecte, qui apparaît lorsqu'on discrimine par rapport à un attribut protégé par l'entremise de variables qui y sont corrélées.

La charte des droits et libertés de la personne mentionne que « Toute personne a droit à la reconnaissance et à l'exercice, en pleine égalité, des droits et libertés de la personne, **sans distinction, exclusion ou préférence fondée sur** la race, la couleur, **le sexe**, l'identité ou l'expression de genre, la grossesse, l'orientation sexuelle, l'état civil, **l'âge sauf dans la mesure prévue par la loi**, la religion, les convictions politiques, la langue, l'origine ethnique ou nationale, la condition sociale, le handicap ou l'utilisation d'un moyen pour pallier ce handicap. Il y a discrimination lorsqu'une telle distinction, exclusion ou préférence a pour effet de détruire ou de compromettre ce droit. » La loi canadienne sur les droits de la personne mentionne quelque chose de similaire.

Cette même charte mentionne aussi à l'article 20.1 que dans un contrat d'assurance, une distinction

sur l'âge est « réputée non discriminatoire lorsque son utilisation est légitime et que le motif qui la fonde constitue un facteur de détermination de risque, basé sur des données actuarielles. »

Néanmoins, à des fins illustratives, il semble justifié d'utiliser l'âge et le sexe comme attributs protégés dans les données. Il est arrivé aux États-Unis que des états interdisent la discrimination sur ces attributs. Les variables gender (variable binaire) et agecat (variable catégorielle à 6 modalités) seraient des attributs protégés dans le jeu de données dataCar. La variable area pourrait possiblement être une variable menant à de la discrimination indirecte étant donné que certains secteurs sont plus pauvres, etc.

2 Définition mathématique de l'équité

Soit

- Y une variable binaire de classification qui vaut 1 si l'individu a fait au moins une réclamation et 0 sinon ;
- M une variable continue qui représente le montant réclamé ;
- A les attributs protégés, par exemple le sexe, la religion, l'ethnie, etc. ;
- X les attributs non protégés ;
- $\hat{Y}$ et $\hat{M}$ les estimations prédites par un modèle entraîné avec les attributs X ;
- R le coût réel de risque ;
- δ le décalage qui prend généralement une valeur supérieure à 0. Plus le décalage est petit, plus le modèle est équitable.

2.1 Mesures d'équités pour variables discrètes

Pour chacune des mesures ci-dessous, il est évident que l'égalité ne sera en pratique jamais respectée. Il existera toujours un décalage $\delta > 0$, qu'on cherche à minimiser. Pour simplifier la notation, au lieu d'écrire une mesure d'équités entre deux probabilités P_1 et P_2 sous la forme $|P_1 - P_2| < \delta$, on l'écrit sous la forme $P_1 = P_2$. Cette notation est intuitive et est utilisée dans la littérature. Nous l'utiliserons donc dans ce rapport.

Parité démographique (PD)

$$\mathbb{P}\left[\hat{Y} = + \mid A = a\right] = \mathbb{P}\left[\hat{Y} = + \mid A = b\right] \quad \forall a, b \in A. \tag{1}$$

Exemple : Le modèle doit prédire la même probabilité de réclamation ($Y = 1$) peu importe le sexe ($A = a$ ou $A = b$). Si A a plus que deux modalités, l'égalité doit être vrai pour tout $a, b \in A$.

Problème de cette métrique : Mosley and Wenman (2022) expliquent le problème en montrant qu'il serait possible d'attribuer aléatoirement les prédictions de réclamation pour les femmes, en s'assurant que la probabilité que $\hat{Y} = 1$ reste la même que pour les hommes.

Lorsqu'on prédit $\hat{Y}$, on pourrait plutôt s'assurer que le modèle est équitable en vérifiant que les taux de vrais positifs soient égaux indifféremment de l'attribut protégé A.

Égalité des chances (EC)

$$\mathbb{P}\left[\hat{Y} = + \mid Y = +, A = a\right] = \mathbb{P}\left[\hat{Y} = + \mid Y = +, A = b\right] \quad \forall a, b \in A. \tag{2}$$

Une version plus stricte de l'égalité des chances est de s'assurer que les taux de faux positifs soient aussi égaux indifféremment de l'attribut protégé A :

Equalized Odds (EO)

$$\mathbb{P}\left[\hat{Y} = + \mid Y = y, A = a\right] = \mathbb{P}\left[\hat{Y} = + \mid Y = y, A = b\right] \quad y \in \{-, +\}, \quad \forall a, b \in A. \tag{3}$$

2.2 Mesures d'équités pour variables continues

Les précédentes définitions ne sont clairement pas adaptées aux variables continues, donc lorsque l'on cherche à prédire le montant réclamé M. Il est donc nécessaire de trouver une autre mesure d'équité.

Parité actuarielle par groupe (PAG)

La parité actuarielle par groupe (PAG) est une métrique proposée par (Dolman and Semenovich, 2018) pour étendre Equalized Odds au contexte d'une variable à prédire continue. L'exemple qu'ils utilisent est la prime d'assurance, mais nous l'utiliserons pour la sévérité (le montant) d'une réclamation en assurance.

$$\left|\mathbb{E}\left[\hat{M} \mid K(n-1) \leq R < Kn, A = a\right] - \mathbb{E}\left[\hat{M} \mid K(n-1) \leq R < Kn, A = b\right]\right| < \delta \tag{4}$$

pour un certain $K \in \mathbb{R}^+$, $\forall n \in \mathbb{N}$ et $\delta \geq 0$.

Pour contextualiser avec les données de dataCar, nous utilisons comme estimation du risque le montant réellement réclamé : $M = \hat{R}$. Dans ce cas particulier, la prédiction $\hat{M}$ est une estimation du vrai risque R, alors que ce n'est pas nécessairement le cas dans la version proposée par Dolman and Semenovich (2018), où M est une fonction monotone de $\hat{R}$ qui pourrait être le montant de la prime chargée à l'individu en fonction de la meilleure estimation possible du vrai risque.

Cette définition de l'équité est imparfaite dans le contexte de l'Actulab. Un K trop grand va mener à une perte de l'information, puisqu'on aura des groupes trop disparates. Un K trop petit va mener à une augmentation du temps de calcul et va potentiellement mener à calculer la moyenne sur des groupes vides. Comme on s'attend à ce que les données soient représentatives de la population, on ne devrait pas avoir un R plus petit que $\min(M)$ ou plus grand que $\max(M)$. On propose alors la métrique d'équité suivante :

Parité Actulab par quartile (PAQ)

$$\frac{\left|\mathbb{E}\left[\hat{M} \mid Q_\alpha \leq R < Q_{\alpha+\frac{1}{4}}, A = a\right] - \mathbb{E}\left[\hat{M} \mid Q_\alpha \leq R < Q_{\alpha+\frac{1}{4}}, A = b\right]\right|}{Q_{\alpha+\frac{1}{4}} - Q_\alpha} < \delta \tag{5}$$

$\forall \alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ et $\delta \geq 0$.

Nous avons décidé d'utiliser les quartiles, puisque dans le cadre de l'atelier d'innovation Actulab, nous utilisons la base de données dataCar. La moitié des réclamations sont pour un montant de moins de 1000$, tandis que M varie entre 200$ et plus de 55000$. On obtient ainsi des groupes contenant le même nombre d'individus avec des risques semblables. Bien entendu, il serait possible d'utiliser n'importe quels quantiles. Notre solution tentera de minimiser le décalage lié à la disparité moindre dans le cas binaire et la parité Actulab par quantiles dans le cas continu.

3 Démarche proposée

Pour éviter la discrimination directe, nous proposons un modèle qui n'utilise pas d'attributs protégés puisqu'on cherche à éviter la discrimination directe. Toutefois, simplement retirer les attributs protégés du modèle sans ajustement supplémentaire n'est pas une bonne approche. Dans la littérature, cette méthode se nomme *fairness through unawareness* et est réputée comme une des moins bonnes méthodes pour gérer l'équité. En effet, il existe souvent des variables corrélées à l'attribut protégé qui vont permettre au modèle de discriminer indirectement.

Revue de l'état de l'art

Depuis 2018, Microsoft, IBM et Google maintiennent une riche documentation des méthodes et des métriques permettant de minimiser et de mesurer la discrimination indirecte. On retient notamment les méthodes suivantes. La repondération est une méthode de pré-apprentissage proposée par Kamiran and Calders (2012) qui permet d'augmenter l'impact des classes d'individus mal représentés sur la formation du modèle. La régularisation d'information mutuelle est une méthode d'apprentissage machine proposée par Kamishima and al. (2012) qui permet de réduire la relation entre les prédictions et les attributs protégés. La correction contradictoire de l'inéquité est une méthode d'apprentissage machine proposée par Zhang and al. (2018) où l'on entraine le modèle à « tromper » un « discriminateur » qui essaie de déterminer à partir de la prédiction du modèle si l'individu appartenait à un groupe protégé où non. La rejection des décisions incertaines est une méthode de post-apprentissage proposée par Kamiran and al. (2012) qui favorise les groupes non privilégiés et défavorise les groupes privilégiés en modifiant les prédictions trop proches de la frontière de classification. L'utilisation d'une méthode plutôt qu'une autre dépend du contexte d'application et des définitions d'équités choisies : il n'y a pas encore de consensus.

Notre méthode fait partie des corrections de discrimination qui corrigent directement le modèle afin de minimiser la discrimination. Plus précisément, nous pénaliserons une fonction de perte donnée

pour minimiser un critère d'inéquité choisi.

Pénalisation de l'apprentissage d'un GLM

Dans le cadre de ce travail, nous utiliserons des modèles linéaires généralisés (GLM) pour leur interprétabilité et puisqu'ils sont bien connus par les actuaires.

En général, en supposant l'indépendance entre les observations de l'échantillon $\boldsymbol{y} = (y_1, \ldots, y_n)$, on cherche à maximiser la vraisemblance des données, soit

$$L(\boldsymbol{\beta}, \phi; \boldsymbol{y}, \boldsymbol{x}) = \prod_{i=1}^{n} f(\boldsymbol{\beta}, \phi; y_i, x_i). \tag{6}$$

Il est avantageux, particulièrement lorque f est une distribution de la famille exponentielle, d'utiliser la log-vraisemblance :

$$\ell(\boldsymbol{\beta}, \phi; \mathbf{y}, \boldsymbol{x}) = \sum_{i=1}^{n} \log\left(f(\boldsymbol{\beta}, \phi; y_i, x_i)\right). \tag{7}$$

Puisque le but du projet est de trouver le meilleur modèle permettant d'éviter la discrimination tout en obtenant des résultats de prédictions intéressants, on propose d'entraîner le modèle en ajoutant un décalage à la log-vraisemblance. Durant l'entraînement, le modèle devra chercher à maximiser la log-vraisemblance tout en minimisant le décalage éthique, ce qui devrait mener à un compromis performance-équité. En résulte le problème de minimisation suivant :

$$\min_{\boldsymbol{\beta}} -\ell(\boldsymbol{\beta}, \phi; \boldsymbol{y}, \boldsymbol{x}) + \lambda \mathcal{D}\left(\boldsymbol{\beta}, \phi; \boldsymbol{y}, \boldsymbol{x}\right), \tag{8}$$

où λ est un paramètre de pénalité qui détermine l'importance entre l'adéquation aux données (vraisemblance) et la mesure d'équité ($\mathcal{D}(\cdot)$). Ce genre de pénalisation est assez typique, puisqu'on l'utilise entre autres pour les régressions LASSO et Ridge.

Dans le cas où $\lambda = 0$, alors il n'y a aucune pénalisation et on obtient le même modèle qu'avec la maximisation de la vraisemblance. Plus λ est grand, plus le poids mis sur la pénalité est important. Lorsque λ tend vers l'infini, on obtient un modèle qui s'entraîne pour être parfaitement équitable sans tenir compte de la vraisemblance sur les données.

Ce genre d'approche est similaire à celle proposée par Williamson and Menon (2019). Toutefois, l'idée de pénalisation n'est pas dérivée de la même façon et n'est pas aussi flexible que notre approche. En effet, notre cadre théorique ainsi que notre implémentation peuvent facilement être adaptés si l'on souhaite utiliser d'autres définitions d'équité. En fait, n'importe quelle méthode d'apprentissage supervisé entraîner par l'optimisation d'une fonction pourrait bénéficier de notre approche. Nous nous sommes toutefois restreints aux modèles linéaires généralisés.

Décalages proposés pour la pénalisation

On propose deux décalages cumulatifs liés aux définitions de l'équité choisie. Tout d'abord, le décalage cumulatif basé sur **Equalized Odds** dans le cas d'une variable binaire et d'une variable protégée catégorielle :

$$\mathcal{D}_{\text{EO}} := \sum_{y\in\{+,-\}} \sum_{\substack{a,b\in A \\ a\neq b}} \left| \mathbb{P}\left[\hat{Y} = + \mid Y = y, A = a\right] - \mathbb{P}\left[\hat{Y} = + \mid Y = y, A = b\right] \right| \tag{9}$$

Ensuite, le décalage cumulatif basé sur la métrique PAQ dans le cas d'une variable réponse continue et d'une variable protégée catégorielle :

$$\mathcal{D}_{\text{PAQ}} := \sum_{\alpha\in\mathbb{Q}} \sum_{\substack{a,b\in A \\ a\neq b}} \frac{\left| \mathbb{E}\left[\hat{M} \mid Q_\alpha \leq R < Q_{\alpha+\frac{1}{4}}, A = a\right] - \mathbb{E}\left[\hat{M} \mid Q_\alpha \leq R < Q_{\alpha+\frac{1}{4}}, A = b\right] \right|}{Q_{\alpha+\frac{1}{4}} - Q_\alpha} \tag{10}$$

avec R le vrai risque, $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$, $Q_{\frac{i}{4}}$ le i-ème quartile et la notation spéciale de $Q_0 = \min R$ et $Q_4 = \max R$.

Dans le cas où l'on veut prédire si la personne va faire une réclamation, on peut utiliser une régression logistique. Dans le cas où l'on veut prédire le nombre de réclamations qu'une personne va faire, on peut utiliser une régression poisson. Dans le cas où on veut prédire le montant de réclamation, on peut utiliser une régression gamma ou un modèle de fréquence-sévérité.

On espère voir dans nos résultats que le fait d'entraîner les modèles avec un biais pour l'équité devrait obtenir des modèles plutôt équitables.

Dans le cas où l'on veut prédire le risque d'une personne, on peut utiliser une régression Poisson-Gamma. On propose d'utiliser un modèle entraîné pour la classification et de le multiplier avec un modèle qui fait de la régression. Toutefois, dans les données fournies, on réalise que seulement 0.4% des données ont plus d'une réclamation. Nous avons donc utilisé une régression logistique pour le modèle de fréquence.

Le modèle fourni sera donc sous la forme : $f_{logistique}(x)f_{gamma}(x)$. Théoriquement, les deux modèles devraient être équitables, ce qui devrait donner un modèle équitable. Toutefois, il n'est pas impossible que des biais soient exploités par la multiplication des deux modèles et qu'on obtienne un modèle qui n'est pas équitable.

Nous posons ainsi la conjecture Actulab :

Conjecture Actulab. *Le produit de deux prédictions équitables est généralement équitable.*

On peut voir le produit d'un modèle de classification avec un modèle de régression comme une décision en deux étapes. Premièrement, on prédit grâce à un modèle de classification équitable si le client va faire une réclamation. Deuxièmement, si l'on prédit que le client va faire une réclamation, on prédit grâce à un modèle de régression équitable le montant de la réclamation.

Puisque chaque décision est prise de façon équitable, nous faisons la conjecture que la combinaison de ces deux décisions devrait être équitable. Pour cette raison, nous allons présenter l'équité du modèle de régression logistique et du modèle gamma. L'implémentation en python permet également de gérer les régressions de Poisson, mais ce ne sera pas présenté ici.

4 Résultats des expériences

Tout à l'exception de la fonction de minimisation a été codé en python à la main. Les expériences ont été faites en R et en Python.

Pour comparer nos résultats, nous considérons ces trois modèles :

- Un modèle direct qui tient compte de l'attribut protégé et qui fait donc une discrimination directe.
- Un modèle indirect qui ne tient pas compte de l'attribut protégé et qui fait possiblement de la discrimination indirecte.
- Un modèle pénalisé entraîné en suivant la démarche proposée.

Pour évaluer la performance prédictive du modèle nous nous évaluer la sensibilité et l'AUC (Area Under the Curve) lorsque la variable réponse est binaire et le NRMSE (*Normalized Root-Mean-Square Error*) lorsque la variable réponse est continue.

Puisqu'il y a très peu de personnes avec plus d'une réclamation, comme on peut le constater au tableau 1, nous construisons un modèle logistique pour détecter la présence d'au moins une réclamation.

Tableau 1 – Répartition du nombre de réclamations dans les données dataCar

Nombre de réclamations	0	1	2	3	4
Pourcentage cumulatif	93.19	99.58	99.971	99.998	100

Nous comparons le modèle de discrimination directe, celui de discrimination indirecte (*fairness through unawareness*) et les différents modèles pénalisés. La figure 1 présente les résultats. Les courbes pour *Equalized Odds* sont en fait le ratio de la probabilité des hommes divisée par celle des femmes. Un ratio de un signifie un équité parfaite selon la métrique. Les lignes rouges correspondent à la « limite » parfois utilisée du ratio de 4/5 et 5/4 (Besse, 2020).

On constate que plus le paramètre λ augmente, plus le ratio des probabilités tend vers 1, ce qui est rassurant, puisque c'est l'objectif de la pénalisation. On voit que le modèle est plus équitable (sur les données de validation).

En termes de performances prédictives, aucun modèle pénalisé ne fait mieux que les deux modèles de référence. On note cependant que les valeurs de l'AUC restent assez stables. On observe une légère diminution de l'AUC lorsque λ augmente. La sensibilité est plus variable. La valeur du paramètre

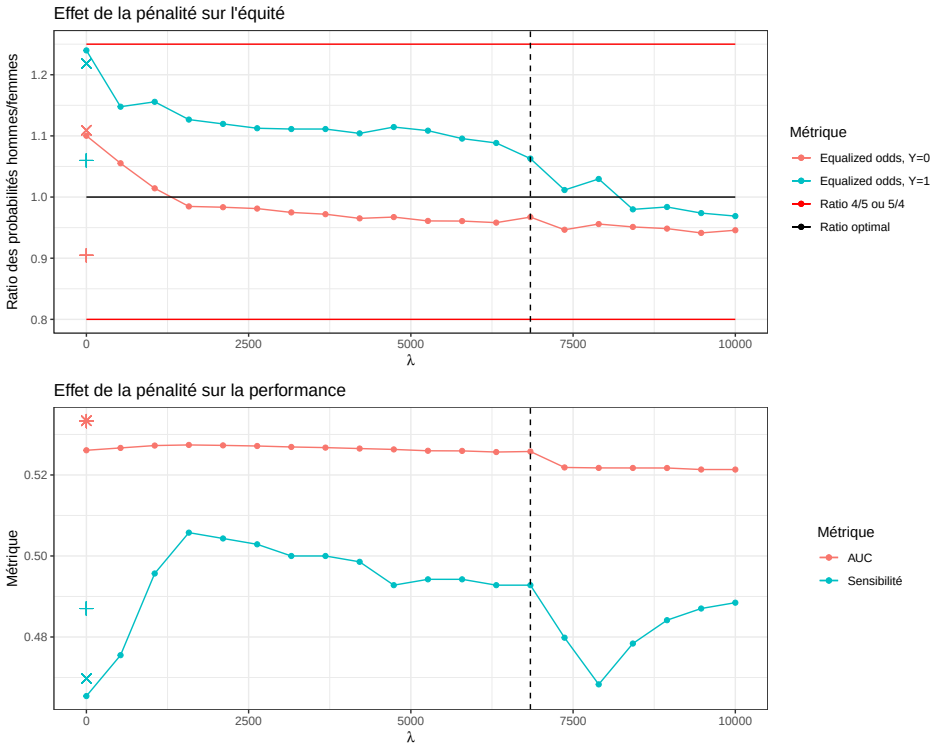

FIGURE 1 – Effet de la pénalité *Equalized odds* sur l'équité et les performances pour la régression logistique

$+$ représente le modèle de discrimination directe. $\times$ représente le modèle de discrimination indirecte.

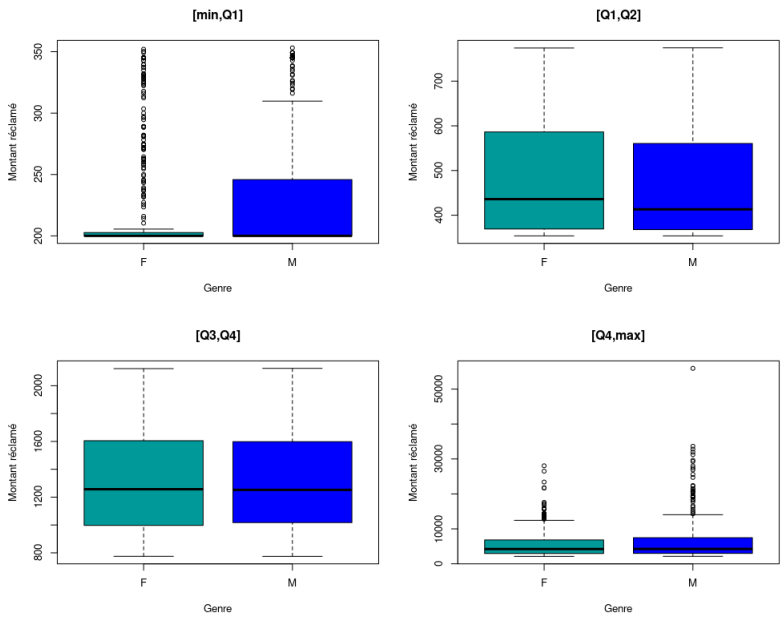

FIGURE 2 – Montant de la réclamation par genre et par quantile

de régularisation peut être sélectionnée pour maximiser la sensibilité, mais n'est pas nécessairement associée à une valeur plus faible ou élevée de λ.

Le modèle associé à la ligne pointillée en noir pourrait être utilisé. L'AUC est un peu plus faible, mais la sensibilité est plus élevée et l'équité selon la métrique choisie l'est également.

Maintenant, concentrons-nous sur la modélisation du montant réclamé. Rappelons que nous nous concentrons sur l'équité du montant de réclamation prédit, indépendamment de l'équité de la prédiction de la réclamation que nous venons de regarder. Nous justifions cela par la **conjecture actulab**. En réalité, cela permet de simplifier les analyses.

Pour faire un GLM avec distribution gamma, nous considérons seulement les montants réclamés supérieurs à 0. Toutes les analyses qui suivent excluent les montants de réclamation nuls.

La figure 2 présente la répartition des montants réclamés pour les hommes et les femmes selon les quantiles. De manière générale, la répartition des montants réclamés est similaire entre les sexes, sauf entre le quatrième quartile et le montant maximal (en bas à droite) où certains hommes ont un montant de réclamation plus élevé.

Nous avons ensuite entraîné deux modèles de référence, encore une fois un modèle qui utilise le genre et un autre qui ne l'utilise pas. La figure 3 illustre le genre de prédictions que nous obtenons avec le modèle de discrimination directe, mais la densité des prédictions est très similaire pour le modèle de discrimination indirecte. Il y a clairement un certain biais statistique des prédictions qui sont souvent plus élevées que la réalité. Ceci est potentiellement causé par les cas extrêmes de montants réclamés et la simplicité du modèle de régression utilisé. Puisque le but de l'exercice est de mesurer

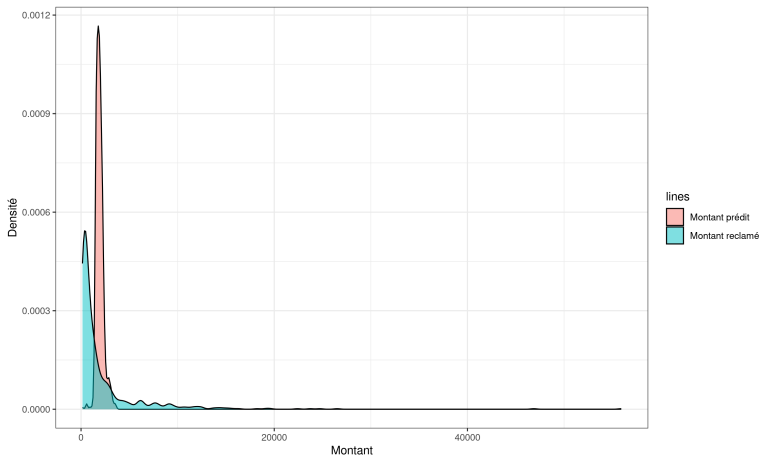

FIGURE 3 – Répartition des montants réclamés réels et prédits, excluant 0, pour le modèle avec discrimination directe

l'équité des modèles, nous nous restreignons à constater ce fait. Nous n'avons pas eu le temps de régler ce problème.

Maintenant, regardons l'équité des modèles. Tout d'abord, regardons la distribution des prédictions pour les hommes et les femmes selon les quantiles des montants réclamés réels. Cela permet de donner une idée intuitive de ce que la PAQ mesure (équation (10)).

On peut voir que pour le modèle qui fait de la discrimination directe, à gauche, entre les quantiles 2 et 3, et les quantiles 3 et 4, on voit clairement que les montants prédits par le modèle qui discrimine directement sont différents entre les hommes et les femmes. Le PAQ devrait être plus élevé pour tenir compte de cette différence. Pour le modèle de discrimination indirecte, la disparité entre les montants prédits est plus faible.

Cette intuition est confirmée au vu de la figure 5. En effet, la somme de la différence de moyenne pour chaque quantile est considérablement plus élevée pour le modèle de discrimination directe.

Nous avons ensuite entraîné plusieurs modèles en faisant varier l'importance de l'équité dans la pénalisation. Les résultats sont présentés à la figure 6.

On peut voir que les valeurs de PAQ ne convergent pas vers 0 comme on aurait voulu, ce qui peut signifier que la métrique est plus difficile à optimiser pour notre algorithme d'optimisation.

Si on prend le modèle associé à la ligne pointillée, on obtient la figure 7, qui présente les résultats sur les données de test. Clairement, le modèle de discrimination directe pénalisé est plus équitable que sans pénalisation. Le modèle de *fairness through unawareness* reste le plus équitable selon la PAQ avec ces données. De plus, on note que notre modèle pénalisé a une erreur quadratique moyenne légèrement plus élevée que les deux autres modèles. Il est possible que la pénalisation impacte les performances de l'algorithme, ou simplement que la distribution des données mène à un biais vers des valeurs supérieures dans l'algorithme.

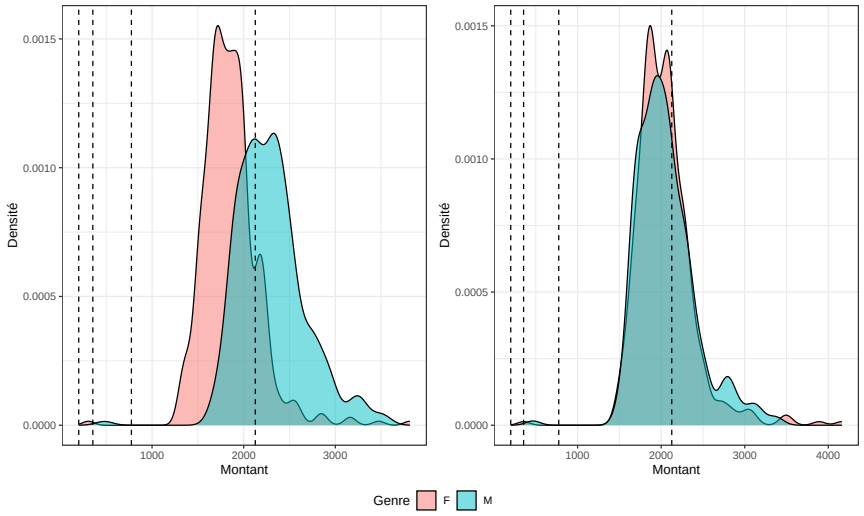

FIGURE 4 – Répartition des prédictions (excluant 0) homme et femme pour la discrimination directe (gauche) et indirecte (droite)

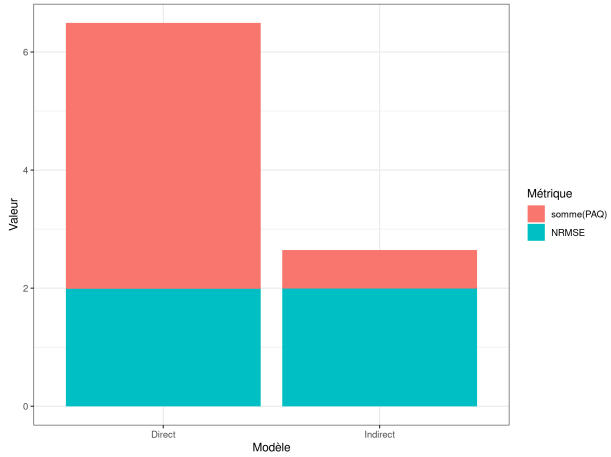

FIGURE 5 – NRMSE et PAQ pour le modèle de discrimination directe et indirecte (excluant 0)

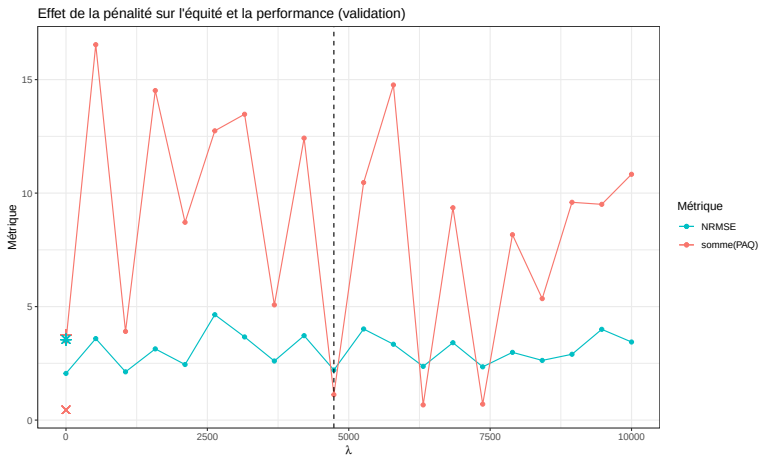

FIGURE 6 – Effet de la pénalité PAQ sur l'équité et les performances pour la régression gamma
+ représente le modèle de discrimination directe. × représente le modèle de discrimination indirecte.

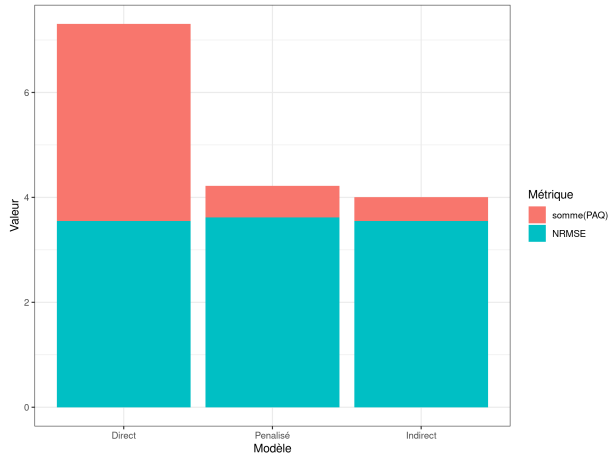

FIGURE 7 – NRMSE et PAQ pour le modèle de discrimination directe, indirecte et modèle pénalisé (test)

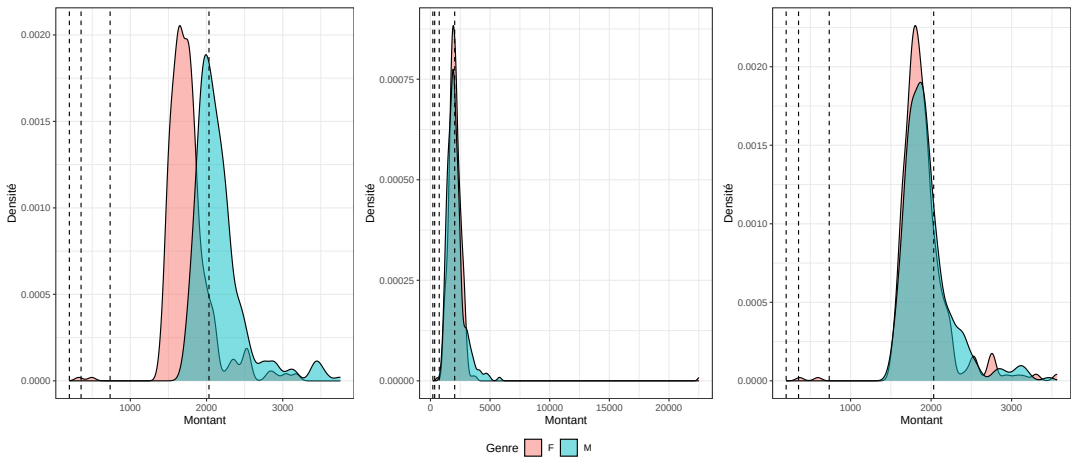

FIGURE 8 – Densité des prédictions des modèles (excluant 0) en test

Gauche : Discrimination directe. Milieu : Discrimination indirecte (*fairness through unawareness*) . Droite : Modèle pénalisé.

Finalement, on peut comparer à nouveau les performances des modèles à l'aide des graphiques permettant de mesurer intuitivement la PAQ. La figure 8 permet encore une fois de constater que le modèle de discrimination directe est moins bon que les autres. Il est plus difficile visuellement de comparer les deux modèles de droite. Néanmoins, on peut voir que le modèle pénalisé prédit plus souvent des montants réclamés élevés, comme indiqué précédemment.

5 Que faire dans le cas où un attribut devient protégé

Dans le cas où un attribut devient protégé et qu'il n'est plus permis de le stocker, il est non seulement interdit d'utiliser la variable dans les modèles, mais il n'est également plus possible de vérifier si les modèles discriminent. On enlève la variable du modèle et on espère ne pas faire de discrimination.

Il serait possible de croiser la base de données avec des données ouvertes, par exemple des données territoriales de Statistique Canada, pour faire une prédiction sur l'attribut protégé, ou utiliser un attribut similaire (proxy) à l'attribut protégé. Par exemple, on pourrait prédire que la personne vivant dans la zone A avec une hatchback est probablement un homme. On peut utiliser cette information pour vérifier qu'on ne discrime pas contre les femmes. On viole toutefois la vie privée de cette personne, en plus de violer l'interdiction d'utiliser des données de ce type pour faire nos prédictions.

Dans le cas où on peut utiliser l'attribut protégé pour vérifier que l'on ne discrimine pas contre cet attribut, notre approche est pertinente. On enlève la variable du modèle (évite la discrimination directe) puis on pénalise le modèle pour éviter de discriminer. On utilise les paramètres trouvés lors du dernier entraînement pour initialiser notre modèle.

Il serait techniquement possible de justifier l'utilisation des attributs protégés dans le modèle. Il est

toutefois nécessaire de faire attention, puisque cela peut-être difficile à justifier pour une personne qui n'est pas familière avec la modélisation statistique. De plus, des métriques telles que la parité démographique a tendance à faire tendre les coefficients associés aux attributs protégés vers 0 dans les tests que nous avons effectués, donc ce ne semble pas être pertinent.

6 Conclusion

Pour répondre à la question comment rendre l'assurance plus équitable, nous avons défini ce qu'est la discrimination en assurance, plusieurs définitions mathématiques de l'équité, fait une petite revue de littérature sur l'état de l'art et introduit une nouvelle métrique Parité Actulab par Quartile (PAQ) inspirée de la Parité actuarielle par Groupe (PAG). Nous avons expliqué l'intuition derrière la PAQ et nous l'avons utilisée pour pénaliser des modèles linéaires généralisés pour réduire la disparité entre les femmes et les hommes dans des groupes avec un risque semblable. Nous avons décidé d'adapter la mesure de Parité actuarielle par groupe de Dolman and Semenovich (2018) pour diminuer le calcul nécessaire et obtenir une partition par quantile mieux adapté aux données. Dans le cas des données de dataCar, nous avons proposé d'utiliser des quartiles.

Pour le modèle discret, la disparité entre les hommes et les femmes était déjà raisonnable dans notre jeu de données selon la métrique equalized odds si nous nous fions à la règle du 4/5. Nous avons toutefois montré que notre modèle pénalisé réussissait à améliorer l'équité du modèle.

Pour le modèle gamma, puisque le modèle de *fairness through unawareness* semblait déjà équitable par rapport à la métrique PAQ, nous avons montré qu'il était possible de diminuer la discrimination d'un modèle utilisant l'attribut protégé pour modéliser le montant réclamé. En effet, nous avons obtenu une réduction de 84% de la PAQ en modifiant la fonction de perte.

Nous avons donc proposé une approche, nous avons programmé un modèle linéaire généralisé équitable de telle façon que le code soit aisément modulable et réutilisable et avons démontré l'efficacité de notre méthode. Même si nous avons fait nos expérimentations dans le domaine de l'assurance automobile, il est clair que nous pourrions généraliser ce modèle à d'autres types d'assurances et même à d'autres domaines entièrement. Dès qu'on cherche à éviter de faire de la discrimination avec un GLM, que ce soit pour la classification ou la régression, notre modèle ainsi que la théorie développée peuvent s'adapter aisément.

Il serait toutefois nécessaire de continuer les expérimentations pour vérifier si la conjecture Actulab tient et s'il est possible de trouver des cas limites où cette conjecture ne tient jamais. Cette conjecture est intuitive, mais il serait nécessaire de quantifier son efficacité avant de l'utiliser dans un cadre réel. Il serait de plus intéressant de voir l'impact de l'utilisation de plusieurs quantiles différents sur la discrimination du modèle.

Références

Besse, P. (2020). Détecter, évaluer les risques des impacts discriminatoires des algorithmes d'ia.

Dolman, C. and Semenovich, D. (2018). Algorithmic fairness : Contemporary ideas in the insurance context.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination.

Kamiran, Faisal, K. and al. (2012). Decision theory for discrimination-aware classification.

Kamishima, T. and al. (2012). Fairness-aware classifier with prejudice remover regularizer.

Khoury, R. (2022). Discrimination dans les systèmes de recommandation.

Mosley, R. and Wenman, R. (2022). Methods for quantifying discriminatory effects on protected classes in insurance.

Williamson, R. and Menon, A. (2019). Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR.

Zhang, B. H. and al. (2018). Mitigating unwanted biases with adversarial learning.